EEG decoding with conditional identification information

Pengfei Sun [1], Jorg De Winne[1], Paul Devos[1] and Dick Botteldooren[1]
[1] Department of Information Technology, WAVES Research Group, Ghent University, Gent, Belgium
E-mail: {pengfei.sun, jorg.dewinne, p.devos, dick.botteldooren} @ugent.be

Summary: Decoding EEG signals is crucial for unraveling human brain and advancing brain-computer interfaces. Traditional machine learning algorithms have been hindered by the high noise levels and inherent inter-person variations in EEG signals. Recent advances in deep neural networks (DNNs) have shown promise, owing to their advanced nonlinear modeling capabilities. However, DNN still faces challenge in decoding EEG samples of unseen individuals. To address this, this paper introduces a novel approach by incorporating the conditional identification information of each individual into the neural network, thereby enhancing model representation through the synergistic interaction of EEG and personal traits. We test our model on the WithMe dataset and demonstrated that the inclusion of these identifiers substantially boosts accuracy for both subjects in the training set and unseen subjects. This enhancement suggests promising potential for improving for EEG interpretability and understanding of relevant identification features.

Keywords: EEG, neural network, classification, human-computer interfaces.

1. Introduction

The interplay between humans and artificial intelligence (AI) remains suboptimal, lacking the depth of engagement and synchrony inherent to human-to-human interactions. In pursuit of bridging this gap, there has been a marked shift towards leveraging neurophysiological insights, particularly through the prism of electroencephalography (EEG), to elucidate underlying cerebral mechanisms and refine the human-computer interface. The WithMe [1] experiment exemplifies this approach by presenting subjects with specific auditory and visual stimuli, thereby enabling the differentiation between target and distractor stimuli, whilst concurrently capturing the resultant EEG data. However, another challenge that arises is decoding the collected EEG signals, and in particular how to effectively decode and analyze the data to extract meaningful information.

Recently, advancements in machine learning have shown notable advantages in extracting intricate information from EEG signals [2][3]. Among these techniques, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) stand out. CNNs process EEG signals as frames, synthesizing this data to make final decisions. RNNs, in contrast, retain information from previous inputs, showcasing an ability to recognize and remember temporal sequences, which is crucial for tasks needing long short-term memory of past events. Initial explorations employing deep neural networks (DNN) and conventional machine learning paradigms have yielded promising outcomes by directly processing EEG signals from WithMe experiment to classify the target/distractor [4]. The majority of neural network solutions for EEG decoding utilize fully supervised learning methods, meaning they refine their parameters based on hard-labeled data. However, this method tends to create models that are highly specialized for the tasks they're trained on, which may not perform well on different tasks or with new individuals [5]. In addition, the heterogeneity in individual brain activity patterns poses a significant challenge to the current deep learning frameworks, particularly in decoding EEG signals from subjects not represented in the training corpus.

Notwithstanding these challenges, the WithMe experiment has unveiled certain individual characteristics, notably sensory dominance [6], that substantially influence experimental outcomes. For instance, participants with auditory dominance exhibited superior performance across various metrics and conditions compared to their visually dominant peers [1]. This observation prompts a reevaluation of the role individual-specific traits play in modulating EEG signals in an attention and working memory task. It raises the intriguing possibility that integrating a compendium of these personal attributes into computational models could potentially enhance their representational capacity. By decoding the latent interplay between personal traits and EEG patterns, this research aspires to not only bolster decoding accuracy for familiar subjects but also extend predictive proficiency to novel individuals.

To tackle this issue and recognize that personal characteristics can impact experiment results, we propose a novel framework that incorporates conditional identification information into the EEG decoding process. Thus, a network employing fully supervised learning can utilize not only the hard label information but also the conditional identification information. This paper is dedicated to investigating the viability of this innovative methodology, with the ultimate aim of advancing human-AI interactions.

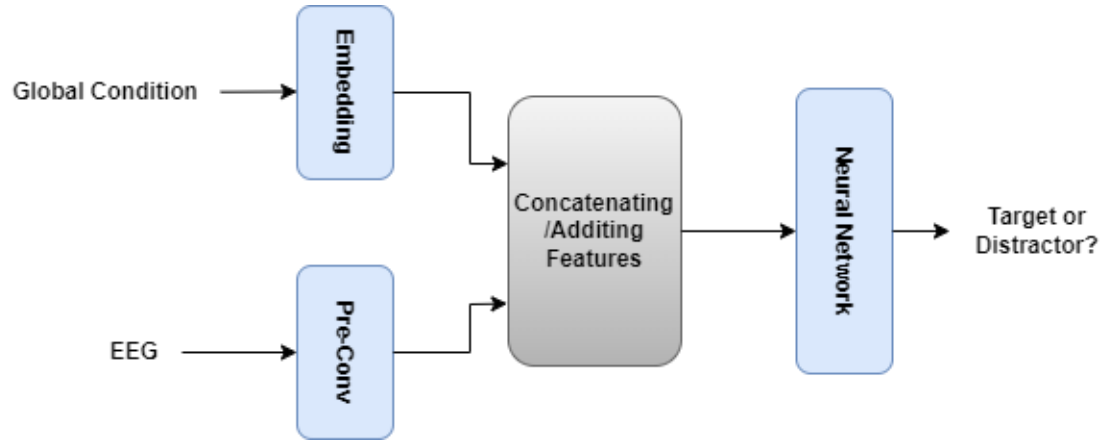


Fig.1. Overview of the Proposed Framework.

2. Overview

2.1. Overview of framework

Figure 1 depicts the structure of our proposed framework, comprising two main components: (1) Embedding the conditional identification information, employing a 16-neuron embedding layer designed to transform conditional identification information. Alongside, the pre-convolution layer, functioning as an identity layer in this study, encodes EEG data. (2) Decoding the integrated features, where this section is capable of utilizing various renowned neural network models to distinguish effectively between target and distractor stimuli.

2.1. Conditional identification information

The conditional identification information of each individual is utilized for target/distractor classification. This auxiliary conditioner employs an embedding layer to encode the identification attributes of each subject, transforming them into a comprehensive subject embedding. These embeddings, along with the EEG patterns, are then synergistically fused and introduced into the neural network. Through this extension, we expect to enhance the model's capability to learn a more generalized representation across diverse individuals, more precisely accounting for the variability in their brain activity characteristics. In this paper, we choose four distinct variables to examine their influence on the outcomes: 'Auditive/Visual Dominance', 'Sex', 'Music Education', and 'Active musician'. The former is assessed with the experiment proposed in [6], the others can be obtained via a simple questionnaire.

3. Experiment and Results

3.1. Dataset

Our model was trained and evaluated using the dataset from the WithMe experiment [1]. This experiment presented target and distractor digits to the subject asking them to remember and rename the targets. Four conditions were tested: simple sequence of visual stimuli, rhythmic presentation of targets, simple presentation supporting targets with a short beep, rhythmic presentation supported by beeps. The dataset encompasses data from a total of 42 participants. For training and internal testing, we randomly selected 38 participants further referred to as Within-subjects. The remaining 4 participants' data were reserved to assess the generalizability of the models and are referred to as Unseen-subjects. Specifically, we partitioned the WithMe data into a training set and two testing sets: Within-subjects, which comprises 18,176 training instances and 4,580 testing instances, and Unseen-subjects, which includes 2,400 testing instances. Preprocessing of the EEG data involved re-referencing each channel to the average activity of the mastoid electrodes. The data were then band-pass filtered between 1 and 30 Hz and subsequently downsampled to 64 Hz. Then, the data were segmented into 1.2 s epochs based on trigger events, with the final preprocessing step normalizing the EEG channel data to ensure zero mean and unit variance for each sample. The data and code can be accessed via https://github.com/sunpengfei1122/Withme-EEG-dataset.

Table 1. The results of three models on WithMe dataset is presented and compared to the models with the global condition id information.

Datasets	***Models***	**Within Accuracy**	**Unseen Accuracy**
WithMe	EEGNet	81.67%	76.42%
	+ IDs	**86.29%**	**79.08%**
	LSTM	80.09%	74.00%
	+ IDs	**81.18%**	**76.00%**
	DMU	81.94%	75.92%
	+ IDs	**82.21%**	**77.21%**

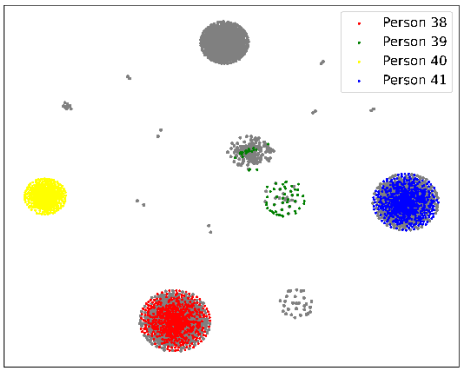


Fig. 2. Illustration of the intrinsic clustering pattern of identification information after embedding layer, as unveiled by t-SNE. The grey clustering pattern represents all the trained subjects, while the colorful pattern denotes the unseen subjects.

3.2. Implementation detail

In our experiment, we use the Adam optimizer to optimize the weights with a constant learning rate of 0.0001 and a minibatch size of 128. The EEGNet architecture features convolutional layers, starting with 16 kernels for initial temporal and spatial feature extraction from EEG signals. This is followed by depthwise and separable convolutions using 32 and 64 kernels, respectively, for efficient feature learning. For the LSTM [7] and DMU [8] models, a single recurrent unit with 64 neurons is utilized. Specifically, for the DMU's delay gate, the total number of delays is set to 20, considering the short duration of each sample. These models are developed within the PyTorch framework, adhering to default training methodologies. All modules are trained and updated in an end-to-end manner.

3.3. Results

Table 1 delineates the performance of three baseline models (EEGNet, LSTM, and DMU) and their counterparts incorporating our conditional identification (IDs) information branch of each participant. Remarkably, the EEGNet model, when enriched with conditional information, exhibits substantial enhancements in performance in both within-subject and unseen-subject. Furthermore, the addition of conditional IDs to LSTM and DMU models also yields marked improvements, particularly in the recognition of unseen subjects, indicating that the network has acquired more generalized representation of EEG. Additionally, t-distributed Stochastic Neighbor Embedding (t-SNE) [9] visualizations across all individuals of the conditional identification embedding layer in Figure 2 reveal a tendency for unseen subjects (person 38 to 41) to gravitate towards familiar centroids. Intriguingly, while up to 14 cluster centers (according to experimental data statistics) are

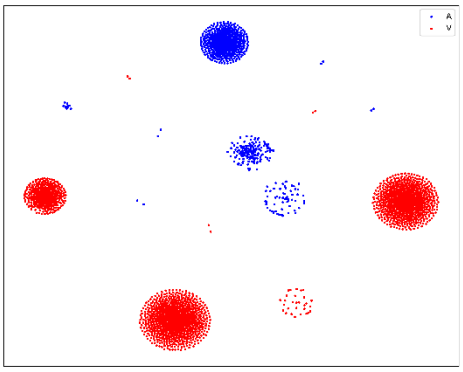


Fig. 3. Illustration of the intrinsic clustering pattern of Audio/Visual (A/V) dominance.

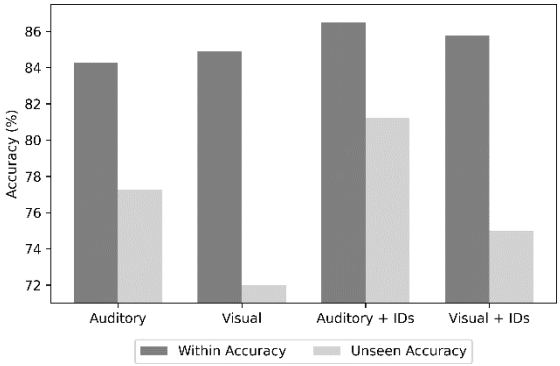


Fig. 4. Classification Performance on the WithMe Dataset based on Auditory/Visual dominance.

theoretically possible given the 4-dimensional input IDs, only 7 prominent clusters emerge, suggesting that not all features exert a significant influence on the model's performance.

3.4. Analysis

To further validate our proposed framework, we focus on a key personal trait: Dominance. The WithMe study [1] demonstrated that participants with auditory dominance outperformed visually dominant individuals in all metrics and scenarios. As illustrated in Fig. 3, this distinction is shown as two distinct clusters based on dominance type. Subsequently, we evaluate EEGNet in two contexts: Auditory vs. Visual dominance for EEG classification. Fig. 4 reveals that, for the vanilla EEGNet, visually dominant individuals slightly outperform their auditory counterparts in within-subject tests but fare worse in unseen situations. However, when incorporating IDs information, the auditory group excels in both scenarios, consistent with our experimental findings. This observation may suggest that participants who performed better in the experiment tended to have clearer representations in their EEG signals that can be recognized by neural networks.

4. Conclusions

In this paper, we investigate the effectiveness of incorporating additional conditional identification information into neural network architectures for the classification of target versus distractor stimuli based on EEG. Through the deployment of an auxiliary global conditioner that utilizes an embedding layer to capture unique individual traits, our methodology not only enhances the model's precision in the same subjects but also amplifies its generalizability to unseen subjects, adeptly navigating the variety of neural responses observed in diverse individuals. Our results suggest that incorporating a personalized and context-aware conditioner is a promising approach to enhance the performance and reliability of EEG classification in real-world scenarios.

Acknowledgements

This work was supported in part by the Flemish Government under the "Onderzoeksprogramma Artificiele Intelligentie (AI) Vlaanderen" and the Research Foundation - Flanders under grant number G0A0220N (FWO WithMe project).

References

[1]. De Winne, J., Devos, P., Leman, M., & Botteldooren, D. (2022). With No Attention Specifically Directed to It, Rhythmic Sound Does Not Automatically Facilitate Visual Task Performance. *Frontiers in Psychology*, *13*, 894366.

[2]. Cai, S., Zhang, R., Zhang, M., Wu, J., & Li, H. (2024). EEG-based Auditory Attention Detection with Spiking Graph Convolutional Network. *IEEE Transactions on Cognitive and Developmental Systems*.

[3]. Cai, S., Li, P., & Li, H. (2023). A bio-inspired spiking attentional neural network for attentional selection in the listening brain. *IEEE Transactions on Neural Networks and Learning Systems*.

[4]. Mortier, Steven, et al. "Classification of targets and distractors in an audiovisual attention task based on electroencephalography." *Sensors* 23.23 (2023): 9588.

[5]. Sarkar, P., & Etemad, A. (2020). Self-supervised ECG representation learning for emotion recognition. *IEEE Transactions on Affective Computing*, *13*(3), 1541-1554.

[6]. Giard, M. H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. *Journal of cognitive neuroscience*, *11*(5), 473-490.

[7]. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

[8]. Sun, P., Wu, J., Zhang, M., Devos, P., & Botteldooren, D. (2023). Delayed memory unit: modelling temporal dependency through delay gate. *arXiv preprint arXiv:2310.14982*.

[9]. Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(11).

Alljoined1 - A dataset for EEG-to-Image decoding

Jonathan Xu [*1 2], Bruno Aristimunha [*3 4], Max Emanuel Feucht [*1 5],
Emma Qian [†1], Charles Liu [†1 2], Tazik Shahjahan [†1 2], Martyna Spyra [†1], Steven Zifan Zhang[1,6],
Nicholas Short[1,6], Jioh Kim[1,6], Paula Perdomo[1,6], Ricky Renfeng Mao[1,2], Yashvir Sabharwal[1],
Michael Ahedor[1], Moaz Shoura[6], Adrian Nestor[6]

Abstract

We present Alljoined1, a dataset built specifically for EEG-to-Image decoding. Recognizing that an extensive and unbiased sampling of neural responses to visual stimuli is crucial for image reconstruction efforts, we collected data from 8 participants looking at 10,000 natural images each. We have currently gathered 46,080 epochs of brain responses recorded with a 64-channel EEG headset. The dataset combines response-based stimulus timing, repetition between blocks and sessions, and diverse image classes with the goal of improving signal quality. For transparency, we also provide data quality scores. We publicly release the dataset and all code at https://linktr.ee/alljoined1.

1. Introduction

In the fields of cognitive neuroscience and medical imaging, advancements in deep learning have led to unparalleled precision in decoding brain activity [7, 21, 38–40]. Researchers have translated the intricate patterns of brain activity during various cognitive processes by utilizing neuroimaging modalities, such as functional Magnetic Resonance Imaging (fMRI) and electroencephalography (EEG).

In this context, one particular area of interest is image reconstruction, which involves the decoding of neural responses to visual stimuli, offering insights into how the brain encodes and processes visual information [7, 10, 11, 25, 37, 39].

While fMRI has traditionally been the primary tool for image reconstruction due to its excellent spatial resolution, its low temporal resolution severely delimits actual clinical usage. On the other hand, EEG is a medical modality available in everyday clinical contexts with an excellent time resolution [35, 41, 42]. As neurons fire at millisecond scales, the high temporal resolution provided by EEG is crucial for real-time monitoring of neural dynamics [13, 18, 47]. Additionally, EEG is portable, more accessible to set up, and much more cost-effective than fMRI, making it suitable for real-world applications, including brain-computer interfaces and clinical diagnostics.

The development of very large fMRI-to-image datasets has proven foundational for recent breakthroughs in deep-learning image reconstruction projects. Inspired by the need for such datasets in the EEG domain, we present **Alljoined1**, a novel, large-scale dataset covering a wide range of naturalistic stimuli that allows for robust, generalizable image reconstruction efforts. Our contributions are as follows:

- We propose a stimulus presentation approach that tailors trial duration and session and block repetitions to maximize the signal-to-noise (SNR) ratio.
- We introduce a diverse dataset of EEG responses to 9k unique naturalistic images for each of the eight participants, with 1k additional images shared between participants.
- We perform qualitative comparisons against current EEG-to-image datasets.

2. Related Work

2.1. EEG-to-Image Datasets

EEG-to-image datasets consist of EEG waveforms recorded while participants watch visual stimuli, enabling the study of neural representations in the brain. However, previous research on EEG-based image reconstruction has often relied on datasets exhibiting severe limitations regarding acquisition design or generalizability to naturalistic stimuli [28, 41, 50].

A popular EEG-image dataset is *Brain2Image* [23], which consists of evoked responses to a visual stimulus from distinct image classes. Each block consists of stimuli corresponding only to a single image class. There are 40 classes, with 50 unique images in each class. This dataset has been criticized for having no train-test separation during recording, block-specific stimuli patterns, and lack of

[*] equal contribution. [†] core contribution. [1]Alljoined [2]The University of Waterloo [3]Université Paris-Saclay, Inria TAU, CNRS, LISN [4]Federal University of ABC [5]Vrije Universiteit Amsterdam [6]University of Toronto

consistency across different frequency bands. These factors can incorrectly boost model performance by giving extraneous proxy information about the block rather than the actual image-specific brain responses [4, 29]. An extensive study highlights how many recent EEG-based image reconstruction attempts depend crucially on their block design, demonstrating how similar analytical approaches are not capable of meaningfully decoding EEG signals in a rapid serial visual paradigm (RSVP) [29], even when collecting large amounts of data for only a single subject [3].

Recent studies achieving impressive reconstruction results have relied on data collected with flawed block designs [6, 24, 27], calling the validity of their results into question. As recommended by [4, 29], the stimuli within each block in our dataset were chosen randomly across a variety of natural images, effectively minimizing the risk of block-class correlations.

The diversity of decoding stimuli further limits current EEG-based image reconstruction datasets. While studies like *Brain2Image* or [3] consist of images belonging to 40 classes, several studies utilize a dataset of visual imagery of characters and objects belonging to only 10 different classes, *ThoughtViz* [48]. Such a discretized representation of real world objects fails to account for the continuous, diverse quality of naturalistic stimuli. The same limitation applies to studies utilizing a severely limited quantity of naturalistic stimuli. Approaches to EEG-based image reconstruction derived from the *ThoughtViz* [34, 41], *Brain2Image* [6, 24, 27], or other equally selective datasets [2, 50], may thus suffer from generalizing well to diverse, real-world stimuli. Moreover, a limited number of image classes may encourage image reconstruction models to generate images class-conditionally, rather than reconstructing images based on (continuous) brain-encoded semantic or perceptual attributes of an image.

To account for the diverse and continuous nature of naturalistic images, Alljoined1 consists of 1) 10,000 images per participant 2) that belong to at least one of 80 MS-COCO [31] object categories. Importantly, each MS-COCO category is broader than a single object class (e.g. the *things* category includes car, skateboard, hat, etc.), and each image can belong to up to 5 classes [30].

There are also existing datasets that include naturalistic stimuli, but compromise in other domains. The *MindBigData* initiative [49], or [3] capture a wide sample of images, but are derived only from a single individual, limiting the potential of training image reconstruction models that generalize to other individuals. The *THINGS-EEG1* [17] and *THINGS-EEG2* [14] datasets were acquired using short image presentation times of 50 and 100 ms, and a stimulus onset asynchrony of 100 and 200 ms.

Although the rapid serial visual presentation [16] paradigm proposes disentangling the temporal dynamics of visual processing and categorical abstraction of non-target stimuli, it is not ideal for capturing cortical image processing beyond early visual activity with low noise. We see that [43] obtained the highest accuracy with their EEG-image classifier when focusing on 320-480 ms after stimulus onset, and [36] is able to extract relevant decoding features even around 550 ms after stimulus onset. This suggests that while it takes 50-120 ms for object recognition of a stimulus to register in the visual cortex, a longer stimulus period is beneficial for accuracy on downstream tasks. Alljoined1 consists of extensive data from eight participants, measured with an inter-stimulus interval of 300 ms, which captures important hallmarks in visual processing while maintaining a high presentation frequency [17, 46]. This setup might furthermore allow us to overcome the limitations in decoding image content from EEG activity in RSVP paradigms, as previously reported in [3].

2.2. fMRI-Image Datasets

The recent development of large functional magnetic resonance imaging (fMRI) datasets has enabled researchers to decode and reconstruct images observed by humans with unprecedented accuracy.

The *Brain, Object, Landscape Dataset* (BOLD5000) [9] contains brain responses from 4 human participants who viewed 5,254 images depicting natural scenes from the Scene UNderstanding (SUN) [51], MS-COCO [31], and ImageNet datasets [12]. Similarly, in the *Generic Object Decoding Dataset* (GOD) [22], 1,200 images from the ImageNet database were cropped and shown to 5 participants, resulting in one of the first datasets to establish methods for decoding generic object categories from brain activity.

The *Natural Scenes Dataset* (NSD) [5] consists of the brain responses of 8 human participants passively viewing 9,000–10,000 color natural scenes from MS-COCO. This magnitudes-larger dataset has fueled leaps in reconstruction accuracy seen in recent work like MindEye2 [40]. However, the adaptation of such impressive achievements to real-life contexts is quite limited, as MRI scanners are notoriously expensive and difficult to access.

3. Methods and Materials

3.1. Participants

We collected data from eight participants (six male, two female), with an average age of 22 $\pm$ 0.64 years, all with normal or corrected-to-normal vision, right-handed. All participants were healthy, with no neurocognitive impairments, except 2 participants who reported a history of mental health disorders (e.g. GAD, ADHD). Each participant provided informed consent. The Research Ethics Board approved the procedures as **suppressed for double-blind review**. We note that there are potential limitations of the

study due to the imbalance between the genders of the participants and the low age disparity, which could influence bias and learning with AI models.

3.2. Stimuli

We use the same visual stimuli as what was shown in the fMRI Natural Scenes Dataset (NSD) [5], consisting of $70,566$ images portraying everyday objects and situations in their natural context. All NSD images are drawn from the MS-COCO dataset [30], including annotations about objects and their corresponding category contained in the image. Each image can contain more than one object and more than one object category. These fine-grained object categories are further grouped into *supercategories*, each of which comprehensively includes all related categories as defined subsets.

The current study uses a subset of the first 960 images in the 1000 images shown across all participants in the NSD study. These images are drawn from the *shared1000* subset of the NSD dataset, which comprises 1000 specially curated images that all participants in the original NSD study were presented with [5]. Within this subset of the NSD dataset, the supercategory *person* was most represented, occurring in 50.94% of all images, followed by animal (23.54%) and vehicle (23.33%). The distribution of the supercategories is shown in Figure 1.

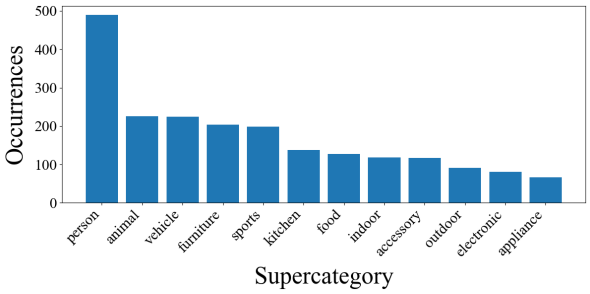


Figure 1. Top 12 most frequently occurring *supercategories* in our dataset.

3.3. Procedure

Images were displayed to participants over the course of multiple one-hour-long sessions. Each session consisted of 16 blocks, wherein images in the first 8 unique blocks were repeated in the second 8 blocks.

The repeated blocks (e.g., blocks 1 & 8, 2 & 9, etc.) contained the same stimuli but in a shuffled order to avoid sequence effects. Within each block, 120 images from the NSD dataset were presented twice, as well as 24 oddball stimuli, amounting to 264 images per block.

Given the within-block and the between-block repetitions of NSD images, each NSD image was presented 4 times to obtain a higher signal-to-noise ratio of the evoked neural responses. Within each trial, an image (NSD or oddball) was presented for 300 ms, followed by 300 ms of a black screen; a white fixation cross was visible on the screen throughout the entire trial.

At the end of each trial, an extra jitter time between 0-50 ms was added for randomness. To ensure focus, participants were prompted to press the space bar when two consecutive trials contained the same image. These oddball trials occurred 24 times within each block; oddballs trials have been discarded from the dataset due to motion artifacts, EEG repetition suppression, and other issues.

3.4. Hardware Setup

We recorded data using a 64-electrode BioSemi ActiveTwo system, digitized at a rate of 512 Hz with 24-bit A/D conversion. The montage was arranged in the International 10-20 System, and the electrode offset was kept below 40 mV. We used a 22 inch Dell monitor at a resolution of 1080p/60Hz to display the visual stimulus. As depicted in Figure 3, the monitor was positioned centrally and placed at a distance of 80 cm to maintain a $3.5°$ visual angle of stimuli. We avoided larger angles to minimize the occurrence of gaze drift.

3.5. Pre-processing

Regarding the dataset pre-processing, we follow recent work on the importance of separating the biomarkers from the central nervous and peripheral systems, as described in [8], and applied the minimum necessary steps. This dataset was pre-processed using the MNE-PYTHON library [15].

Filtering Initially, we applied band-pass filtering with a low frequency of 0.5 Hz and a high frequency of 125 Hz with overlap-add finite impulse response filtering, with range based on [45]. We then apply a notch filter at 60Hz to eliminate power line noise.

Independent Component Analysis (ICA) Next, we performed an ICA decomposition using a FastICA model [1, 19] to separate non-gaussian biological artifacts noise from the signal source. We used a decomposition that retained 95% of the variance and excluded ICs corresponding to eye blinks on the raw data.

Epoching We segmented the continuous data into *epochs*, with each epoch starting at -50 ms onset stimulus and ending at the end of each trial at 600 ms, as described in Section 3.2. The inter-trial jitter periods were excluded from the epochs.

Artifact correction We used the AUTOREJECT algorithm [20] to identify and handle artifact-heavy epochs. Autoreject employs a peak-to-peak threshold criterion separately

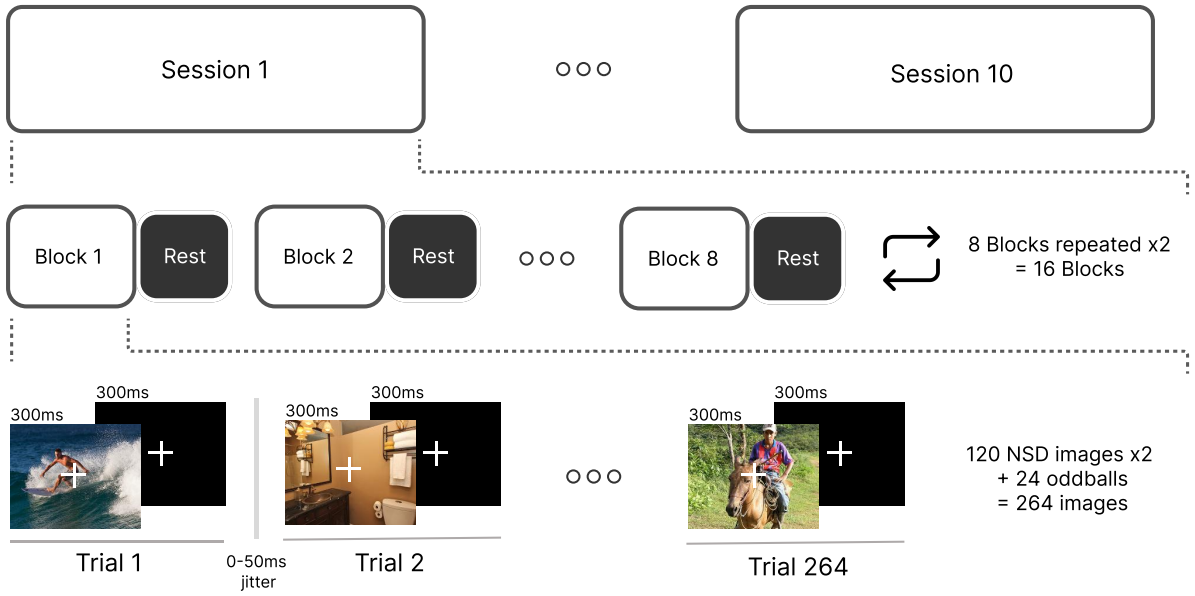


Figure 2. Schematic overview of the structure of trials, blocks, and sessions. Each of the 120 block-specific NSD images is presented twice within each block, and each of the 8 session-specific blocks is presented twice within each session. Each participant performed two sessions on different days. Each of the 10 sessions thus consists of 960 NSD images repeated four times within and across blocks, totaling 9600 unique NSD images per participant.

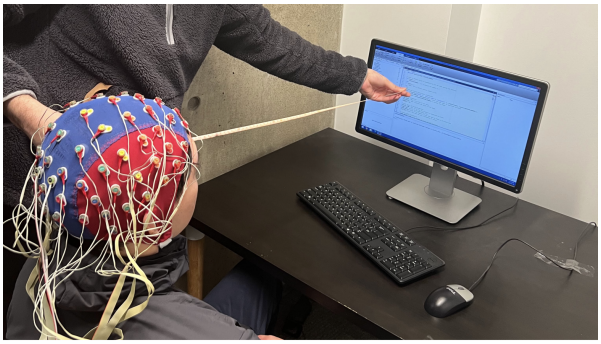


Figure 3. Experimental setup with monitor 80 cm from participant.

for each sensor to determine whether an epoch should be (i) repaired by interpolating the affected sensors using neighboring sensors, or (ii) entirely excluded from further analysis. It performs grid search to determine appropriate values for ρ, the number of channels to interpolate, and κ, the percentage of channels that must agree as a fraction of total channels for consensus. By looking at the number of erroneous sensors per trial, this approach allows correction on a per-trial basis instead of applying a single global threshold to all trials. A mean of 130.75 epochs was dropped per session, with a standard deviation of 260.44.

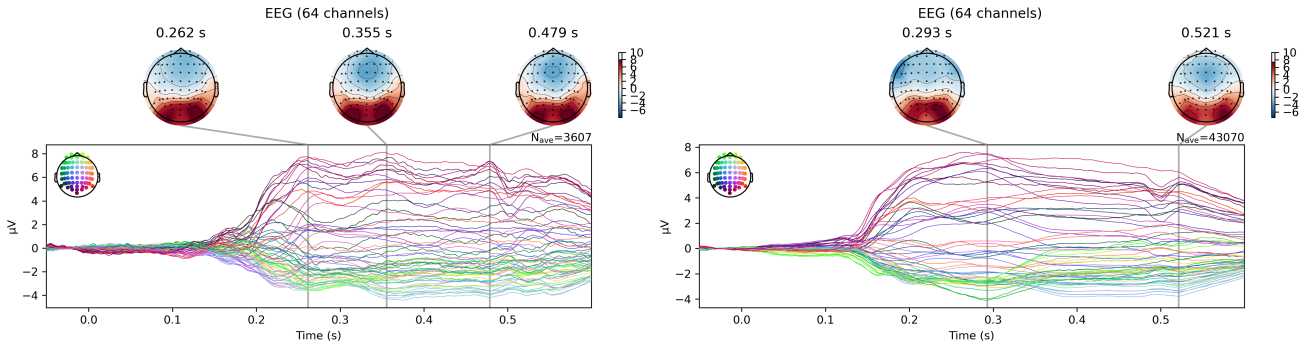
Baseline correction Finally, we re-reference our channels using an average reference scheme, before applying a baseline correction window from -50 ms to 0 ms relative to stimulus onset, following recommendations from [44] for ERP baseline. The epoch data subtracts the average activation during the baseline interval to remove noise from the signal.

4. Analysis

4.1. ERP Analysis

The distribution of event-related potentials (ERPs) across all 64 channels is displayed for a single session of one participant and averaged over all participants and sessions in Figure 5. We observe a strong consistent rise in activity beginning after 150 ms, with a peak between 250 and 300 ms. Note that this latency aligns with the timing parameters of our experimental design, which involves a 300ms presentation followed by a 300ms rest period, with an additional 0-50ms jitter. Both participant- and cohort-level activity exhibits a sustained high level of activity up until 500 ms after stimulus onset, where a consistent dip in activity is observed for both the single participant as well as the whole cohort.

Topographies of activation additionally reveal a strong concentration of positive activation at occipital, parietal, and partially temporal electrode locations and a consistently



(a) First session of the fifth participant

(b) All sessions across all participants

Figure 4. EEG topographic maps and corresponding signals at all 64 electrodes averaged over a) 3823 events for the fifth participant (left) and b) across all sessions for all participants (43070 events) in the Alljoined1 dataset (right), highlighting individual and common brain activity patterns associated with image presentation. An *event* is defined as a specific time point in the experiment.

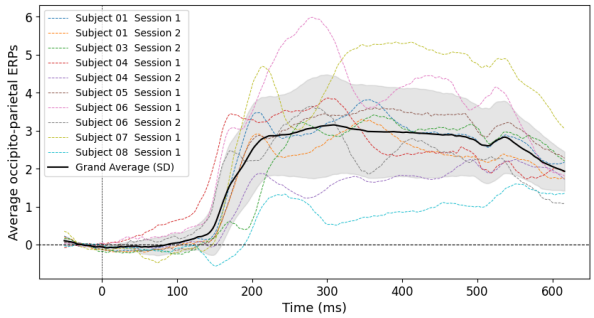


Figure 5. ERPs averaged over occipital and parietal electrodes for all participants and sessions. Shaded areas around the grand average ERP indicate standard deviations at all timepoints.

negative activation at central and frontal areas. This topographical distribution was stable across the duration of the ERP and corresponded well between the single participant and the cohort. Given the strong peak in activity at the occipital and parietal areas, we further investigated the distribution of ERPs across individual participants and sessions at the occipital and parietal electrodes, as displayed in figure 5. While the magnitude of activation differs between participants, we conclude a by-and-large consistent activation pattern across participants and sessions.

4.2. SNR Analysis

The Signal-to-Noise Ratio (SNR) serves as a pivotal metric in evaluating the efficacy of our dataset. To ascertain the SNR, we employ the Standardized Measurement Error (SME) as a gauge for noise assessment [33]. We choose SME as our metric of choice as it is able to robustly quantify the data quality for each participant at each electrode site [32]. The SME is determined by calculating the standard deviation of the aggregated waveform average for each event type across all trials and then dividing this by the square root of the event type's occurrence count. The SNR is subsequently derived by dividing the mean signal values by their corresponding SME.

Figure 6 compares the average SNR across all events in a single session for participant 5 with the average SNR across all events for both sessions concatenated. We see that the SNR is noticeably lower in the multi session graph. This is due to the increased number of repetitions for a given event at different timepoints. This leads to a disproportionately higher standard deviation value and consequently a higher SME and lower SNR. However, that is not to say that the quality of the data is worse. It actually reflects more accurate SNR values as there are more data points, distributed across different sessions. It is also observed that the single sessions graph is more volatile across time, demonstrating a greater variance in SNR values which are captured by having a less accurate metric for noise with less trials to average between. This fluctuation underscores the limited accuracy of noise metrics derived from fewer trials, thus highlighting the critical importance of incorporating repeated measures across sessions or blocks for robust SNR evaluation.

Furthermore, we observe a strong SNR increase 150 ms after stimulus onset. Note that this increase in SNR exhibits the same spiking timing we see in our earlier topographic maps and averaged ERP graphs, suggesting that meaningful activity starts to surface with a considerable delay with respect to stimulus onset.

4.3. Discussion

The ERP and topography analyses, as well as our analysis of the SNR reveal and reinforce several benefits of the acquired dataset, with regard to the stimulus design and timing.

1. **Stimulus duration:** A 300 ms presentation window as well as a subsequent 300 ms rest period allows the capture of both early and late cognitive processes, as ev-

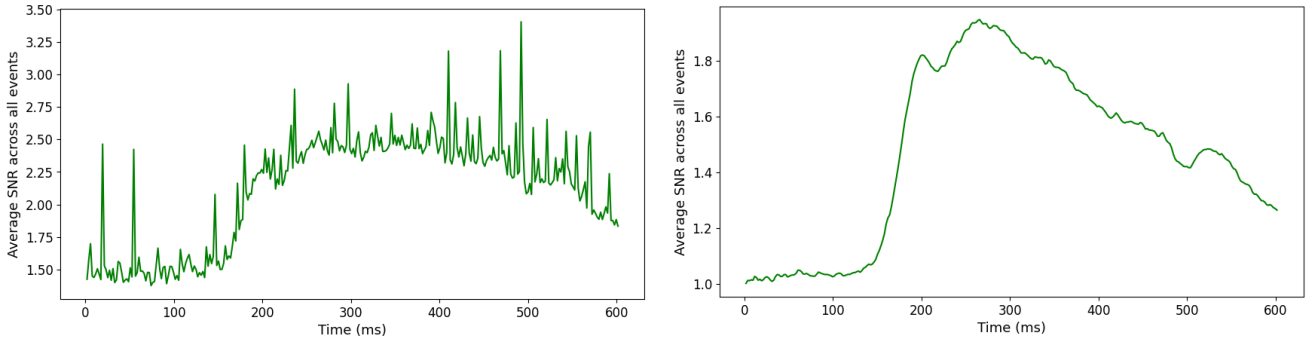


Figure 6. Signal to Noise Rate (SNR) averaged across each session, across each block, and within each block for participant 5. Left: SNR for only the first session 1, Right: SNR for all sessions.

idenced by the single subject peaks at around 262 ms up to 479 ms in Figure 6 a), and the averaged peaks at 293 ms and 521 ms in Figure 6 b), respectively. The duration of 300 ms for image presentation is sufficient for the brain to engage in both perceptual encoding and initial stages of memory processing, which may not be as effectively captured with shorter presentation times. The subsequent 300ms rest period provides a window to measure the brain's higher-level visual and semantic response to the stimuli. The whole ERP thus not only reflects the initial feed-forward transfer of sensory information to visual cortical areas but also the subsequent recurrent interactions involved in attention and semantic analysis, that unfold over hundreds of milliseconds after stimulus onset. The relevance of longer presentation times and longer stimulus-onset asynchrony is additionally supported by the sustained ERP activation presented in Figure 5, as well as the latency of SNR increase and peak in Figure 6.

2. **Comparison with prior studies:** Presentation times of only 100 ms, or stimulus onset asynchronies of only 200 ms fail to capture the rich neural dynamics associated with image processing, involving both lower and higher level processing. In THINGS EEG2 [14], with a shorter 100ms presentation time followed by 100ms of rest, the stimulus exposure may have been insufficient to elicit the full range of cognitive processes to occur. The limited time window could explain the lesser degree of neural activity in the corresponding time window. Similarly, THINGS EEG1 [17] employed a shorter 50ms presentation window followed by 50ms rest, which, while suitable for examining the earliest stages of sensory processing in the visual cortex, likely precluded the phases of the cognitive processes that unfold over a longer period. This includes higher-order mechanisms such as selective attention, working memory updating, and retrieval of semantic associations from long-term memory stores [26].
3. **Phase locking mitigation:** The inclusion of a jitter ranging from 0-50ms helps mitigate phase locking, a phenomenon where the participant's alpha-wave activity becomes synchronously aligned with the pattern of the stimuli after repeated presentations.
4. **Anticipatory bias minimization:** Additionally, the jitter prevents the participants from predicting the exact onset of the next stimulus, thus reducing the potential for anticipatory neural activity that could confound the data.

In conclusion, we choose a 300ms latency as it provides a good trade-off between capturing long-term neural activity whilst maintaining a high presentation frequency. The ideal timing of our experiment ensures the acquisition of a comprehensive ERP waveform, contributing to a more nuanced understanding of cognitive processes and neural dynamics as compared to the shorter intervals used in THINGS EEG2 and THINGS EEG1.

5. Conclusion

We introduce Alljoined1, an EEG-image dataset that uses well-timed stimuli, repetitions between blocks and sessions, and a wide distribution of natural images to create an improved dataset for image decoding tasks. We believe that its size, diversity, and quality will help promote work to better understand the mechanisms of visual processing, and in decoding visual responses in clinical and consumer brain-computer interface (BCI) contexts.

Future directions: We are eager to explore high-density EEG recordings of exclusively the occipital and parietal regions to better target regions of the brain most responsive to visual stimuli. We are also interested in conducting ablation studies on the generalizability of responses to imagined mental imagery. We further believe there is great potential in exploring continuous data collection in natural environments with a wireless headset.

Data availability: Both the raw and preprocessed EEG dataset is available on OSF. Labels to the corresponding NSD image IDs are included in the object files.

Code availability: The stimulus and preprocessing code to reproduce all the results is available on Anonymous GitHub here and here.

6. Acknowledgements

This work was sponsored by Z Fellows, Hack Grants, Moth Fund and Fiona Leng.Bruno's work was supported by DATAIA Convergence Institute as part of the "Programme d'Investissement d'Avenir", (ANR-17-CONV-0003) operated by LISN-CNRS. We would like to thank Dr Sylvain Chevallier for his valuable feedback on this manuscript.

References

[1] Pierre Ablin, Jean-François Cardoso, and Alexandre Gramfort. Faster ICA under orthogonal constraint. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4464–4468. IEEE, 2018. 3

[2] Hajar Ahmadieh, Farnaz Gassemi, and Mohammad Hasan Moradi. Visual image reconstruction based on EEG signals using a generative adversarial and deep fuzzy neural network. *Biomedical Signal Processing and Control*, 87: 105497, 2024. 2

[3] Hamad Ahmed, Ronnie B. Wilbur, Hari M. Bharadwaj, and Jeffrey Mark Siskind. Object classification from randomized eeg trials. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3844–3853, 2021. 2

[4] Hamad Ahmed, Ronnie B. Wilbur, Hari M. Bharadwaj, and Jeffrey Mark Siskind. Confounds in the Data—Comments on "Decoding Brain Representations by Multimodal Learning of Neural Activity and Visual Features". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9217–9220, 2022. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. 2

[5] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022. 2, 3

[6] Yunpeng Bai, Xintao Wang, Yan-pei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. Dreamdiffusion: Generating high-quality images from brain eeg signals. *arXiv preprint arXiv:2306.16934*, 2023. 2

[7] Yohann Benchetrit, Hubert Banville, and Jean-Remi King. Brain decoding: toward real-time reconstruction of visual perception. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[8] Philipp Bomatter, Joseph Paillard, Pilar Garces, Jörg Hipp, and Denis Engemann. Machine learning of brain-specific biomarkers from EEG. *bioRxiv*, 2024. 3

[9] Nadine Chang, John A Pyles, Austin Marcus, Abhinav Gupta, Michael J Tarr, and Elissa M Aminoff. BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Scientific data*, 6(1):49, 2019. 2

[10] Zijiao Chen, Jonathan Xu, Jiaxin Qing, Ruilin Li, and Juan Helen Zhou. Structure-Preserved Image Reconstruction from Brain Recordings. In preparation, 2023. 1

[11] Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[13] Nadine Dijkstra, Pim Mostert, Floris P de Lange, Sander Bosch, and Marcel AJ van Gerven. Differential temporal dynamics during visual imagery and perception. *Elife*, 7: e33904, 2018. 1

[14] Alessandro T. Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M. Cichy. A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage*, 264: 119754, 2022. 2, 6

[15] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. MEG and EEG data analysis with MNE-Python. *Frontiers in neuroscience*, 7:70133, 2013. 3

[16] Tijl Grootswagers, Amanda K Robinson, and Thomas A Carlson. The representational dynamics of visual objects in rapid serial visual processing streams. *NeuroImage*, 188: 668–679, 2019. 2

[17] Tijl Grootswagers, Ivy Zhou, Amanda K Robinson, Martin N Hebart, and Thomas A Carlson. Human EEG recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Scientific Data*, 9(1):3, 2022. 2, 6

[18] Assaf Harel, Iris IA Groen, Dwight J Kravitz, Leon Y Deouell, and Chris I Baker. The temporal dynamics of scene processing: A multifaceted EEG investigation. *Eneuro*, 3(5), 2016. 1

[19] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. Independent component analysis, adaptive and learning systems for signal processing, communications, and control. *John Wiley & Sons, Inc*, 1:11–14, 2001. 3

[20] Mainak Jas, Denis A Engemann, Yousra Bekhti, Federico Raimondo, and Alexandre Gramfort. Autoreject: Automated artifact rejection for MEG and EEG data. *NeuroImage*, 159: 417–429, 2017. 3

[21] Vinay Jayaram and Alexandre Barachant. MOABB: trustworthy algorithm benchmarking for BCIs. *Journal of neural engineering*, 15(6):066011, 2018. 1

[22] Tomoyasu Horikawa & Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 2017. 2

[23] Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, and Mubarak Shah. *Brain2Image*: Converting Brain Signals into Images. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1809–1817, Mountain View California USA, 2017. ACM. 1

[24] Nastaran Khaleghi, Tohid Yousefi Rezaii, Soosan Beheshti, Saeed Meshgini, Sobhan Sheykhivand, and Sebelan Danishvar. Visual Saliency and Image Reconstruction from EEG

Signals via an Effective Geometric Deep Network-Based Generative Adversarial Network. *Electronics*, 11(21):3637, 2022. Number: 21 Publisher: Multidisciplinary Digital Publishing Institute. 2

[25] Jean-Rémi King, Laura Gwilliams, Chris Holdgraf, Jona Sassenhagen, Alexandre Barachant, Denis Engemann, Eric Larson, and Alexandre Gramfort. Encoding and Decoding Framework to Uncover the Algorithms of Cognition. In *The Cognitive Neurosciences*. The MIT Press, 2020. 1

[26] Yixuan Ku. Selective attention on representations in working memory: cognitive and neural mechanisms. *PeerJ*, 6:e4585, 2018. 6

[27] Yu-Ting Lan, Kan Ren, Yansen Wang, Wei-Long Zheng, Dongsheng Li, Bao-Liang Lu, and Lili Qiu. Seeing through the Brain: Image Reconstruction of Visual Perception from Human Brain Signals, 2023. arXiv:2308.02510 [cs, eess, q-bio]. 2

[28] Lynn Le, Luca Ambrogioni, Katja Seeliger, Yağmur Güçlütürk, Marcel Van Gerven, and Umut Güçlü. Brain2pix: Fully convolutional naturalistic video reconstruction from brain activity. *BioRxiv*, pages 2021–02, 2021. 1

[29] Ren Li, Jared S. Johansen, Hamad Ahmed, Thomas V. Ilyevsky, Ronnie B. Wilbur, Hari M. Bharadwaj, and Jeffrey Mark Siskind. The perils and pitfalls of block design for eeg classification experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):316–333, 2021. 2

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 3

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[32] Steven J Luck and Emily Kappenman. A new metric for quantifying erp data quality. https://erpinfo.org/blog/2020/4/28/data-quality, 2020. 5

[33] Steven J Luck, Andrew X Stewart, Aaron Matthew Simmons, and Mijke Rhemtulla. Standardized measurement error: A universal metric of data quality for averaged event-related potentials. *Psychophysiology*, 58(6):e13793, 2021. 5

[34] Rahul Mishra, Krishan Sharma, R. R. Jha, and Arnav Bhavsar. NeuroGAN: image reconstruction from EEG signals via an attention-based GAN. *Neural Computing and Applications*, 35(12):9181–9192, 2023. 2

[35] Dan Nemrodov, Matthias Niemeier, Ashutosh Patel, and Adrian Nestor. The neural dynamics of facial identity processing: insights from EEG-based pattern analysis and image reconstruction. *Eneuro*, 5(1), 2018. 1

[36] Dan Nemrodov, Shouyu Ling, Ilya Nudnou, Tyler Roberts, Jonathan S. Cant, Andy C. H. Lee, and Adrian Nestor. A multivariate investigation of visual word, face, and ensemble processing: Perspectives from EEG-based decoding and feature selection. *Psychophysiology*, 57(3):e13511, 2020. 2

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[38] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of Neural Engineering*, 16(5): 051001, 2019. 1

[39] Paul Steven Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Cohen Ethan, Aidan James Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, and Tanishq Mathew Abraham. Reconstructing the Mind's Eye: fMRI-to-Image with Contrastive Learning and Diffusion Priors. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1

[40] Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. MindEye2: Shared-Subject Models Enable fMRI-To-Image With 1 Hour of Data. *arXiv preprint arXiv:2403.11207*, 2024. 1, 2

[41] Prajwal Singh, Pankaj Pandey, Krishna Miyapuram, and Shanmuganathan Raman. EEG2IMAGE: Image Reconstruction from EEG Brain Signals, 2023. arXiv:2302.10121 [cs, q-bio]. 1, 2

[42] Prajwal Singh, Dwip Dalal, Gautam Vashishtha, Krishna Miyapuram, and Shanmuganathan Raman. Learning Robust Deep Visual Representations from EEG Brain Recordings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7553–7562, 2024. 1

[43] Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Nasim Souly, and Mubarak Shah. Deep learning human mind for automated visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6809–6817, 2017. 2

[44] Darren Tanner, James JS Norton, Kara Morgan-Short, and Steven J Luck. On high-pass filter artifacts (they're real) and baseline correction (it'sa good idea) in ERP/ERMF analysis. *Journal of neuroscience methods*, 266:166–170, 2016. 4

[45] Yunzhe Tao, Tao Sun, Aashiq Muhamed, Sahika Genc, Dylan Jackson, Ali Arsanjani, Suri Yaddanapudi, Liang Li, and Prachi Kumar. Gated transformer for decoding human brain EEG signals. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 125–130. IEEE, 2021. 3

[46] Lina Teichmann, Martin N Hebart, and Chris I Baker. Multidimensional object properties are dynamically represented in the human brain. *bioRxiv*, 2023. 2

[47] Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *nature*, 381(6582): 520–522, 1996. 1

[48] Praveen Tirupattur, Yogesh Singh Rawat, Concetto Spampinato, and Mubarak Shah. ThoughtViz: Visualizing Human Thoughts Using Generative Adversarial Network. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 950–958, Seoul Republic of Korea, 2018. ACM. 2

[49] David Vivancos and Felix Cuesta. MindBigData 2022 A Large Dataset of Brain Signals. *arXiv preprint arXiv:2212.14746*, 2022. 2

[50] Suguru Wakita, Taiki Orima, and Isamu Motoyoshi. Photorealistic Reconstruction of Visual Texture From EEG Signals. *Frontiers in Computational Neuroscience*, 15, 2021. 1, 2

[51] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016. 2

EEGDiR: Electroencephalogram denoising network for temporal information storage and global modeling through Retentive Network

Bin Wang[a,1], Fei Deng[a,*,1] and Peifan Jiang[b,2]

[a]College of Computer Science and Cyber Security, Chengdu University of Technology, ChengDu, 629100, SiChuan, China
[b] College of Geophysics, Chengdu University of Technology, ChengDu, 629100, SiChuan, China

ARTICLE INFO

Keywords:
Electroencephalogram (EEG) Denoising
Retentive Network
Deep Learning
Signal Embedding

ABSTRACT

Electroencephalogram (EEG) signals are pivotal in clinical medicine, brain research, and neurological disorder studies. However, their susceptibility to contamination from physiological and environmental noise challenges the precision of brain activity analysis. Advances in deep learning have yielded superior EEG signal denoising techniques that eclipse traditional approaches. In this research, we deploy the Retentive Network architecture—initially crafted for large language models (LLMs)—for EEG denoising, exploiting its robust feature extraction and comprehensive modeling prowess. Furthermore, its inherent temporal structure alignment makes the Retentive Network particularly well-suited for the time-series nature of EEG signals, offering an additional rationale for its adoption. To conform the Retentive Network to the unidimensional characteristic of EEG signals, we introduce a signal embedding tactic that reshapes these signals into a two-dimensional embedding space conducive to network processing. This avant-garde method not only carves a novel trajectory in EEG denoising but also enhances our comprehension of brain functionality and the accuracy in diagnosing neurological ailments. Moreover, in response to the labor-intensive creation of deep learning datasets, we furnish a standardized, preprocessed dataset poised to streamline deep learning advancements in this domain.

1. Introduction

Electroencephalography (EEG) measures neural activity as potential fluctuations on the scalp, primarily emanating from the brain's gray matter [42] [26]. This neural activity is detected via an array of electrodes strategically placed on the scalp [35] [34]. Analyzing EEG data yields a broad range of physiological, psychological, and pathological insights [29]. Yet, the high temporal resolution characteristic of EEG signals renders them vulnerable to diverse and complex noise sources such as cardiac, ocular, and muscular artifacts, as well as environmental interference [14]. The prevalent intrusion of these noises during the acquisition phase significantly hampers the isolation of unadulterated EEG signals, thus severely restricting advancements in EEG-related research and practical applications [24] [33]. Consequently, there is an imperative need for a robust EEG denoising technique that effectively reduces noise without compromising critical signal information, which is vital for advancing EEG research.

A multitude of traditional denoising techniques for EEG signal enhancement has been advanced, encompassing both regression-based and adaptive filter-based methodologies. Regression-based strategies involve estimating the noise component with a pre-established noise model and subsequently subtracting this estimate to purify the EEG data, thereby producing a cleaner signal [21] [8]. Conversely, adaptive filtering operates on a fundamentally different principle, adjusting filter coefficients in real-time based on incoming EEG data to attenuate noise [10] [16]. However, these conventional techniques have their limitations. The tuning of hyperparameters in these methods critically affects the denoising performance, requiring expert judgment to optimize settings. Moreover, there is a risk of losing vital EEG information during noise reduction, which could detrimentally affect further analytical work.

The EEG signal is a complex waveform characterized by nonlinear features crucial for its analysis. Therefore, denoising methods must preserve these nonlinear features while eliminating noise [29]. The advancement in computer processing power and the expansion of EEG datasets [42] have spurred recent research endeavors to leverage deep learning for EEG signal denoising. Commonly employed architectures for EEG denoising networks include feedforward neural networks (FNN) [4] [40], convolutional neural networks (CNN) [1] [29], and recurrent neural networks (RNN) [41], along with their variations, such as long and short-term memory networks (LSTM) [22] [42]. As EEG is collected in the time dimension, establishing a temporal relationship between sampling points, basic network architectures have demonstrated significant improvement in denoising performance compared to traditional methods. However, they face challenges either in retaining temporal information or lacking global modeling capability while preserving temporal information. To address this, some studies have explored integrating the Transformer model [36] into EEG denoising tasks, as it [25] effectively preserves temporal information and enables efficient data parallel computation, yielding

*Corresponding author. College of Computer Science and Cyber Security, Chengdu University of Technology, ChengDu, 629100, China
woldier@foxmail.com (B. Wang); dengfei@cdut.edu.cn (F. Deng); jpeifan@qq.com (P. Jiang)
ORCID(s):
[1]Bin Wang and Fei Deng are with the College of Computer Science and Cyber Security, Chengdu University of Technology, Chengdu 610059, China(e-mail: woldier@foxmail.com, e-mail: dengfei@cdut.edu.cn).
[2]Peifan Jiang are with the College of Geophysics, Chengdu University of Technology, Chengdu 610059, China (e-mail: jpeifan@qq.com).

notable results. In EEG signal processing, it is critical to accurately capture temporal order information, as patterns and features in these signals are often synchronized with specific neurophysiological events. Although the Transformer provides a powerful algorithm to process sequential data through an attentional mechanism, it does not explicitly encode the temporal order of the signals, but rather introduces this information indirectly through positional encoding. This may lead to insufficient sensitivity to the temporal dynamics of EEG signals in some cases.

In recent years, with the rapid advancement of large language model (LLM), a novel network called Retentive Network [30] has emerged. Retentive Network gains its intuitive understanding of the temporal order in a sequence by introducing the decay mask Retention mechanism. With this mechanism, the input sequences are made naturally chronologically sequential, thus providing an effective way to model the inherent temporal dynamics of EEG signals. This approach is more suitable for processing EEG data as it is able to capture and utilize changes in bioelectrical activity over time, and thus may perform better than the standard Transformer model in practical applications. Retentive Network exhibits a favorable disposition towards temporal information, boasts robust global modeling capabilities for nonlinear features, and demonstrates commendable performance. However, when applied directly to denoise EEG signals, a challenge arises. This stems from the fact that EEG signals possess temporal characteristics and encompass global nonlinear features. Unfortunately, using Retentive Network directly for EEG denoising is unfeasible due to a misalignment in the dimensional requirements. Retentive Network, designed for two-dimensional input, conflicts with the one-dimensional nature of EEG signals. Unlike the approach in [25], reshaping a 1D signal into a 2D format results in a fixed sum of input dimensions after reshaping, compromising subsequent network feature extraction. To address this issue, we propose a signal embedding method capable of transforming the signal into a sequence of arbitrary length and embedding dimensions, enhancing network flexibility. Additionally, while EEGdenoiseNet [42] introduces a standard deep learning EEG dataset, expediting the development of EEG denoising methods, the dataset remains unprocessed, lacking sample pairs. This necessitates mixing various noise types (muscle artifacts and eye artifacts) during preprocessing. Divergent data preprocessing approaches may yield disparate network results, impeding the comparison of methodologies. To mitigate this, we curated an open-source dataset using the huggingface datasets library [17] from preprocessed data. [1] [2]

The main contributions of this paper can be summarized as follows:

(1) **Proposal of Signal Embedding.** In order to efficiently extract features from EEG signals, we introduce a method called signal embedding, which adds an embedding dimension to EEG signals. This method achieves the enhancement of feature information of EEG signals through the embedding strategy. The introduction of this method not only enhances the adaptability of the network, but also positively impacts the overall improvement of the system performance.

(2) **Introducing Retentive Network for EEG Signal Denoising.** For the first time, we introduce the Retentive Network architecture into the field of EEG signal denoising to address the temporal nature of EEG signals, providing a new way to explore the intersection of EEG signals and natural language processing and expanding the scope of related research. It provides a new way to explore the intersection of EEG signal and natural language processing, and also expands the scope of related research.The introduction of Retentive Network allows us to take full advantage of its time-series information friendly and global modeling, thus realizing a new denoising method for EEG signal.

(3) **Provide open source datasets.** When examining the standard deep learning EEG dataset provided by EEGDenoiseNet, we observe that the raw nature of the dataset and the absence of pairs of training samples hinder the comparison of different methods. To address this limitation, we create an open-source dataset using preprocessed data. This not only eliminates challenges related to noise and ensures data consistency but also facilitates the exploration of deep learning-based denoising methods for EEG signals.

2. Related work

2.1. Traditional Methods

Regression methods typically depend on exogenous reference channels such as EOG, EMG, or ECG [40] to model and eliminate associated noise [16]. The efficacy of these methods is contingent upon the quality of the reference channels. Poor quality or absence of these channels significantly undermines the performance of the regression model. Furthermore, many regression methods, especially linear regression, presuppose a linear relationship between the data. However, the association between EEG signals and noise is often nonlinear, particularly when noise sources are complex, such as muscle activity or eye movement. This complexity necessitates more sophisticated nonlinear models for effective noise removal.

The Wavelet Transform method [38] is used to convert time-domain signals into time and frequency domains. This method is favored over the Fourier Transform due to its better tunable time-frequency tradeoff and its capability to analyze non-stationary signals. It operates by mapping the signal into the wavelet domain, where distinct properties of wavelet coefficients generated by signal and noise at various scales are utilized [5]. The primary goal is to eliminate noise-generated wavelet coefficients while preserving those from the actual signals. However, this method may lack sensitivity to the specific time-frequency characteristics of the noise in complex EEG signals.

[1] https://github.com/woldier/EEGDiR
[2] https://huggingface.co/datasets/woldier/eeg_denoise_dataset

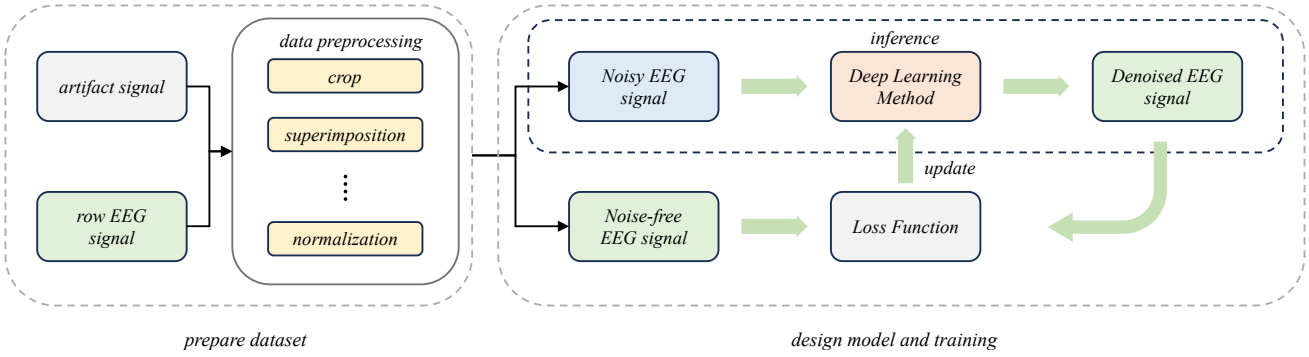


Figure 1: The diagram illustrates the training procedure of the deep learning method. It includes the Data Processing stage, where raw EEG data and various artifacts undergo preprocessing to create a suitable dataset for deep learning. The dataset comprises sample pairs, namely Noisy EEG signal and Noise-free EEG signal, representing EEG signals with and without noise, respectively. Throughout the training process, the Noisy EEG signal serves as the network input. The network, in turn, produces the Denoised EEG signal, which is utilized as the input for the subsequent training steps. The Loss Function calculates the disparity between the Denoised EEG signal and the Noise-free EEG signal, facilitating the optimization of the network. The black dashed box delineates the inference phase of the network, omitting the optimization component. This inference stage can be likened to the end-to-end output where the Noisy EEG signal is input into the network, yielding the Denoised EEG signal as the output.

To overcome these challenges, deep learning methods have become a promising alternative, thanks to their robust feature learning and representation capabilities, achieving significant successes in EEG denoising tasks.

2.2. Deep Learning Methods

In recent years, deep learning has significantly advanced in fields such as natural language processing [36] [32] [23] and computer vision [9] [27]. Notably, its application to signal processing has demonstrated remarkable efficacy in signal denoising [29] [40] [42] [25] [37] [40] [2].

To address ocular artifacts, and muscle artifacts in EEG signals, Yang et al. [40] introduced DLN, a straightforward and efficient fully-connected neural network that surpasses traditional EEG denoising methods in processing efficiency, requiring no human intervention. Sun et al. [29] proposed a one-dimensional residual CNN (1D-ResCNN) model based on convolutional neural network CNN, showcasing superior denoising performance compared to DLN by adeptly employing various convolutional kernel sizes (1×3, 1×5, 1×7) and integrating a residual layer [9]. Zhang et al. [42] presented a comprehensive EEG dataset, reducing the dataset collection challenge, and outlined four fundamental network models utilizing fully connected neural networks (FCNN), convolutional neural networks (CNN), and recurrent neural networks (RNN) for the removal of ocular and muscle artifacts. Additionally, Pu et al. [25] introduced EEGDnet, leveraging the Transformer model, which outperforms prior networks in both nonlocal and local self-similarity within the model architecture. On Zhang et al.'s benchmark EEG dataset, EEGDnet surpasses previous networks in eliminating ocular artifacts, and muscle artifacts. This body of work provides valuable references and innovations to propel the advancement of EEG deep learning.

The temporal information in EEG signals is inherently long-term and characterized by numerous temporal correlations. Traditional methods often encounter difficulties in handling extensive time-series data. However, the integration of deep learning methods proves advantageous in accommodating the temporal intricacies of EEG signals. As EEG signals emanate from the entire brain, comprehensive global modeling becomes imperative for enhanced comprehension and processing. Despite the simplicity and efficiency of the DLN model, its fully-connected structure may exhibit limitations when dealing with prolonged time-series information and global modeling. While the 1D-ResCNN model surpasses DLN in denoising performance, its dependence on a single convolutional kernel size might present constraints. The model could face challenges in addressing multi-scale features and intricate temporal information. In the case of EEGDnet, its incorporation of the Transformer model demonstrates superior architectural performance concerning nonlocal and local self-similarity. However, given the diverse frequencies present in EEG signals, effective feature capture across different scales becomes crucial.

2.3. Dataset Preparation

In order to prepare data for deep learning, the raw EEG data and various artifacts undergo preprocessing, as illustrated in the left half of Fig. 1. This includes cropping, signal stacking, normalization, and other operations to create a dataset suitable for deep learning. The creation of this dataset is foundational to our study, as it provides immediate access to preprocessed EEG data for training deep learning models. However, data preprocessing is laborious and time-consuming, underscoring the importance of providing an out-of-the-box (i.e., no data preprocessing required) dataset readily available to researchers.

During the training phase of the deep learning method, employ the framework depicted in the right half of Fig. 1,

where noisy EEG signals and noiseless EEG signals form pairs for training. The objective of training is to model the noisy EEG signals to produce outputs closely resembling the noise-free state by learning the network's weighting parameters. Initially, the noisy EEG signal is inputted into the network to generate the corresponding denoised EEG signal. The discrepancy between this denoised EEG signal and the actual noise-free EEG signal is quantified as a loss, calculated by the loss function.

The black dashed box in the figure delineates the network's inference process. During inference, the network directly takes the noisy EEG signal as input and outputs the denoised EEG signal, bypassing the optimization of the weighting parameters. This end-to-end inference capability enables the network to denoise new and unknown EEG signals in practical applications. The primary objective of the entire training process is noise suppression by optimizing the network parameters to extract true signal features amidst noise interference. This design facilitates the network to learn more effective representations, enhancing its denoising performance and providing robust support for real EEG signal processing.

Given the shared structure between the training and inference processes, and provide an out-of-the-box dataset readily available to researchers. Researchers can concentrate their efforts on investigating the network's architecture. This facilitates a more profound exploration of the application of deep learning in EEG signal processing.

3. Method

3.1. Overall structure of the EEGDiR network

In this paper, we present EEGDiR, a novel network model tailored for EEG signal denoising. Our model incorporates Retentive Network into the realm of EEG signal denoising, introducing innovative perspectives to signal processing tasks. The overall network structure (see Fig. 2(a)) involves processing the noisy signal y through signal embedding, elaborated later, to augment the embedding dimension. Subsequently, the noise signal y undergoes processing via stacked DiR Blocks at multiple levels, with the final output linearly projected to match the input dimension. EEGDiR operates as an end-to-end model, taking a noisy input signal y and generating the corresponding noiseless signal $\hat{x}$. Fig. 2(b) depicts the structure of the DiR Block, which begins with pre-Norm, followed by multi-scale Retention and a Residual Connection [9]. By using skip connections, residual learning enables the network to learn residual mappings, which can help mitigate the degradation problem associated with increasing network depth. These residual connections provide shortcuts for gradient flow, making it easier for the network to optimize the deeper layers and improve overall performance. The output of the residual join undergoes pre-Norm once more before serving as the input for the Fully Functional Network (FFN). The term "pre-Norm" refers to Layer Normalization, employed for normalization before each submodule. Layer Normalization (LN) [3] is another normalization technique that serves as an alternative to traditional Batch Normalization (BN) [12]. While BN can face challenges in performance when dealing with small batch sizes, Layer Normalization aims to address these issues by normalizing at the layer level. It is crucial to note that an FFN typically includes a fully-connected module with a hidden layer that doubles the hidden dimension. The FFNs output layer reduces the hidden dimension to align with the input. The dimensions of input and output vectors remain constant across DiR and its submodules. Figure 2(c) illustrates the signal embedding structure, the input sequence is segmented into new sequences of length $l_s//patch\_size$ based on the patch size. The initial hidden dimension of these sequences is equal to the patch size after reshaping. Through linear projection, the hidden dimension of the projected sequences matches the final hidden dimension. Given that EEG signals may encompass multiple frequencies and require effective feature capture at different scales, signal embedding is introduced to intelligently handle temporal information at varying scales. It enhances context preservation and temporal relationship retention by merging consecutive samples into a patch.

3.2. Muti-Scale Retention

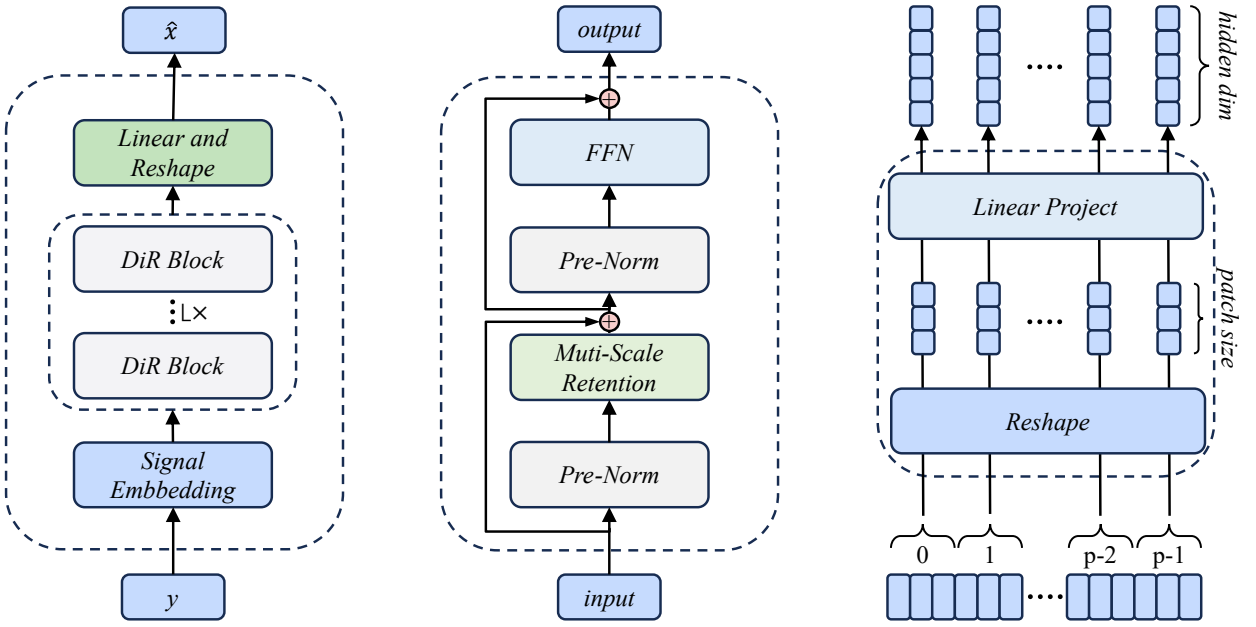
In this study, we explore the integration of the Retentive Network model, originally designed for Natural Language Processing, into the realm of EEG signal denoising. It is important to emphasize that while the Retentive Network has demonstrated remarkable success in natural language processing, our investigation centers on its applicability to EEG denoising. The Retentive Network is comprised of L identical modules arranged in a stacked fashion, featuring residual connectivity and pre-LayerNorm akin to the Transformer architecture. Each Retnet module comprises two sub-modules: the Multi-scale Retention (MSR) module and the Feedforward Network (FFN) module. For a given input sequence $s = s_1 s_2 s_3 \dots s_{l_s}$ (where l_s denotes the length of the sequence), the input vector is initially transformed to $X^0 = \left[x_1, x_2, \dots, x_{l_x}\right] \in \mathbb{R}^{l_x \times d_{model}}$, where d_{model} is the hidden dimension. Subsequently, the Retnet Block can be computed for each layer, denoted as $X^l = Retnet(X^{l-1}), l \in [1, L]$. The Simple Retention layer is defined as follows [30]:

$$Q = (XW_Q) \odot \Theta, \quad K = (XW_K) \odot \bar{\Theta}, \quad V = XW_V \tag{1}$$

$$\Theta_n = e^{in\theta}, \quad D_{mn} = \begin{cases} \gamma^{n-m}, & n \geq m \\ 0, & n < m \end{cases} \tag{2}$$

$$Retention(X) = (QK^T \odot D)V \tag{3}$$

where $\bar{\Theta}$ is the complex conjugate of Θ [31] [28], $W_Q, W_K, W_V \in \mathbb{R}^{l_x \times l_x}$, $D \in \mathbb{R}^{l_x \times l_x}$ combines causal masking and exponential decay of relative distances into one matrix.



(a) EGGDiR (b) DiR Block (c) Signal Embbedding

Figure 2: (a) illustrates the architecture of the EEGDiR network. This network generates hidden dimensions through Signal Embedding and obtains the output via linear projection and transformation following multi-level DiR Block processing. EEGDiR operates as an end-to-end model, taking a noisy signal as input and producing a noise-free signal, denoted as $\hat{x}$. The DiR Block, depicted in (b), comprises Pre-Norm, Multi-scale Retention, and Residual Connection, with Pre-Norm utilizing Layer Normalization. The Signal Embedding structure, outlined in (c), involves segmenting the input sequence into new sequences based on the patch size. The hidden dimension after reshaping aligns with the patch size, and after linear projection, it matches the final hidden dimension.

To achieve a multichannel-like effect, input sequences can be projected to lower dimensions d times, akin to the multiple-header mechanism in Transformer. This method is employed in each Retention layer with multiple headers $h = \frac{d_{model}}{d}$, where d represents the length of the sequences in each header. Each header utilizes distinct $W_Q, W_K, W_V \in \mathbb{R}^{d\times d}$, constituting the Muti-Scale Retention (MSR) block. Different γ hyperparameters are assigned to various heads in MSR, maintaining simplicity with the same γ across different layers. Additionally, a swish [13] gate is incorporated to enhance nonlinear features in each layer. Given an input X, the mathematical representation of the Muti-Scale Retention is provided as follows:

$$\gamma = 1 - 2^{-5-arrange(0,h)} \in R^h \tag{4}$$

$$head_i = Retention(X, \gamma_i) \tag{5}$$

$$Y = GroupNorm_n(Concat(head_1, \dots, head_h)) \tag{6}$$

$$MSR(X) = (swish(XW_G) \odot Y)W_o \tag{7}$$

Here, $W_G, W_O \in \mathbb{R}^{d_{model}\times d_{model}}$ are learnable parameters, and GroupNorm normalizes the output of each head. Group Normalization (GN) [39] is an alternative to traditional Batch Normalization (BN) [12] that addresses the issue of poor performance with small batch sizes. By normalizing within groups, GN provides a more robust estimation of statistics and helps mitigate the negative impact of small batch sizes or imbalanced data distribution.

3.3. Signal Embedding

In our extended investigation, it was observed that when the input sequence $s = s_1 s_2 s_3 \dots s_{l_s} \in \mathbb{R}^{l_s\times 1}$ is relatively short, direct embedding is feasible. However, in general scenarios, where the time-series information of the signal is usually lengthy, direct embedding incurs high computational complexity, hindering effective network training. To address this, we propose the introduction of a concept termed "patch", involving the amalgamation of a series of consecutive samples into a single input feature. This concept is inspired by speech signal processing, where a solitary sample may inadequately represent the current word, while a segment of samples offers more semantic expressiveness. It is noteworthy that EEG signals frequently encompass extensive temporal information, and signal embedding intelligently captures this temporal data. By grouping consecutive samples into patches, the network better retains context

and temporal relationships in the signal, enhancing denoising effectiveness. This approach aligns with speech signal processing, where context is pivotal for accurate speech comprehension. Consequently, this paper introduces signal embedding, a more efficient process tailored to the characteristics of EEG signals.

The complete signal embedding process is illustrated in Figure 2(c). Assuming a given patch size, the original sequence is divided accordingly, reducing the sequence length to $l_s//patch\_size$. Subsequently, each patch undergoes reshaping and linear projection to attain the desired hidden dimension. This process not only mitigates computational complexity but also preserves timing information more effectively. It's crucial to note that the signal embedding used here does not employ positional encoding. This is because the Retention mechanism already incorporates positional encoding considerations, obviating the need for additional positional encoding. Mathematically, it can be expressed as follows:

$$s' = Pathchfiy(s) = s'_1 s'_2 s'_3 \dots s'_{\frac{l_s}{p}} \in \mathbb{R}^{\frac{l_s}{p} \times p} \tag{8}$$

$$X^0 = Embedding(s'; \omega) = [x_1, x_2, \dots, x_{|x|}] \in \mathbb{R}^{\frac{l_s}{p} \times d_{model}} \tag{9}$$

$$X^0 = SignalEmbedding(s; \omega) \tag{10}$$

Equation (8) delineates the patching process, wherein the original sequence s is segmented into smaller sequences through patching, with each patch serving as an input feature. This operation not only preserves the feature information of the input (EEG signal) but also significantly truncates the length of the input sequence, thereby diminishing the computational complexity of subsequent operations. Here, s' denotes the sequence post the patch operation. In Equation (9), we illustrate the feature embedding, signifying that after the patch sequence s', linear projection of the feature size results in the generation of the larger feature size X^0. This dispersion of signal features is conducive to the subsequent network's extraction of diverse features. This process allows the signal features to be spread out, facilitating the network in extracting distinct feature types. Here, ω denotes the learnable parameter, and X^0 remains consistent with the preceding section. If we conceptualize patching and embedding as an end-to-end operation, it can be expressed as (10). In other words, the input s' can be derived from the signal embedding module to yield X^0, with ω serving as the learnable parameter, akin to (9).

4. Experiments and results

4.1. Preliminary

Signals disturbed by noise are acquired through the linear combination of the electrooculogram (EOG) or electromyogram (EMG) with the pristine electroencephalogram (EEG). This procedure can be mathematically represented as Equation (11) [42]. The mixed EEG noise signal is denoted as $y \in \mathbb{R}^{l_y}$, where y represents the sequence length. The noise-free EEG signal, denoted as $x \in \mathbb{R}^{l_x}$, serves as the ground truth, and $n \in \mathbb{R}^{l_n}$ represents ocular artifacts or muscle artifacts. It is important to note that the lengths of each sequence l_x, l_y, l_n are equal. To control the noise level during mixing, we introduce the hyperparameter λ, regulating the signal-to-noise ratio (SNR) of the noisy signal. Adjustment of different λ values enables effective control of SNR magnitude to adapt to various noise environments. The SNR is calculated using Equation (12), while λ is determined by Equation (13), where $RMS(\cdot)$ denotes the root mean square of the sample, $RMS(x)$ is the root mean square of the noiseless EEG signal x, and RMS $RMS(\lambda \cdot n)$ is the root mean square of the mixed noise $\lambda \cdot n$. These formulas provide flexible adjustment of the signal-to-noise balance to meet diverse signal quality requirements in specific application scenarios.

$$y = x + \lambda \cdot n \tag{11}$$

$$SNR = 10log(\frac{RMS(x)}{RMS(\lambda \cdot n)}) \tag{12}$$

$$\lambda = \frac{RMS(x)}{RMS(n) \cdot (\frac{SNR}{10})^{10}} \tag{13}$$

In the context of deep learning applied to EEG signal denoising, the denoising process can be conceptualized as a nonlinear mapping function. This function, denoted as $\hat{x} = F(y; \theta)$, maps the EEG signal y with noise to the corresponding noise-free signal $\hat{x}$. Here, $F(\cdot)$ represents the nonlinear mapping function, our neural network model, and θ is the model's learnable parameter. To facilitate better parameter learning, we employ the mean square error (MSE) as the loss function. The MSE is defined by calculating the squared difference between the predicted value $\hat{x}_i$ and the true value x_i for each sample point i of the signal, summing these differences, and dividing by the number of samples n. Mathematically, this is expressed as Equation (14).

$$\mathbb{L}(x, \hat{x}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x}_i)^2 \tag{14}$$

4.2. Experiments Detail

4.2.1. Datasets

To assess the denoising efficacy of the proposed EEGDiR model, we utilized the EEGDenoiseNet dataset [5], a widely adopted dataset in deep learning for EEG signal denoising, for both training and testing. The dataset encompasses various signal categories, including 4515 pristine EEG signals, 3400 ocular artifacts, and 5598 muscle artifacts. Each sample has a sampling time of 2 seconds at a rate of 256 samples per second. Pure EEG signals are denoted as x

in Equation (11), while ocular artifacts or muscle artifacts are denoted as n in Equation (11).

For signals contaminated with ocular artifacts, 3400 samples were randomly chosen from pure EEG signals and all 3400 ocular artifact signals. Subsequently, the training and test sets were constructed in an 8:2 ratio, respectively. At specified signal-to-noise ratio (SNR) levels (-7 dB to 2 dB), pure EEG signals were linearly combined with ocular artifacts to generate ocular artifact-contaminated signals y. Notably, the parameter λ for superimposing eye movement artifacts in Equation (13) was directly calculated based on the given SNR value to obtain y. This dataset is referred to as the EOG dataset.

The signals contaminated with muscle artifacts were derived from pure EEG signals, and all 4515 samples were utilized along with the 5598 samples from the EOG artifact signals. To maintain consistency in the number of samples from pure EEG signals and EMG artifact signals, some samples from the pure EEG signal were reused. The resulting dataset was partitioned into training and test sets in an 8:2 ratio. Similarly, using specified SNR levels, pure EEG signals were randomly combined with EMG artifacts to generate EMG artifact-contaminated signals y. This dataset is denoted as the EMG dataset.

To validate our proposed network thoroughly, we utilize a semi-simulated EEG dataset (SS2016) [15], contaminated with ocular artifacts, alongside clean EEG signals. This dataset is notable for containing both clean EEG signals and their contaminated counterparts. Collected from 54 participants during a closed-eye experiment, the signals are devoid of ocular artifacts. Electroencephalogram (EEG) signals were recorded from 19 electrodes placed according to the International 10-20 system, with each experiment lasting approximately 30 seconds and sampled at 200 SPS[20]. The dataset, comprising pure and contaminated EEG signals, comprises 54 matrices, each corresponding to one participant. Each matrix consists of 19 channels, each representing the signal recorded by an electrode. The number of sampling points for channel signals ranged from 5600 to 8400. Data synthesis involved overlaying EOG data onto EEG signals, resulting in recordings of about 30 seconds in duration but with varying sample counts. For ease of processing, we segmented the data without overlap into small segments of 512. Post-segmentation, we obtained 11495 samples, with Figure 3 displaying segments 1-4 for participant 1.

Subsequently, we calculated the SNR values of each sample, rounding them, and depicted the results in Figure 4. Upon observation, we noted SNR values ranging between -10 to 20, with a relatively large number of samples falling within the -5 to 15 range, and fewer samples at other SNR values, posing challenges for network learning. To address this, we followed the data processing approach of the EEG-Denoise dataset, separating the noise (n) from the pure signal (x) and contaminated signal (y). We then multiplied n by different λ values to attain noise levels ranging from -7dB to 2dB. However, we observed that when the signal-to-noise ratio of signal pairs in the original dataset was high, such

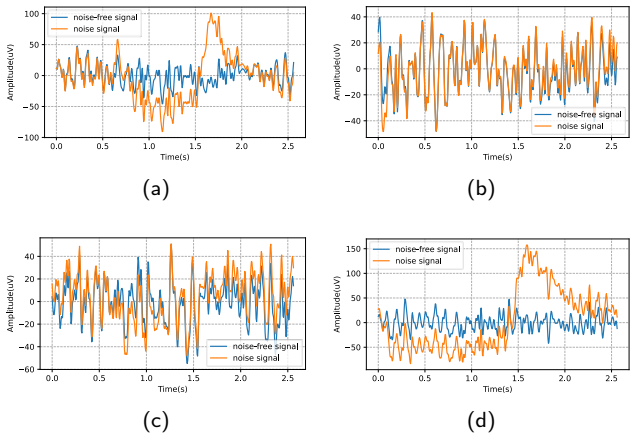


Figure 3: The recorded signals from the first electrode of Participant 1 in SS2016 are displayed in sequence as segments 1 through 4, denoted as (a), (b), (c), and (d) respectively. Each segment comprises 512 samples, with a sampling rate of 200 SPS.

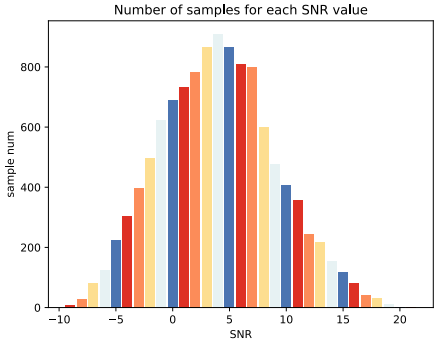


Figure 4: The SNR distribution of the SS2016 dataset after signal segmentation is depicted in the figure. It's important to highlight that the SNR values are rounded up for statistical simplicity. The horizontal axis represents various SNR levels ranging from -10 to 20, while the vertical axis indicates the number of segments corresponding to each SNR level.

as Figure 3b and Figure 3b, excessively large λ values were required to achieve low signal-to-noise levels, resulting in signal amplification beyond realistic levels. Data like this not correspond to actual noisy signals and were deemed unsuitable for training samples. Thus, we discarded pairs of samples with SNR higher than 5dB from the original dataset, resulting in a final sample count of 6716. We assigned different λ values to n corresponding to each signal, obtaining noise levels from -7dB to 2dB. This dataset is denoted as the SS2016 EOG dataset. The resulting dataset was partitioned into training and test sets in an 8:2 ratio.

In order to facilitate the learning procedure, we normalized the input contaminated EEG segment and the ground-truth EEG segment by dividing the standard deviation of contaminated EEG segment according to Equation (15), where σ_y is the standard deviation of y (artifact contaminated

signal).

$$\hat{x} = \frac{x}{\sigma_y}, \hat{y} = \frac{y}{\sigma_y} \tag{15}$$

However, it is important to note that the EEGDenoiseNet dataset solely provides raw data, and the data provided by SS2016 is not segmented, necessitating researchers to conduct their own processing. This procedure is relatively intricate, posing inconvenience for the exploration of EEG signal denoising through deep learning methodologies. To address this challenge, this paper undertakes the preprocessing of the dataset and subsequently shares the processed dataset as an open-source resource, now accessible on the Hugging Face Hub. The primary objective of this endeavor is to facilitate researchers access to and utilization of processed data, enabling them to concentrate more on the investigation of EEG deep learning denoising methods without being encumbered by the intricacies of data processing. By providing this dataset as an open-source entity, our aim is to stimulate increased research in EEG signal processing and offer a more streamlined resource for the academic community.

4.2.2. Train details

In this investigation, we opted to implement the EEGDiR model utilizing the PyTorch deep learning framework, renowned for its widespread usage and adaptability. To enhance the training efficiency of the network, we employed the AdamW [43] optimizer, a proficient choice for managing large-scale deep learning models. The learning rate was set to $5e{-}4$, and the betas parameter ranged from $(0.5, 0.9)$, with meticulous adjustment to optimize the network training process.

Throughout the training phase, the network underwent 5000 epochs to ensure comprehensive learning of dataset features. Furthermore, we configured the batch size to 1000, a standard choice that balances memory utilization and training efficacy. Notably, for accelerated training, we harnessed the computational capabilities of an NVIDIA GeForce RTX 4090 Graphics Processing Unit (GPU). Leveraging the parallel computing power of GPUs significantly augmented the speed of deep learning model training, facilitating rapid experimentation and tuning for researchers.

4.3. Results

4.3.1. Comparing method

We conducted comparative experiments to assess the efficacy of our proposed EEGDiR model against established state-of-the-art deep learning EEG denoising networks, encompassing the following models:

(1) Simple Convolutional Neural Networks (SCNN) [42]:
Network Structure: Four 1D convolutional layers with 1×3 convolutional kernels, a 1-step size, and 64 channels.
Interlayer Structure: Batch Normalization and ReLU activation functions follow each convolutional layer.
Output: Features linearly projected through fully connected layers to match input dimensions.

(2) One-dimensional Residual Convolutional Neural Networks (1D-ResCNN) [29]:
Network Structure: Utilizes three distinct convolutional kernels (1×3, 1×5, 1×7) ResBlocks for parallel feature extraction.
ResBlock Structure: Each ResBlock comprises four 1D convolutional layers, with every two forming a residual block.
Interlayer Structure: Activation through Batch Normalization and ReLU functions for each residual block.
Output: Concatenation of the three ResBlocks' outputs, linearly projected through fully connected layers to maintain input dimensions.

(3) Long Short Term Memory (LSTM) Network:
A Long Short-Term Memory (LSTM) network, adapted from [11], is considered the benchmark for recurrent neural networks (RNNs) [41]. LSTM is capable of learning long-term dependencies, which aids in distinguishing long-term features in noise and EEG signals. Each EEG sample is sequentially fed into LSTM cells, and the output is derived from the state of each cell through a fully-connected network.

(4) EEG Denoise Network (EEGDnet) [25]:
Network Structure: Incorporates a Transformer infrastructure with four Transformer layers, featuring an Attention module and a FeedForward Network (FFN) module in each layer.
Module Structure: Layer Normalization applied to the input of each module.
Output: Linear projection to maintain input dimensions.

These benchmark networks represent diverse EEG denoising methodologies, serving as benchmarks to validate the superiority of our proposed EEGDiR model in denoising performance. These comparisons aim to offer readers a comprehensive understanding of the EEGDiR model's performance.

4.3.2. Evaluation measures

We assess the denoising outcomes using three methods: Relative Root Mean Squared Error in the temporal domain ($RRMSE_t$), $RRMSE$ in the spectral domain ($RRMSE_s$), and the correlation coefficient (CC) [42]. This selection is grounded in a profound understanding of EEG signal characteristics. Firstly, considering the temporal significance of the EEG signal, we employ $RRMSE_t$ to quantify the relative error between the denoised and original signals in the time domain. This method exhibits sensitivity to denoising techniques preserving temporal information. Secondly, as EEG signals encapsulate rich spectral information with research often focusing on specific frequency ranges, $RRMSE_s$ is employed to ensure the preservation of features in the frequency domain. Lastly, acknowledging the synergistic activities between different brain regions in EEG signals, CC serves as an evaluation metric. CC reflects the linear relationship between the denoised and original signals, crucial for maintaining vital information about interregional correlation. The combined use of these methods facilitates a

comprehensive evaluation of denoising performance across time and frequency domains, considering their adaptability to EEG signal characteristics. The mathematical expressions for $RRMSE_t$, $RRMSE_s$, and CC are represented in Equation (16), (17), and (18) respectively, where $RMS(\cdot)$ denotes root mean square, $PSD(\cdot)$ denotes power spectral density, and $Cov(\cdot)$, $Var(\cdot)$ represent covariance and variance, respectively.

$$RRMSE_t = \frac{RMS(F(y) - x)}{RMS(x)} = \frac{RMS(\hat{x} - x)}{RMS(x)} \quad (16)$$

$$RRMSE_s = \frac{RMS(PSD(F(y)) - PSD(x))}{RMS(PSD(x))} = \frac{Cov(\hat{x}, x)}{Var(\hat{x})Var(x)} \quad (17)$$

$$CC = \frac{Cov(F(y), x)}{\sqrt{Var(F(y))Var(x)}} = \frac{Cov(\hat{x}, x)}{\sqrt{Var(\hat{x})Var(x)}} \quad (18)$$

4.3.3. Ablation study

In this study, we pioneer the application of the Retnet network to EEG signal denoising and introduce the innovative concept of signal embedding. To assess the impact of each hyperparameter on denoising performance, we conduct ablation experiments, marking the inaugural exploration of these techniques. Our initial focus is on the influence of patch size and hidden dimension, investigating their effects on network performance. Table 1 presents quantized results for the network applied to EOG, EMG and SS2016 EOG datasets under various configurations of patch size and hidden dimension hyperparameters. Notably, we observe an enhancement in denoising performance with decreasing patch size while maintaining a consistent hidden dimension. This improvement is attributed to the impact of patched sequence length on the network's feature extraction capability—smaller patch sizes result in larger mini sequence lengths, preserving more information and thereby improving denoising effectiveness. However, a cautious approach is essential as blindly reducing patch size escalates computational complexity due to increased sequence length. Thus, a delicate balance between denoising performance and computational efficiency is imperative. Furthermore, maintaining the same patch size, an increased hidden dimension corresponds to improved denoising performance, aligning with the intuitive understanding that higher dimensionality facilitates enhanced feature extraction.

Subsequently, we assess the impact of varying the number of block layers L on network performance. Table 2 displays the denoising quantization results for the model applied to EOG, EMG and SS2016 EOG datasets, with fixed parameters patch size 16, hidden dimension 512, and heads 8. The observations indicate a gradual enhancement in denoising performance with an increasing number of layers. This improvement is ascribed to the benefits of residual connections, whereby a higher number of layers does not lead to overfitting. The increased network depth contributes to superior feature extraction capabilities.

Following the ablation study, optimal performance is achieved when employing a patch size of 16, hidden dimension of 512, 8 heads, and 4 layers. Consequently, this configuration is chosen as the benchmark for subsequent comparisons with other networks.

4.3.4. Denoising effect of each method at all noise levels

Table 3 illustrates the denoising efficacy of various methods on EOG, EMG and SS2016 EOG datasets. The outcomes in this table lead to the following:

(1) Due to its relatively simplistic structure comprising only four convolutional layers and lacking residual connections, SCNN exhibits suboptimal denoising effects, potentially prone to overfitting.
(2) Featuring a more intricate architecture incorporating diverse convolutional kernels for multi-scale feature extraction and alleviating overfitting through the introduction of residual connections, 1D-ResCNN surpasses SCNN, significantly enhancing denoising outcomes. Thanks to the temporal information added to the input by the LSTM structure, LSTM exhibits better denoising results than 1D-ResCNN.
(3) Leveraging the transformer architecture, EEGDnet excels in denoising, benefitting from the global modeling prowess of the attention mechanism, complemented by residual connections and layer normalization. This results in substantial denoising improvements compared to SCNN, 1D-ResCNN and LSTM.
(4) Capitalizing on the Retentive Network, EEGDiR achieves superior denoising performance by comprehensively understanding input temporal information and exhibiting robust global modeling capabilities. The incorporation of residuals and multiple normalizations (layer norm, group norm) further distinguishes EEGDiR, outperforming other networks. Moreover, guided by our proposed signal embedding, EEGDiR intelligently processes temporal information. This strategy adeptly captures the contextual and temporal relationships within EEG signals, aligning with their prolonged temporal characteristics. The signal embedding strategy contributes to optimized denoising performance, reinforcing EEGDiR's exceptional superiority over alternative networks.

4.3.5. Denoising effect of each method at different noise levels

In the subsequent section, we present the quantitative benchmarking results ($RRMSE_t$, $RRMSE_s$, CC) of diverse methods across varying SNR levels in the test set. Figures 5, Figures 6 and Figures 7 showcase the test outcomes on the EOG, EMG and SS2016 EOG test dataset, pivotal for evaluating the denoising efficacy of the methods.

Table 1
The effect of patch size and hidden dim on the noise reduction performance of EEGDiR. Note mini sequence length must be $l_s//patchsize$, where $l_s = 512$, with 8 Heads and 4 DiRBlock layers.

Patch size	Mini seq. length	Hidden dim	EOG dataset			EMG dataset			SS2016 EOG dataset		
			$RRMSE_t$	$RRMSE_s$	CC	$RRMSE_t$	$RRMSE_s$	CC	$RRMSE_t$	$RRMSE_s$	CC
32	16	512	0.339	0.367	0.928	0.556	0.561	0.793	0.357	0.392	0.932
32	16	256	0.353	0.377	0.911	0.569	0.575	0.791	0.397	0.466	0.916
32	16	64	0.382	0.371	0.909	0.572	0.577	0.790	0.420	0.525	0.903
16	32	512	0.327	0.361	0.932	0.532	0.501	0.807	0.315	0.362	0.948
16	32	256	0.348	0.371	0.925	0.598	0.573	0.776	0.357	0.395	0.932
16	32	64	0.374	0.378	0.912	0.654	0.593	0.701	0.401	0.443	0.913

Table 2
The effect of the layers of EEGDiR on the noise reduction performance. Note that patch size ,hidden dim and N Heads equal to 16, 512 and 8 respectively.

Pathch size	EOG dataset			EMG dataset			SS2016 EOG dataset		
	$RRMSE_t$	$RRMSE_s$	CC	$RRMSE_t$	$RRMSE_s$	CC	$RRMSE_t$	$RRMSE_s$	CC
4	0.327	0.361	0.932	0.532	0.501	0.807	0.315	0.362	0.948
3	0.356	0.380	0.925	0.578	0.583	0.789	0.381	0.447	0.922
2	0.372	0.383	0.917	0.591	0.596	0.781	0.396	0.461	0.917
1	0.394	0.429	0.908	0.613	0.594	0.766	0.411	0.478	0.911

(1) Primarily, the performance of all methods exhibits a decline as the SNR level decreases. This negative correlation arises due to the gradual increase in noise level, posing a greater challenge for the methods in noise removal.

(2) Among the methods, SCNN displays the highest $RRMSE_t$ and $RRMSE_s$, along with the lowest CC. This indicates SCNN inferior denoising performance, attributed to its relatively simple network structure hindering effective input feature extraction. In contrast, the more intricate 1D-ResCNN and LSTM yields significantly improved denoising outcomes. However, compared to EEGDnet with a Transformer model and global modeling capability, there are discernible performance gaps. The EEGDiR, incorporating Rententive Network and signal embedding, achieves the lowest $RRMSE_t$ and $RRMSE_s$, coupled with the highest CC. It excels in denoising tasks across varying noise levels.

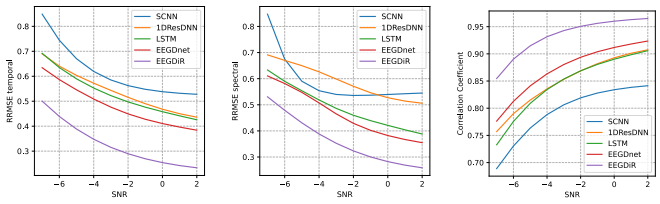
(3) Analyzing the $RRMSE_t$ results on the EOG and SS2016 EOG dataset, denoising performance improves with decreasing noise levels and increasing SNR levels across all methods. However, the performance gap between methods persists, potentially due to EOG noise being more easily removed than EMG noise. On the EMG dataset, the performance gap diminishes as noise levels decrease (SNR levels increase), particularly evident for SCNN, 1D-ResCNN, LSTM, and EEGDnet. Nevertheless, EEGDiR maintains superior denoising performance.

(4) Evaluation of the $RRMSE_s$ results on the EOG and SS2016 EOG dataset indicates weaker denoising performance for SCNN, 1D-ResCNN and LSTM, possibly due to limited global modeling capability. Conversely, EEGDnet and EEGDiR exhibit superior denoising performance owing to their robust global modeling ability. On the EMG dataset, despite decreasing differences in performance as noise levels decrease, EEGDnet and EEGDiR consistently outperform SCNN, 1D-ResCNN and LSTM.

(5) Examination of CC results on the EOG and SS2016 EOG dataset reveals improved denoising performance for all methods as noise levels decrease, with relatively

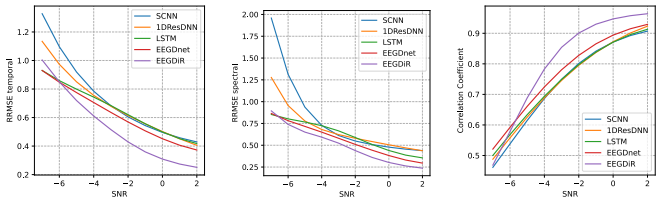
Table 3
Average performances of all SNRs (from −7 dB to 2 dB). The smaller $RRMSE_t$ and $RRMSE_s$, and the larger CC, the better denoising effect. Note that all the models are trained and tested on the same data set. The baseline of EEGDiR consists of 4 layers and 8 Heads with patch size 16 and hidden dim 512. For $RRMSE_t$, $RRMSE_s$, the lower the better. For CC, the higher the better. The best result is shown in bold.

Model	EOG dataset			EMG dataset			SS2016 EOG dataset		
	$RRMSE_t$	$RRMSE_s$	CC	$RRMSE_t$	$RRMSE_s$	CC	$RRMSE_t$	$RRMSE_s$	CC
SCNN	0.6176	0.5905	0.7938	0.7342	0.7977	0.7364	0.5893	0.6724	0.8156
1D-ResCNN	0.5409	0.5900	0.8503	0.6921	0.6848	0.7434	0.5523	0.5804	0.8552
LSTM	0.5290	0.4894	0.8449	0.6560	0.6092	0.7461	0.4823	0.5573	0.8747
EEGDnet	0.4819	0.4647	0.8725	0.6200	0.5565	0.7711	0.4594	0.5267	0.8875
EEGDiR(ours)	**0.3279**	**0.3616**	**0.9329**	**0.5322**	**0.5004**	**0.8072**	**0.3146**	**0.3613**	**0.9488**



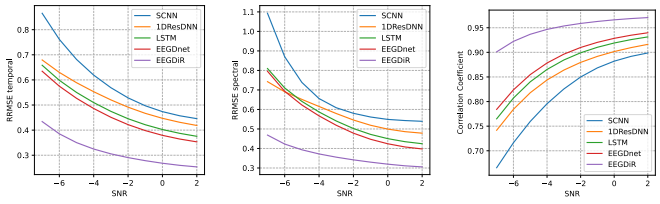
(a) $RRMSE_{temporal}$ (b) $RRMSE_{spectral}$ (c) CC

Figure 5: Performance of four deep-learning networks at different SNR levels with EOG dataset artifact removal. The smaller $RRMSE_t$ and $RRMSE_s$, and the larger Correlation Coefficient(CC), the better denoising effect. The denoising performance increases as the SNR increases.



(a) $RRMSE_{temporal}$ (b) $RRMSE_{spectral}$ (c) CC

Figure 6: Performance of four deep-learning networks at different SNR levels with EMG dataset artifact removal. The smaller $RRMSE_t$ and $RRMSE_s$, and the larger Correlation Coefficient(CC), the better denoising effect. The denoising performance increases as the SNR increases.



(a) $RRMSE_{temporal}$ (b) $RRMSE_{spectral}$ (c) CC

Figure 7: Performance of four deep-learning networks at different SNR levels with SS2016 EOG dataset artifact removal. The smaller $RRMSE_t$ and $RRMSE_s$, and the larger Correlation Coefficient(CC), the better denoising effect. The denoising performance increases as the SNR increases.

stable performance differences. In the EMG dataset, SCNN exhibits poorer performance due to the dataset's more complex noise. Conversely, the denoising performance of the remaining three networks improves as noise levels decrease, with consistent performance differences. Notably, EEGDiR maintains excellent denoising performance throughout.

Figures 8, Figures 9 and Figures 10 illustrate the ANOVA results for models evaluated on the EOG, EMG and SS2016 EOG datasets. Drawing conclusions from the provided information and ANOVA analyses, the following observations emerge:

(1) $RRMSE_t$ Rancking: The denoising performance across the four methods is observed as follows: SCNN < 1D-ResCNN < LSTM < EEGDnet < EEGDiR. ANOVA analysis indicates significant differences in $RRMSE_t$, with marked distinctions in the EOG dataset and relatively modest differences in the EMG dataset. EEGDiR significantly outperforms other methods in time-domain denoising for both EOG, EMG and SS2016 EOG datasets, followed by EEGDnet, LSTM and 1D-ResCNN, while SCNN exhibits the least efficacy.

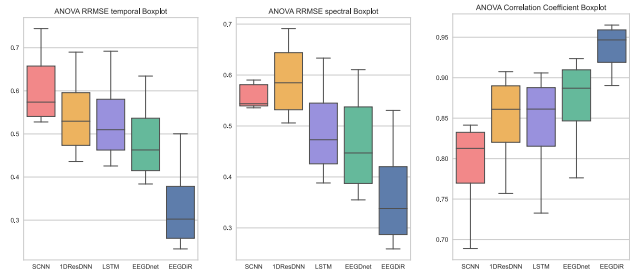
(2) $RRMSE_s$ Rancking: The denoising performance order for $RRMSE_s$ is 1D-ResCNN < SCNN < LSTM < EEGDnet < EEGDiR. The occurrence of 1D-ResCNN < SCNN is attributed to 1D-ResCNN superior extraction of features in the time domain, leading to diminished denoising performance in the spectral features. ANOVA results show a significant difference in $RRMSE_s$ among methods on the EOG and SS2016 EOG dataset, while the difference is relatively weak on the EMG dataset. EEGDiR significantly outperforms other methods in spectral denoising, followed by EEGDne, LSTM and SCNN, while 1D-ResCNN is less effective.

(3) CC Metric Ranking: The denoising performance sequence on the CC metric remains SCNN < 1D-ResCNN < LSTM < EEGDnet < EEGDiR. ANOVA analysis reveals a significant difference in CC between methods for the EOG and SS2016 EOG dataset, while the difference is relatively weak for the EMG dataset. Comparing mean values, EEGDiR excels in correlation, followed by EEGDnet and 1D-ResCNN, while SCNN exhibits poor CC performance.

In summary, the outstanding denoising performance of the EEGDiR method can be attributed to multiple factors. The Rentetive Network architecture provides enhanced global modeling capability, enabling a more accurate restoration of input timing information. The proposed signal embedding method adeptly handles the prolonged temporal information of EEG signals, capturing context and temporal relationships intelligently through the combination of successive sampling points into patches. This advantage enables EEGDiR to achieve superior denoising effects in both the time domain and spectral characteristics. Additionally, the synergy of residual connectivity and multiple normalization methods (layer norm, group norm) enhances EEGDiR denoising performance and robustness to noise. The advanced Retnet architecture, skillful embedding strategy, and enhanced network design collectively contribute to EEGDiR exceptional performance in time-domain and spectral denoising, as well as correlation.

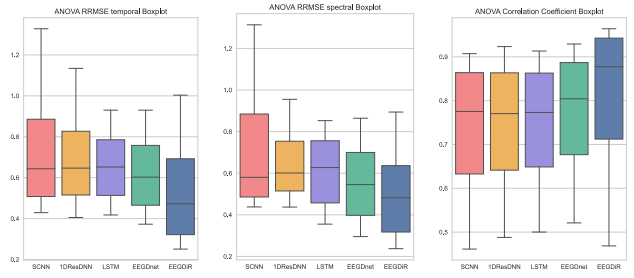
4.3.6. Visualization of denoising results for each method on EOG and EMG datasets

The visualization results depicting the impact of EOG and EMG noise on EEG signals are presented in Figure 11a, Figure 11b and Figure 11c, yielding the following observations. It is noteworthy that the dataset has undergone



(a) $RRMSE_{temporal}$ (b) $RRMSE_{spectral}$ (c) CC

Figure 8: Performance of four DL networks (SCNN, 1D-ResDNN, LSTM, EEGDnet, EEGDiR) in EOG dataset artifact removal. The smaller $RRMSE_t$ and $RRMSE_s$, and the larger Correlation Coefficient(CC), the better denoising effect. EEGDiR models robustly outperform other model for EEG denosing.



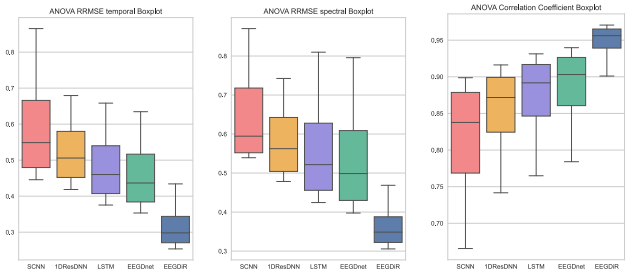
(a) $RRMSE_{temporal}$ (b) $RRMSE_{spectral}$ (c) CC

Figure 9: Performance of four DL networks (SCNN, 1D-ResDNN, LSTM, EEGDnet, EEGDiR) in EMG dataset artifact removal. The smaller $RRMSE_t$ and $RRMSE_s$, and the larger Correlation Coefficient(CC), the better denoising effect. EEGDiR models robustly outperform other model for EEG denosing.

variance normalization. When presenting the visualization results, Equation (15) is employed to scale down the results to the original data scale, enhancing the accuracy of showcasing the noise effect on EEG signals.

(1) All methods exhibit some degree of noise suppression in noisy signals, underscoring the variability among different denoising approaches. Particularly notable is the substantial difference between the denoising results of the SCNN method and the noisy signal. This divergence may be attributed to the relatively simplistic network structure of SCNN, hindering comprehensive feature extraction.

(2) The relatively complex structure and residual connectivity of LSTM and 1D-ResCNN result in an improvement in denoising compared to SCNN, emphasizing the impact of network architecture on denoising performance. Leveraging the global modeling and feature extraction capabilities facilitated by the Transformer's attention mechanism, EEGDnet outperforms SCNN, LSTM and 1D-ResCNN in denoising. This underscores the enhancement of network performance with the introduction of the attention mechanism.



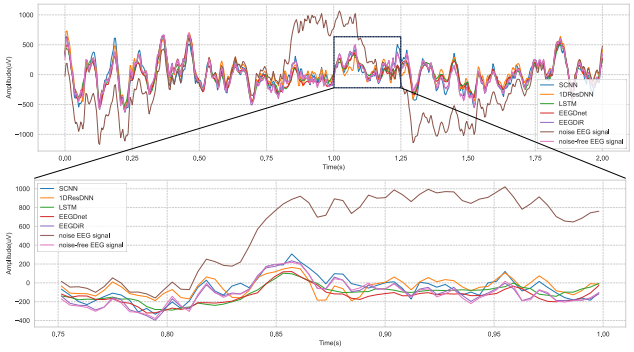
(a) $RRMSE_{temporal}$ (b) $RRMSE_{spectral}$ (c) CC

Figure 10: Performance of four DL networks (SCNN, 1D-ResDNN, LSTM, EEGDnet, EEGDiR) in SS2016 EOG dataset artifact removal. The smaller $RRMSE_{temporal}$ and $RRMSE_{spectral}$, and the larger Correlation Coefficient(CC), the better denoising effect. EEGDiR models robustly outperform other model for EEG denosing.

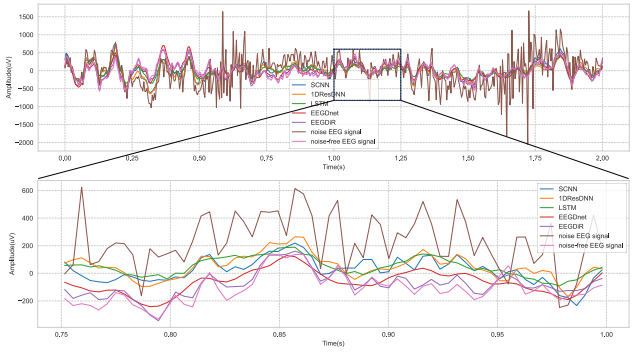
(3) Overall, the denoising effect achieved by EEGDiR closely approaches that of a noise-free signal. This notable advantage can be attributed to the synergistic effect of the Retentive Network architecture and our proposed signal embedding, tailored to match the characteristics of EEG signals. Firstly, the Retentive Network architecture enhances understanding of timing information in EEG signals through its robust global modeling capability. This enables the network to accurately capture the complex time-domain structure, thereby improving denoising performance. Secondly, our signal embedding method adeptly addresses the challenge of handling long temporal information in EEG signals. By intelligently grouping consecutive sampling points into patches, this method effectively preserves the context and temporal relationships of the signal, facilitating the network in learning and restoring features more efficiently. The sensitivity to long temporal information aligns with the characteristics of EEG signals, forming the basis for EEGDiR outstanding denoising effect. Therefore, the performance of EEGDiR in approaching noise-free signals arises not only from the superior processing of temporal information by the Retention mechanism but also from the mutually reinforcing capabilities of Retentive Network and signal embedding. This synergy enables the network to better comprehend and process the intricate structure of EEG signals.

5. Discussions

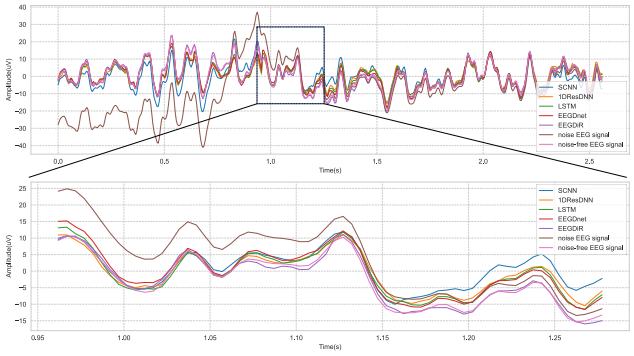
In our analysis, we compared various denoising methodologies by examining their efficacy on the EOG, EMG, and SS2016 EOG datasets as presented in Table 3. Our findings underscore the significant advancements our EEGDiR



(a) Denoising outcomes on EOG dataset.



(b) Denoising outcomes on EMG dataset.



(c) Denoising outcomes on SS2016 EOG dataset

Figure 11: Visualization of denoising outcomes for various state-of-the-art models: (a) Denoising outcomes on EOG dataset. (b) Denoising outcomes on EMG dataset. (b) Denoising outcomes on SS2016 EOG dataset. A closer inspection can be facilitated by zooming in for a more detailed view. It is important to highlight that we have restored the network output's amplitude back to the original data scale through back-normalization. The temporal domain operates at a sampling rate of 256 SPS for EOG and EMG dataset. The temporal domain operates at a sampling rate of 200 SPS for SS2016 EOG dataset. From the provided results, it is evident that the denoising results achieved by the proposed EEGDiR model in this study closely approximate the true signal.

model, incorporating the Retentive Network and signal embedding techniques, makes over other methods.

Comparative Analysis of Denoising Methods. The data reveals that simpler models like SCNN, despite their utility, fall short in more complex noise environments due to their basic structures and lack of advanced features like residual connections. Conversely, 1D-ResCNN improves upon this by utilizing diverse convolutional kernels and residual connections to mitigate overfitting, thereby enhancing denoising results. Notably, the LSTM model, which integrates temporal information directly into its structure, outperforms 1D-ResCNN by better handling the dynamics within EEG signals. EEGDnet, leveraging the global modeling capabilities of the transformer architecture, enhanced by residual connections and layer normalization, significantly surpasses both SCNN and 1D-ResCNN. The most robust performance, however, is exhibited by EEGDiR. This model not only understands complex temporal sequences better due to the Retentive Network but also optimizes the handling of these sequences through our innovative signal embedding approach. This dual strategy is particularly effective in preserving the contextual and temporal integrity of EEG signals, which is crucial given their prolonged time-series nature.

Performance Across Different Noise Levels. Our study also evaluated the performance of these methods across varying SNR levels, as depicted in Figures 5, 6, and 7. All methods demonstrated declining performance with decreasing SNR, highlighting the challenges posed by increased noise levels. However, EEGDiR consistently outperformed other methods at all noise levels, achieving the lowest RRMSE and highest correlation coefficients. This suggests that EEGDiR's architecture and signal processing strategies are well-suited to effectively reduce noise while preserving the integrity of the EEG signal.

Broader Implications. The success of the Retentive Network in this context not only paves the way for its use in EEG signal denoising but also suggests its applicability to other types of temporal signals, such as electromagnetic and seismic data. The ability of the Retentive Network to process temporal information effectively, coupled with our signal embedding technique, offers a robust framework for denoising tasks across various domains requiring detailed temporal analysis. Moreover, this framework holds significant potential for enhancing downstream EEG tasks, such as classification [19], motor imagery [7], brain recognition [18] and fatigue detection [6], suggesting broad applicability in enhancing the accuracy and effectiveness of these complex applications.

6. Conclusion

By incorporating the Retentive Network architecture and employing signal embedding for processing EEG signals, this study introduces an innovative methodology aiming to leverage Retentive Network comprehensively for EEG signal denoising. The integration of Retentive Network architecture enhances the understanding and processing of temporal information in EEG signals, while the utilization of signal embedding underscores the processing of prolonged temporal information and feature extraction. Experimental

results showcase the outstanding denoising performance of our proposed EEGDiR network on EOG, EMG and SS2016 EOG datasets. In comparison to traditional EEG denoising methods, EEGDiR demonstrates notable enhancements in temporal information processing and global modeling.

The global modeling prowess of EEGDiR, coupled with its favorable handling of temporal information, positions it as an optimal choice for processing EEG signals. The incorporation of signal embedding further refines the representation of EEG signals, preserving context and temporal relationships more effectively. The synergistic application of the Retentive Network and signal embedding strategy yields a substantial improvement in the denoising performance of the EEGDiR network.

This study holds significant implications as a guide for integrating deep learning into neuroscience, offering valuable insights to enhance the efficacy and application potential of EEG signal processing. By providing an out-of-the-box deep learning dataset, our contribution enables subsequent researchers to expedite EEG signal denoising research by eliminating the need for extensive data preprocessing. This accelerates the development of EEG signal denoising methods.

CRediT authorship contribution statement

Bin Wang: Conceptualization of this study, Methodology, Software.

References

[1] Albawi, S., Mohammed, T.A., Al-Zawi, S., 2017. Understanding of a convolutional neural network, in: 2017 international conference on engineering and technology (ICET), Ieee. pp. 1–6.

[2] Aynali, E., 2020. Noise reduction of eeg signals using autoencoders built upon gru based rnn layers .

[3] Ba, J.L., Kiros, J.R., Hinton, G.E., 2016. Layer normalization. arXiv preprint arXiv:1607.06450 .

[4] Bebis, G., Georgiopoulos, M., 1994. Feed-forward neural networks. Ieee Potentials 13, 27–31.

[5] Burger, C., Van Den Heever, D.J., 2015. Removal of eog artefacts by combining wavelet neural network and independent component analysis. Biomedical Signal Processing and Control 15, 67–79.

[6] Gao, D., Li, P., Wang, M., Liang, Y., Liu, S., Zhou, J., Wang, L., Zhang, Y., 2023. Csf-gtnet: A novel multi-dimensional feature fusion network based on convnext-gelu-bilstm for eeg-signals-enabled fatigue driving detection. IEEE Journal of Biomedical and Health Informatics .

[7] Gao, D., Yang, W., Li, P., Liu, S., Liu, T., Wang, M., Zhang, Y., 2024. A multiscale feature fusion network based on attention mechanism for motor imagery eeg decoding. Applied Soft Computing 151, 111129.

[8] Gratton, G., Coles, M.G., Donchin, E., 1983. A new method for off-line removal of ocular artifact. Electroencephalography and clinical neurophysiology 55, 468–484.

[9] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

[10] He, P., Wilson, G., Russell, C., 2004. Removal of ocular artifacts from electro-encephalogram by adaptive filtering. Medical and biological engineering and computing 42, 407–412.

[11] Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.

[12] Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International conference on machine learning, pmlr. pp. 448–456.

[13] Jahan, I., Ahmed, M.F., Ali, M.O., Jang, Y.M., 2023. Self-gated rectified linear unit for performance improvement of deep neural networks. ICT Express 9, 320–325.

[14] Jiang, X., Bian, G.B., Tian, Z., 2019. Removal of artifacts from eeg signals: a review. Sensors 19, 987.

[15] Klados, M.A., Bamidis, P.D., 2016. A semi-simulated eeg/eog dataset for the comparison of eog artifact rejection techniques. Data in brief 8, 1004–1006.

[16] Klados, M.A., Papadelis, C., Braun, C., Bamidis, P.D., 2011. Reg-ica: a hybrid methodology combining blind source separation and regression techniques for the rejection of ocular artifacts. Biomedical Signal Processing and Control 6, 291–300.

[17] Lhoest, Q., del Moral, A.V., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., et al., 2021. Datasets: A community library for natural language processing. arXiv preprint arXiv:2109.02846 .

[18] Li, P., Zhang, Y., Liu, S., Lin, L., Zhang, H., Tang, T., Gao, D., 2023. An eeg-based brain cognitive dynamic recognition network for representations of brain fatigue. Applied Soft Computing 146, 110613.

[19] Lin, L., Li, P., Wang, Q., Bai, B., Cui, R., Yu, Z., Gao, D., Zhang, Y., 2024. An eeg-based cross-subject interpretable cnn for game player expertise level classification. Expert Systems with Applications 237, 121658.

[20] Mashhadi, N., Khuzani, A.Z., Heidari, M., Khaledyan, D., 2020. Deep learning denoising for eog artifacts removal from eeg signals, in: 2020 IEEE Global Humanitarian Technology Conference (GHTC), IEEE. pp. 1–6.

[21] McMenamin, B.W., Shackman, A.J., Maxwell, J.S., Bachhuber, D.R., Koppenhaver, A.M., Greischar, L.L., Davidson, R.J., 2010. Validation of ica-based myogenic artifact correction for scalp and source-localized eeg. Neuroimage 49, 2416–2432.

[22] Memory, L.S.T., 2010. Long short-term memory. Neural computation 9, 1735–1780.

[23] Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 .

[24] Molla, M.K.I., Islam, M.R., Tanaka, T., Rutkowski, T.M., 2012. Artifact suppression from eeg signals using data adaptive time domain filtering. Neurocomputing 97, 297–308.

[25] Pu, X., Yi, P., Chen, K., Ma, Z., Zhao, D., Ren, Y., 2022. Eegdnet: Fusing non-local and local self-similarity for eeg signal denoising with transformer. Computers in Biology and Medicine 151, 106248.

[26] Qin, Z., Li, Q., 2018. High rate bci with portable devices based on eeg. Smart Health 9, 115–128.

[27] Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

[28] Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y., 2024. Roformer: Enhanced transformer with rotary position embedding. Neurocomputing 568, 127063.

[29] Sun, W., Su, Y., Wu, X., Wu, X., 2020. A novel end-to-end 1d-rescnn model to remove artifact from eeg signals. Neurocomputing 404, 108–121.

[30] Sun, Y., Dong, L., Huang, S., Ma, S., Xia, Y., Xue, J., Wang, J., Wei, F., 2023. Retentive network: A successor to transformer for large language models. arXiv preprint arXiv:2307.08621 .

[31] Sun, Y., Dong, L., Patra, B., Ma, S., Huang, S., Benhaim, A., Chaudhary, V., Song, X., Wei, F., 2022. A length-extrapolatable transformer. arXiv preprint arXiv:2212.10554 .

[32] Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. Advances in neural information processing systems 27.

[33] Tiwari, A., Chaturvedi, A., 2019. A multiclass eeg signal classification model using spatial feature extraction and xgboost algorithm, in:

2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE. pp. 4169–4175.

[34] Tiwari, A., Chaturvedi, A., 2023. Automatic eeg channel selection for multiclass brain-computer interface classification using multiobjective improved firefly algorithm. Multimedia Tools and Applications 82, 5405–5433.

[35] Turnip, A., Junaidi, E., 2014. Removal artifacts from eeg signal using independent component analysis and principal component analysis, in: 2014 2nd International Conference on Technology, Informatics, Management, Engineering & Environment, IEEE. pp. 296–302.

[36] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

[37] Wang, B., Deng, F., Jiang, P., Wang, S., Han, X., Zheng, H., 2024. Witunet: A u-shaped architecture integrating cnn and transformer for improved feature alignment and local information fusion. arXiv preprint arXiv:2404.09533 .

[38] Weidong, Z., Yingyuan, L., 2001. Eeg multiresolution analysis using wavelet transform, in: 2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE. pp. 1854–1856.

[39] Wu, Y., He, K., 2018. Group normalization, in: Proceedings of the European conference on computer vision (ECCV), pp. 3–19.

[40] Yang, B., Duan, K., Fan, C., Hu, C., Wang, J., 2018. Automatic ocular artifacts removal in eeg using deep learning. Biomedical Signal Processing and Control 43, 148–158.

[41] Zaremba, W., Sutskever, I., Vinyals, O., 2014. Recurrent neural network regularization. arXiv preprint arXiv:1409.2329 .

[42] Zhang, H., Zhao, M., Wei, C., Mantini, D., Li, Z., Liu, Q., 2020. Eegdenoisenet: A benchmark dataset for end-to-end deep learning solutions of eeg denoising. arXiv preprint arXiv:2009.11662 .

[43] Zhuang, Z., Liu, M., Cutkosky, A., Orabona, F., 2022. Understanding adamw through proximal methods and scale-freeness. arXiv preprint arXiv:2202.00089 .

Unveiling Thoughts: A Review of Advancements in EEG Brain Signal Decoding into Text

Saydul Akbar Murad[1], *(IEEE Member)* Nick Rahimi[1], *(IEEE Member)*

***Abstract*—The conversion of brain activity into text using electroencephalography (EEG) has gained significant traction in recent years. Many researchers are working to develop new models to decode EEG signals into text form. Although this area has shown promising developments, it still faces numerous challenges that necessitate further improvement. It's important to outline this area's recent developments and future research directions. In this review article, we thoroughly summarize the progress in EEG-to-text conversion. Firstly, we talk about how EEG-to-text technology has grown and what problems we still face. Secondly, we discuss existing techniques used in this field. This includes methods for collecting EEG data, the steps to process these signals, and the development of systems capable of translating these signals into coherent text. We conclude with potential future research directions, emphasizing the need for enhanced accuracy, reduced system constraints, and the exploration of novel applications across varied sectors. By addressing these aspects, this review aims to contribute to developing more accessible and effective Brain-Computer Interface (BCI) technology for a broader user base.**

***Index Terms*—EEG, Neuroscience, Signal Decoding, Deep Learning, BCI.**

I. Introduction

EEG-based brain-to-text communication is a revolutionary field that explores using EEG to decode brain signals into actual text. EEG measures electrical activity on the scalp, and researchers are currently devising techniques to convert particular patterns into letters, words, and even sentences. This technological advancement exhibits significant promise for persons who experience speech or motor disabilities, offering a direct communication channel that bypasses traditional methods. BCIs play a pivotal role in EEG-based brain-to-text communication. BCIs bridge the gap between the brain and external devices by interpreting brain signals to facilitate control or communication. They have evolved significantly, from early systems focusing on simple commands to more advanced interfaces capable of recognizing complex thought patterns. This evolution is important for brain-to-text communication, as researchers rely on BCIs to translate the intricate neural correlates of language into meaningful text.

In recent years, there has been a significant increase in research focused on decoding EEG signals for the purpose of direct brain-to-text communication [1]. The increasing interest can be attributed to various things. Firstly, the progress in machine learning algorithms facilitated the researchers to analyze complex EEG patterns with greater accuracy [2]. Secondly, the potential applications of this technology are vast, particularly in the realm of communication assistance [3]. Finally, the non-invasive nature of EEG makes it a promising option for individuals who are unable to employ conventional modes of communication. For individuals with conditions like amyotrophic lateral sclerosis (ALS), stroke, or severe cerebral palsy, traditional communication methods can be severely limited or even impossible [1]. Brain-to-text communication presents a promising prospect, as it gives a direct means for individuals to articulate their thoughts and requirements [4]. This technology has the potential to revolutionize their lives, enabling them to interact with the world, express themselves creatively, and regain a sense of independence.

Early research in EEG primarily concentrated on areas like emotion recognition and neurological condition studies. For instance, in [5], researchers reviewed how EEG signals could be linked to understanding human emotions by analyzing brain activities. A Similar study [6] extended this focus to emotion recognition from EEG data, exploring a range of techniques from initiating emotions through the preprocessing of EEG signals to extracting and classifying features. This research also critically evaluated the strengths and weaknesses of these varied methods. Moving away from emotional analysis, the study [7] shifted the focus toward diagnosing Alzheimer's Disease. It offered an in-depth look into the use of EEG for detecting Alzheimer's, including a detailed analysis of the complexity found in the EEG signals associated with the disease. However, there's still a gap in the literature: no review articles have yet focused on converting EEG signals to text. This gap is significant, considering the potential applications and advancements this research could bring.

Considering the significance of EEG-to-text conversion, this review compiles and examines recent research in this burgeoning field. We aim to outline and discuss the techniques used in transforming EEG signals into text, aiming to guide future research directions. The contributions of our review are the following:

- Firstly, we address the various challenges faced in decoding EEG signals, providing insight into the complexities inherent in this process.
- Secondly, we design a comprehensive taxonomy that categorizes and discusses the range of techniques used in this domain, from initial data collection to the intricacies of model development.
- Lastly, we explore potential avenues for future research, identifying gaps in current knowledge and proposing areas that hold promise for further investigation.

[1] School of Computing Sciences & Computer Engineering, University of Southern Mississippi, Hattiesburg, MS, USA. (e-mail: saydulakbar.murad@usm.edu, nick.rahimi@usm.edu)

This review aims not just to go over what's been done but also to spark new ideas and paths in the field of EEG-to-text conversion.

The rest of this paper is structured as follows: In Section II, we delve into the challenges of EEG signal decoding, detailing six distinct types of challenges encountered in this field. Section III presents a taxonomy, concentrating on the techniques employed, from the initial stages of data collection to the intricacies of model development. In Section IV, we highlight potential directions for future research. The paper concludes with Section V, summarizing our findings and final thoughts.

II. Challenges in EEG Decoding

Despite the immense potential of EEG decoding for text generation, translating brain activity into written language presents significant challenges. This section delves into the complexities of this process, highlighting the key hurdles researchers face at each stage. Figure 1 presents the challenges of decoding the EEG signals into text. This figure provides a roadmap to understanding the complexities of EEG decoding for text generation. It outlines the various stages involved, from data acquisition to model building, and highlights the key challenges encountered at each step.

A. Data Acquisition

Signal Acquisition and Quality: Acquiring clean and high-quality EEG signals is essential for accurate text decoding. However, this process faces several hurdles. The inherent weakness of brain signals compared to background electrical noise, particularly from muscles and power lines, necessitates sophisticated noise reduction techniques [8], [9]. Furthermore, significant variability in brain activity patterns between individuals due to anatomical and cognitive differences poses a challenge for developing generalized decoding models that work effectively for everyone [10]. Even slight head movements can introduce artifacts, further muddying the signal [8]. Additionally, the limited spatial resolution of EEG, which measures activity from a large brain area, makes it difficult to pinpoint the exact source of language-related activity [11]. Finally, user comfort and training can be hurdles. Wearing an EEG cap with multiple electrodes can be uncomfortable for some users, and extensive training sessions may be required to learn how to control their brain activity for optimal decoding results [12].

Inter-subject Variability: Despite the potential of EEG for text decoding, a major hurdle lies in inter-subject variability. Unlike fingerprints, brain activity patterns are highly individualistic. This variation arises from several sources. Firstly, anatomical differences in brain structure and neuron distribution between people lead to diverse electrical activity patterns [12]. Secondly, cognitive styles influence which brain regions are activated during thought processes. Some individuals may rely heavily on visual processing, while others favor auditory or kinesthetic pathways [10]. Finally, even slight variations in how EEG electrodes are placed on the scalp can significantly impact the recorded signals and decoding accuracy across users [13]. As a result, decoding models may need to be personalized for each user, as a model trained on one person's EEG data might not perform well for someone with a different brain or cognitive style.

B. Data Preprocessing and Feature Selection

Non-Stationary Nature of EEG Signals: Understanding the intricate connection between brain activity and written language is further complicated by the non-stationary nature of EEG signals. Unlike stationary signals with consistent statistical properties over time, EEG signals exhibit dynamic changes [8]. These variances can arise from internal cognitive changes as users direct their attention towards different elements of the text they intend to produce [14]. Additionally, external factors like fatigue or slight head movements might cause temporary fluctuations in the signal [12]. This non-stationary nature poses challenges for data preprocessing and feature selection. Traditional techniques assuming stationary signals may not effectively capture the time-varying information crucial for accurate text decoding. Researchers are exploring methods like time-frequency analysis and adaptive filtering techniques to account for the non-stationary characteristics of EEG signals and extract the most relevant features for successfully decoding brain activity into text [15].

Identifying Informative Features: In the context of transforming EEG data to text, an important step in data preprocessing and feature selection involves identifying informative features. This process is important because EEG data is characteristically high-dimensional and contains a significant amount of non-informative or redundant information. Efficient feature selection can significantly improve the performance of machine learning models used in this domain. The challenge lies in distinguishing relevant features that are most representative of the underlying cognitive or neural processes from the irrelevant ones. Techniques such as PCA, ICA, and mutual information-based methods have been employed to address this issue [16], [17]. These methods aim to reduce the dimensionality of EEG data while retaining the most significant information, facilitating the subsequent machine learning tasks such as classification or regression required for converting EEG signals into text representations. This step is important for applications in BCI, where the goal is to translate neural activity into actionable commands or textual forms [18].

C. Model Building and Decoding

Limited training data: The scarcity of training data poses a major obstacle in the creation of efficient EEG-to-text algorithms. Due to the inherent complexity and variability of EEG signals, coupled with the difficulty in collecting large datasets, machine learning models often suffer from inadequate training, leading to poor generalization and performance. This problem is most noticeable in EEG-to-text applications, where the model needs to interpret intricate brain signals and produce accurate textual results. Transfer learning and data augmentation are often used techniques to address this difficulty. Transfer learning involves fine-tuning a pre-trained model on a smaller dataset, while data augmentation artificially

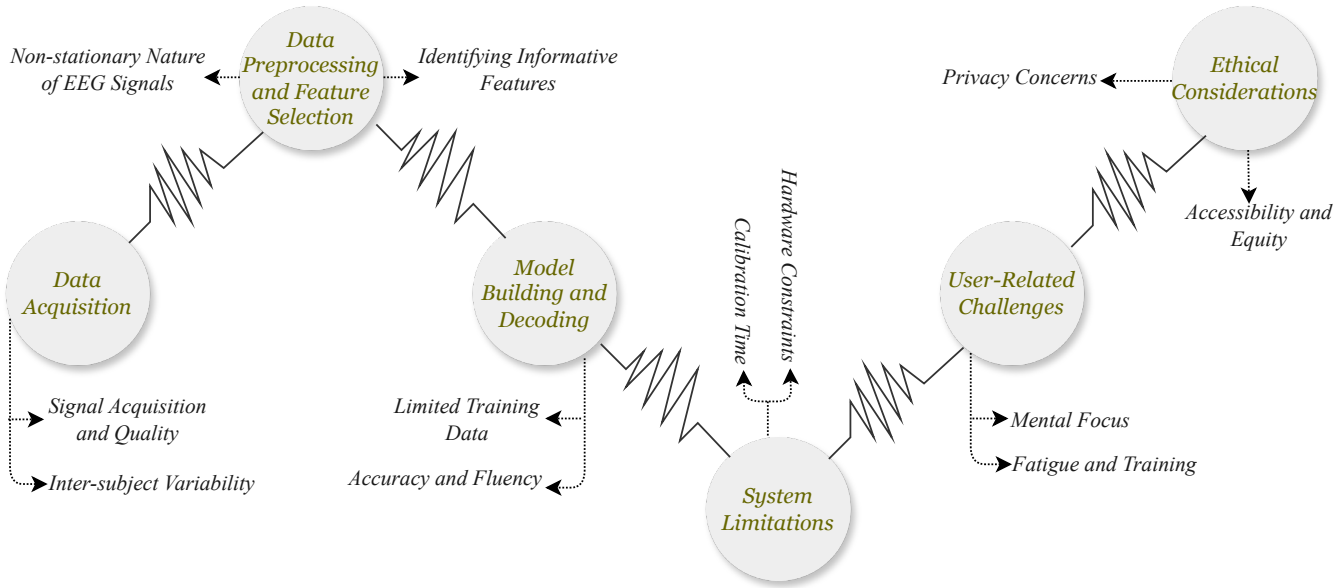


Figure 1: Challenges in EEG Signal Decoding for Text Generation.

enhances the quantity and variability of training datasets [19]. Furthermore, the use of generative models to create synthetic EEG data, which resembles real recordings, has shown promise in enhancing the training process (Hartmann et al., 2018). These approaches aim to overcome the limitations imposed by scarce EEG data, thus improving the accuracy and reliability of EEG-to-text conversion models crucial for applications in neural prosthesis and BCI [20].

Accuracy and Fluency: Attaining a high level of accuracy and fluency is still a difficult task when it comes to constructing models and decoding for EEG-to-text conversion. The inherent complexity of interpreting EEG signals, which are often noisy and highly individualized, poses significant difficulties in accurately translating these signals into coherent and fluent text. The accuracy of a model refers to its ability to accurately interpret the EEG signals, while fluency pertains to the naturalness and readability of the generated text. These dual objectives often require sophisticated algorithms that can handle the intricate patterns in EEG data. Researchers have studied the possibility of deep learning models, specifically recurrent neural networks (RNNs) and attention mechanisms, to improve both accuracy and fluency in this field [21]. These models can capture temporal dependencies and contextual nuances in the EEG signals, which are crucial for producing precise and fluent text output. However, the trade-off between accuracy and fluency remains a key area of research, as improving one aspect can sometimes be at the expense of the other. This balancing act is critical in applications like real-time communication aids for individuals with speech impairments, where both accuracy and fluency are essential for effective interaction [22], [23].

D. System Limitations

Hardware Constraints: Hardware limitations pose a substantial obstacle in the creation and execution of EEG-to-text systems. The capabilities of the EEG recording equipment directly affect the quality and resolution of EEG signals, which are essential for proper decoding. The availability of high-resolution EEG devices, which offer intricate neurological data, is often restricted due to their high cost and limited accessibility. Consequently, the general utilization of advanced EEG-to-text applications is hindered [24]. Moreover, these sophisticated devices can be unwieldy and intrusive, impeding their suitability for daily usage. However, portable and user-friendly EEG equipment typically have lower resolution and are more prone to noise and aberrations. These factors can have a negative impact on the performance of EEG-to-text models [25]. The difficulty is in achieving a harmonious equilibrium between the excellence of EEG data capture and the feasibility and availability of the technology. In addition, the processing and decoding of EEG data in real-time necessitate high-performance processing units, which may not be practical for portable or wearable devices [26], [27]. The hardware limitation plays a crucial role in assessing the practicality and efficiency of EEG-to-text systems, especially in the context of assistive communication devices designed for individuals with speech or mobility disabilities.

Calibration Time: Calibration time poses a substantial challenge in developing and implementing EEG-to-text models. Calibrating EEG devices to individual users is essential for accurately decoding neural signals. However, this process can be time-consuming and demands substantial human exertion. The brainwave patterns of each individual are distinct, and the EEG system must be precisely calibrated to these patterns to achieve successful communication or control. The calibration process often requires the user to engage in specific cognitive tasks regularly, enabling the system to learn and adjust to their EEG signals [27], [28]. The calibration process for EEG-to-text systems might be challenging in terms of time and effort, especially when considering their applicability to individuals with disabilities or in time-critical scenarios. Efforts to decrease the time required for calibration while also

upholding or enhancing the precision and dependability of the system are a crucial focus of research. Researchers are currently investigating emerging methods, such as adaptive algorithms and transfer learning, to tackle this difficulty. These methods involve using calibration data from one user to help calibrate the system for another user [26], [29]. The objective of these techniques is to enhance the user-friendliness and accessibility of EEG-to-text systems, hence expanding their range of potential applications.

E. User Related Challenges

Mental Focus: The use of EEG-to-text systems faces a notable user-related barrier in terms of mental attention. The success of these systems heavily depends on the user's capacity to sustain unwavering mental concentration, as fluctuations in attention can result in significant deviations in EEG signal patterns, thus impacting the precision of signal interpretation and text generation. This is particularly challenging because maintaining a high concentration level over extended periods is difficult for most individuals. Fatigue, distraction, and cognitive load can all adversely impact the user's ability to generate stable and clear EEG signals [30]. Moreover, the requirement for sustained mental focus can be especially demanding for users with cognitive or attention impairments, limiting the accessibility and usability of EEG-to-text technologies for these populations. Researchers are exploring various strategies to address this challenge, such as developing more robust algorithms that can cope with fluctuating levels of user attention and integrating adaptive learning systems that adjust to the user's mental state [27]. The purpose of these developments is to improve the robustness of EEG-to-text systems against fluctuations in mental concentration, thus increasing their practicality and efficacy for a wider spectrum of users.

Fatigue and Training: User-related issues in EEG-to-text communication systems include fatigue and the requirement for substantial training. Users often experience fatigue during prolonged use of EEG systems, as the process of generating consistent and accurate neural signals for text communication requires considerable mental effort and concentration. Fatigue can result in a deterioration of the quality of the EEG signals, which in turn affects the function of the system [31]. Furthermore, the requirement for extensive training to use EEG-to-text systems efficiently can be a barrier to their widespread adoption, particularly for individuals with disabilities or those who lack the time and resources for lengthy training sessions. The training process involves users learning to generate distinct neural patterns that the system can recognize and translate into text, which can be a time-consuming and demanding task [32]. To address these challenges, researchers are focusing on developing more intuitive and user-friendly interfaces, as well as adaptive algorithms that require less user training and are more resilient to the effects of fatigue [27].

F. Ethical Considerations

Privacy Concerns: Privacy concerns constitute a critical ethical challenge in the realm of EEG-to-text technology. EEG data, which might disclose private and confidential information about a person's mental condition, thoughts, or intentions, presents substantial concerns regarding privacy [33]. Preserving the confidentiality and security of this data is of utmost importance, as breaches could result in unauthorized access and exploitation of personal information. This is particularly relevant in the context of BCIs, where EEG data is used for direct communication or control. The risk of eavesdropping or hacking into these systems poses a serious threat to user privacy [34]. To address these concerns, researchers and developers are exploring various data protection strategies, such as advanced encryption methods and strict access controls, to safeguard against unauthorized data access and ensure compliance with privacy regulations [35]. Furthermore, ethical guidelines and frameworks are being developed to guide the responsible use of EEG data and protect individuals' privacy rights in the use of EEG-to-text and other BCI technologies [36].

Accessibility and Equity: Ensuring accessibility and equity are crucial ethical considerations while developing and using EEG-to-text technology. The potential of these systems to provide communication aids for individuals with disabilities highlights the need for inclusive design and equitable access. Nevertheless, there is a potential danger that these technologies may exacerbate the disparity between individuals who have access to sophisticated medical and assistive technologies and those who do not, owing to factors such as expense, technological proficiency, and the presence of healthcare facilities [37]. Furthermore, the development of EEG-to-text systems frequently prioritizes the preferences of a typical user, possibly neglecting the distinct demands of individuals with different abilities and backgrounds. This lack of inclusivity can result in technologies that are not equally accessible or beneficial to all potential users [38]. Addressing these issues, it is essential to focus on developing EEG-to-text and other BCI technologies using universal design principles. This involves considering the varied requirements and situations of different user groups. Furthermore, policy measures and funding initiatives could play a significant role in ensuring equitable access to these technologies, particularly for underserved and marginalized communities [39].

III. EEG to Generative Model Pipeline

Extracting meaningful information from electroencephalogram (EEG) signals is crucial for various applications in neuroscience, brain-computer interfaces, and clinical diagnosis. This process typically involves several stages, as illustrated in Figure 2. The initial stage focuses on data acquisition, where electrodes record EEG signals from the scalp. Subsequently, preprocessing techniques are employed to remove noise and artifacts, ensuring the quality of the data for further analysis. After the EEG signals have been preprocessed, feature extraction is a key step in turning them into useful features that capture the right characteristics for the application. This article delves into various feature extraction techniques commonly employed in EEG signal processing.

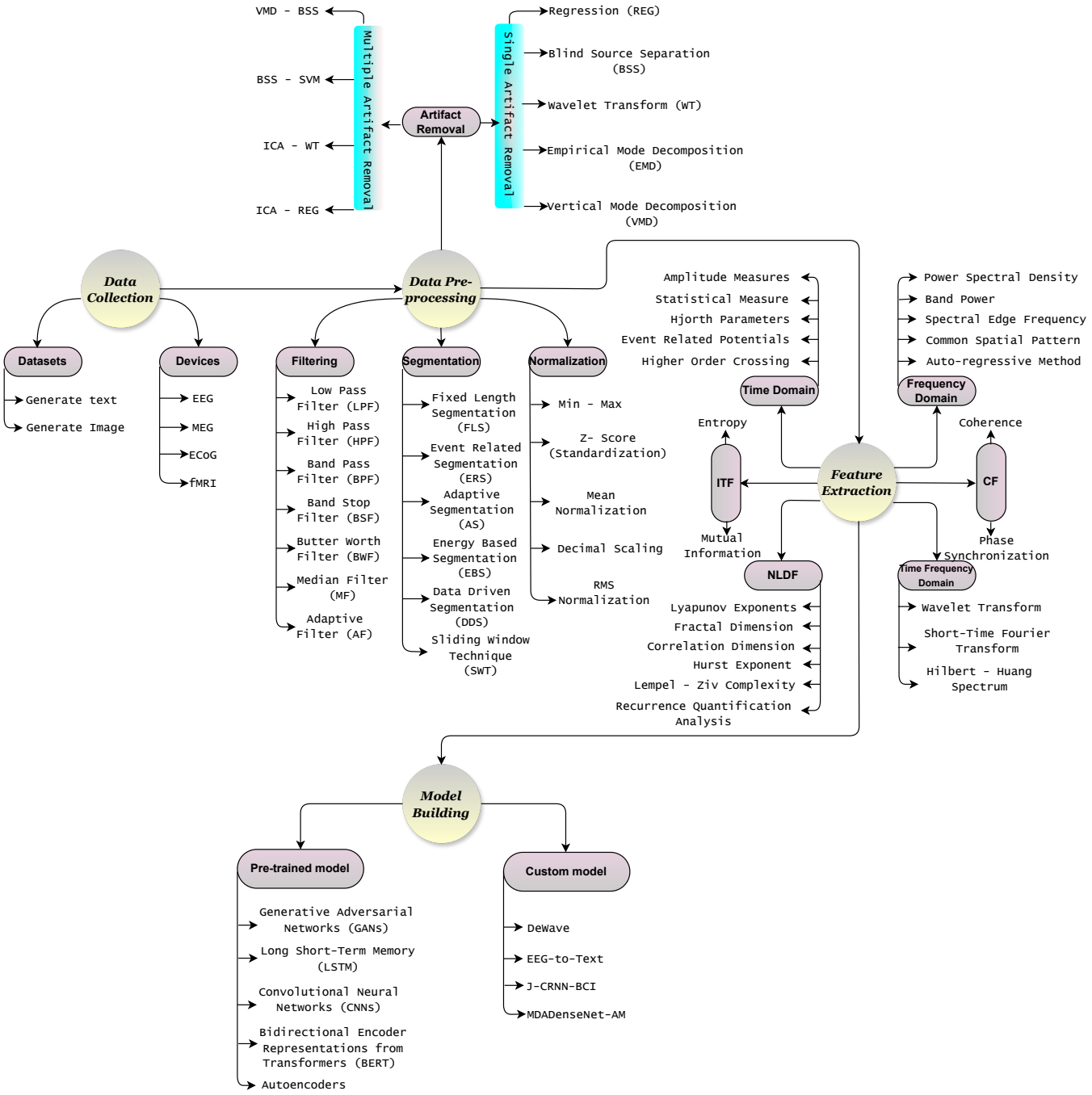


Figure 2: Taxonomy of EEG Signal Processing for Text and Image Generation.

A. Data Acquisition

The data acquisition process encompasses two key components: the dataset and the device. The dataset aspect involves an examination of existing data previously utilized for converting brain signals into text and images. This part of the process focuses on how this data was initially collected and processed to facilitate the transformation of neural activity into comprehensible formats like text and visual imagery. Conversely, the device segment delves into the various instruments employed to gather data directly from the brain. This includes an array of technologies ranging from non-invasive tools like EEG, Magnetoencephalography (MEG), and Functional MRI (fMRI) to more invasive methods such as Electrocorticography (ECoG).

1) Datasets:

a) Generate Text : Many researchers have made significant contributions to the field of brain signal decoding into text, resulting in the publication of various datasets. These datasets have been gathered using both invasive techniques, like ECoG, and non-invasive techniques, including EEG and fMRI. Notably, ZuCo 1.0 [40] and ZuCo 2.0 [41] are prominent EEG-based datasets collected using setups with 128 channels. In contrast, when utilizing fMRI for data collection [42], [43], researchers often employ 32-channel head coils.

Additionally, ECoG, a widely used method for brain signal acquisition [44], [45], typically involves the use of 16 channels during experiments. Table I lists the datasets utilized for decoding human brain signals into text.

b) Generate Image : Many other researchers have focused on creating images from brain signals, primarily utilizing non-invasive methods such as EEG [46]–[48] and fMRI [46], [49]–[51]. In these studies, the stimuli primarily consist of images, though a few studies have also explored the use of text as stimuli. This exploration into brain signal-based image generation is a growing field, delving into the complex relationship between neural activity and visual representation. By analyzing brain responses to visual stimuli, these studies aim to reconstruct or generate images that correlate with the observed brain activity, providing insights into how the brain processes and interprets visual information. Table II provides a comprehensive list of datasets used to convert human brain signals into images.

2) Devices : Various companies offer a variety of devices for monitoring human brain signals, utilizing both invasive and non-invasive methods to collect data. In non-invasive techniques, electrodes are positioned on the scalp without surgical intervention. This approach is commonly used due to its safety and ease of application. In contrast, invasive techniques involve a surgical process to place electrodes directly on or within the brain, providing more direct and often more detailed brain signal readings. Table III provides a comprehensive overview of the most popular devices currently used. This includes various organizations such as NeuroScan [52], Brain Products [53], BioSemi [54], Emotiv [55], NeuroSky [56], ANT Neuro [57], ABM [58], and OpenBCI [59]. Each organization designs devices with varying ranges of electrical channels and frequencies.

B. Data Pre-processing

1) Artifact Removal: During EEG signal collection, artifacts are a common issue. These artifacts can be of two types: psychological and non-psychological. Non-psychological artifacts originate from external sources such as electrode malfunctions, the movement of cables, or poor connections in channels. On the other hand, physiological artifacts are caused by internal electrical signals within the body. It is necessary to remove these unwanted signals. Some physiological artifacts, like those caused by skin and sweat, can be eliminated during neural activity recording. For example, wearing a cooling ventilation vest helps regulate body temperature during physical activities. Power line artifacts can be effectively filtered out using notch filtering techniques. Moreover, movement-related distortions in the EEG signal, caused by head or body motion and the resultant electrode and cable movement, can be reduced by employing a double-layer cap.

In the field of EEG signal processing, various strategies have been developed to eliminate artifacts, as depicted in Figure 1. These techniques are broadly categorized into two groups: single artifact removal and multiple artifact removal methods. Common single artifact removal methods utilized by researchers include techniques such as Regression (REG) [60], Blind Source Separation (BSS) [61], Wavelet Transform (WT) [62], [63], Empirical Mode Decomposition (EMD) [64], and Vertical Mode Decomposition (VMD) [65]. In recent times, there has been a growing trend among researchers to combine two single artifact removal methods to enhance the efficacy of artifact removal from EEG signals. This approach has led to the development of sophisticated hybrid techniques such as VMD-BSS [66], BSS-SVM [67], ICA-WT [68], and REG-ICA [69]. These multiple artifact removal techniques offer a more robust and comprehensive approach, addressing a broader range of artifacts compared to single artifact removal methods. Consequently, they have gained popularity in the research community for their improved ability to clean EEG signals from various types of unwanted interference. Table IV delineates a range of artifact mitigation strategies employed in EEG signal analysis as explored by various researchers.

2) Filtering : Filtering is another important step of EEG signal processing that is used to enhance the quality of the signal by minimizing noise and interference that can obscure the actual signals. There are many ways from which the unwanted signals come, such as power line noise, environmental electromagnetic interference, and physiological artifacts like muscle movements or eye blinks. By removing these unwanted signals, filtering makes it possible to separate the meaningful brainwave patterns that are useful in research. This isolation is particularly important because EEG signals are typically weak and can be easily contaminated by extraneous noise.

EEG signals can be categorized into two primary types: linear and nonlinear. Linear filters, such as low-pass [70], high-pass [71], band-pass [72], band-stop [73], and butter-worth filters [74], are commonly used to remove frequency components outside of the desired range. For instance, low-pass filters are employed to filter out high-frequency disturbances, while high-pass filters are designed to eliminate low-frequency noise. Conversely, bandpass filters are adept at removing noise that lies outside a designated frequency band. Nonlinear filters, on the other hand, include adaptive filters [75] and median filters [76], which are more complex and can be tailored to the specific characteristics of the EEG signal. Adaptive filters are particularly useful in scenarios where the signal or noise characteristics are changing over time, as they can dynamically adjust their filtering parameters. The choice of filter type and settings mainly depends on the characteristics of the EEG data and the goals of the analysis. This flexibility and specificity in filtering ensure that the most relevant and accurate information is extracted from the EEG data, significantly enhancing the quality of the signals.

3) Segmentation: Segmentation in EEG signal processing is an important preprocessing step that involves dividing continuous EEG data into smaller, more manageable segments. It's important for many different reasons. Firstly, it enhances noise reduction, as segmenting the signal allows for easier identification and elimination of artifacts and interference [75]. Secondly, it facilitates event-related analysis, particularly in cognitive or sensory studies, by enabling precise examination of brain responses to specific stimuli [77]. Moreover, considering the non-stationary nature of EEG signals, segmentation helps analyze data within shorter, more statistically uniform

Table I: Description of Publicly Available Datasets Used for Text Generation

Datasets	Participants				Channels	Stimulus
	Male	*Female*	*Total*	*Age*		
ZuCo 1.0 [40]	7	5	12	22-54	EEG (128)	1107 English sentences
ZuCo 2.0 [41]	9	10	19	23-54	EEG (128)	739 English sentences
fMRI Image [42]	-	-	5	21-50	32-channel head coil	15 subjects with 540 scans
fMRI Image [43]	-	-	-	-	fMRI	180 fMRI Images
ECoG Data [44]	-	-	7	-	Eight16-channel g.USBamp biosignal amplifiers	4381 second voice recording
ECoG Data [45]	1	4	5	29-49	16 Channel	4053 English sentences

Table II: Description of Publicly Available Datasets Used for Image Generation

Datasets	Participants				Channels	Stimulus
	Male	*Female*	*Total*	*Age*		
EEG + fMRI [46]	-	-	5	25-27	EEG(64), 3.0-Tesla Siemens MAGNETOM Verio scanner	50 images in 20 category
EEG data [47]	-	-	23	15-40	EEG (14)	20 text and10 non-text items
EEG data [48]	5	1	6	-	EEG (128)	2000 image
fMRI [49]	2	1	3	23-33	3.0-Tesla Siemens MAGNETOM Verio scanner	1200 natural images
fMRI [50]	-	-	2	-	4T INOVAMR scanner	1870 images
NSD [51]	-	-	8	-	fMRI	27750 fMRI-image

Table III: Name of Companies Offering Brain Signal Acquisition Devices with Varied Functionalities

Name	*Sampling Rate*	*Channel*	*Devices*		
			EEG	*MEG*	*fMRI*
NeuroScan [52]	500 Hz	32, 64, 128, 256	✓	✓	✓
Brain Products [53]	1000 Hz, 500 Hz, 250 HZ	8, 16, 32, 64	✓	✗	✓
BioSemi [54]	2048 Hz, 1024 Hz	32, 64, 128, 256	✓	✗	✗
Emotiv [55]	2048 Hz, 128 Hz	2, 5, 14, 32	✓	✗	✗
NeuroSky [56]	512 Hz	12	✓	✗	✗
ANT Neuro [57]	16 kHz, 2048 Hz	32, 64, 128, 256	✓	✓	✓
ABM [58]	4 kHz, 2048 Hz	24, 32, 64	✓	✗	✗
OpenBCI [59]	125 Hz, 200 Hz, 250 Hz	4, 8, 16	✓	✗	✗

intervals, ensuring more accurate interpretations. Finally, from a practical standpoint, segmentation increases computational efficiency by breaking down lengthy EEG recordings into smaller, more computationally manageable units. This preprocessing step is particularly important in studies where temporal precision and data quality are paramount, such as in cognitive neuroscience research, clinical diagnostics, and real-time BCI systems.

Researchers employ various segmentation techniques to mitigate noise in EEG signal processing, each with its own unique approach and application. Fixed-Length Segmentation (FLS) [78] is widely used for its simplicity, segmenting the EEG data into equal, predetermined lengths, thus providing a uniform framework for analysis. Event-related Segmentation (ERS) [79] is tailored for cognitive and sensory studies, where it segments the signal based on specific stimuli or events. Adaptive Segmentation (AS) [75] offers flexibility by adjusting the segment length in response to the signal's characteristics, making it ideal for non-stationary data. Energy-Based Segmentation (EBS) [80] relies on the signal's energy content to determine segment boundaries, which is particularly useful in detecting and analyzing high-energy neural events. Data-Driven Segmentation (DDS) [81] employs algorithms to segment the data based on inherent features of the EEG signal itself, thus adapting to the signal's natural structure. Finally, the Sliding Window Transform (SWT) [82] is a complex method that uses wavelet transforms to deal with non-stationary data. It offers a multi-resolution analysis that is useful for pulling out complex EEG signals' nuanced features. Each of these techniques offers distinct advantages and is chosen based on the specific requirements of the study.

4) Normalization : Normalization is another important step in EEG signal processing, mainly due to the high variability of EEG signals both within and between individuals. Normalization helps to mitigate inherent noise in EEG data, such as electrical interference or artifacts from muscle movements, by scaling the EEG signals. EEG signals ensure that outliers or variations in amplitude do not dominate the signal; normalization facilitates the extraction of meaningful biomarkers from the EEG data, which are vital for subsequent analysis and interpretation.

Researchers employ various techniques to normalize EEG signals, each with its own distinct approach to standardizing the data. Min-Max [83] normalization rescales the data to a fixed range, typically [0, 1], ensuring that each signal falls within a standardized band. Z-score [84] normalization, also known as standard score normalization, centers the data around the mean with a unit standard deviation, thereby addressing scale and distribution shape issues. Mean normalization [85] adjusts the data values to revolve around the mean, effectively balancing the dataset around a central value. Decimal normalization [86] shifts the decimal point of values, standardizing them based on their magnitude, which is particularly useful for varying signal amplitudes. Lastly, RMS (Root Mean Square) normalization [87] scales the signal by the square root of the mean squared value, often used to maintain the signal's energy across different conditions. These normalization methods collectively enhance EEG data's comparability and consistency, facilitating more accurate analyses.

C. Feature Extraction

Feature extraction is a critical step in EEG signal processing due to the inherent complexity and high dimensionality of the

Table IV: Comparative Overview of Artifact Removal Techniques in EEG Signal Processing

Ref.	Type of Artifact	Artifact Removal Technique	Single Artifact Removal	Multiple Artifact Removal	Application Area	Performance Metrics
[62]	Motion, Eye blinking, EMG Artifact	WT	✓	✗	Clinical Monitoring	Signal-to-noise ratio (SNR)
[63]	Eye movement and blinking, Swallowing, Chewing, Limb movement, Body movement artifact	WT	✓	✗	Brain-Computer Interface (BCI)	Signal-to-noise ratio (SNR), root mean square error (RMSE), Lambda
[61]	Ocular, cardiac, muscle and powerline artifacts	BSS	✓	✗	Epileptic spike and seizure detection and brain-computer interfaces (BCIs)	Signal-to-artifact ratio (SAR)
[67]	Eye blinks and heart rhythm artifacts	BSS - SVM	✗	✓	Neurology and brain research	Signal-to-Artifact Ratio (SAR)
[68]	Electrocardiographic	ICA - WT	✗	✓	Biomedical engineering	Signal-to-Artifact Ratio (SAR)
[69]	Eye movements and blink artifacts	REG - ICA	✗	✓	Biomedical signal processing	Artifact to signal ratio (ASR), Root mean square error (RMSE), Power spectral density (PSD)
[60]	Eye blinking artifacts	REG	✓	✗	Clinical Neurology	Root Mean Square Error (RMSE), Mutual Information (MI)
[64]	Muscle artifacts	EMD	✓	✗	Patients with movement disorders	Relative Root Mean Square Error (RRMSE)
[65]	Eye blinking, flutters and lateral eye movements artifacts	VMD	✓	✗	Neurophysiology and clinical neurology	Multiscale modified sample entropy (mMSE)
[66]	Muscular activity, heartbeat, and eye movements	VMD - BSS	✗	✓	N/A	Euclidean Distance (ED), Spearman Correlation Coefficient (SCC)

Table V: Comparison of Feature Extraction Methods for EEG Signal Processing.

Ref.	Time Domain	Frequency Domain	Time Frequency Domain	NLDF	ITF	CF
[88]	✓	✓	✓	✗	✗	✗
[89]	✗	✗	✓	✗	✗	✗
[90]	✗	✓	✓	✗	✗	✗
[91]	✗	✓	✗	✗	✗	✗
[92]	✗	✗	✗	✗	✗	✓
[93]	✗	✗	✗	✗	✓	✗
[94]	✗	✗	✗	✗	✗	✓
[95]	✗	✗	✗	✓	✗	✗
[96]	✓	✗	✗	✗	✗	✗
[97]	✗	✓	✗	✗	✗	✗
[98]	✗	✗	✓	✗	✗	✗

data. It acts as a bridge between raw EEG recordings and meaningful insights. Extracting relevant and informative features from various domains (time, frequency, time-frequency, and space) compresses the data while preserving key information crucial for subsequent analysis and interpretation. This allows researchers to effectively utilize ML algorithms and unlock the hidden potential of EEG data for various applications in neuroscience, BCI, and clinical diagnosis. Figure V shows a comparison of different feature extraction methods that were used in previous research.

1) Time Domain : Time-domain analysis is a fundamental approach for extracting features from EEG signals. It focuses on characterizing the signal's behavior directly over time, offering insights into its amplitude, variability, and rhythmicity. This domain proves valuable for capturing transient events and quantifying specific characteristics within the signal. Common time domain techniques include Amplitude measures [96], Statistical measures [99], Hjorth parameters [100], [101], Event-related potentials (ERPs) [102], and Higher-order crossing (HOC) [101]. Each method contributes uniquely to the comprehensive analysis of EEG signals, making them indispensable in extracting meaningful information from complex brain activity.

2) Frequency Domain : The frequency domain approach is another highly effective technique for feature extraction in EEG signal processing. This method transforms EEG signals to analyze their frequency components, providing a different perspective compared to time-domain analysis. This domain proves valuable for understanding the dominant rhythms associated with various brain activities and identifying event-related changes in the frequency spectrum. Key methods utilized in this domain include Power Spectral Density [97], Band Power [103], Spectral Edge Frequency [104], Common Spatial Pattern [105], and the Auto-regressive Method [106]. Each of these plays a significant role in analyzing and interpreting the complex frequency-based characteristics of EEG data.

3) Time Frequency Domain : While both time-domain and frequency-domain analyses offer valuable insights into EEG signals, they each provide a limited perspective. Time-domain analysis excels at capturing temporal dynamics but lacks resolution in the frequency domain. Conversely, frequency-domain analysis excels at revealing spectral characteristics but fails to capture how these characteristics change over time.

To overcome these limitations, researchers often employ time-frequency domain analysis. This approach provides a comprehensive understanding of EEG signals by simultaneously analyzing their temporal and spectral information. By decomposing the signal into its time-frequency components, researchers can gain insights into how the frequency content of the signal evolves over time. Among the most prominent techniques in this domain are the Wavelet Transform [62], [63], Short-Time Fourier Transform (STFT) [**?**], and Hilbert-Huang Spectrum [98].

4) Non-linear Dynamic Features (NLDF) : NLDF is a new way to look at EEG signals beyond the usual time-domain, frequency-domain, and time-frequency methods. This method explores the inherent non-linearity and complexity of brain activity, which linear models are unable to capture fully. NLDF techniques focus on extracting features that characterize the dynamic behavior of the EEG signal over time. These

features often involve measures of complexity, chaos, and synchronization, providing insights into the underlying mechanisms of brain function. NLDF such as Lyapunov Exponents [107], Fractal Dimension [95], Correlation Dimension [108], Hurst Exponent [108], Lempel-Ziv Complexity [109], and Recurrence Quantification Analysis [110] are pivotal in EEG signal processing for capturing the complex, dynamic behavior of the brain's electrical activity.

5) Information Theoretic Features (ITF) : In recent years, ITF has emerged as a powerful tool for extracting meaningful information from EEG signals. Entropy [93], a fundamental ITF measure, evaluates the unpredictability or randomness of the signal, offering insights into the complexity of neural dynamics. It is particularly useful in assessing the regularity and predictability of EEG signals, which can be pivotal in differentiating between different neurological states or conditions. On the other hand, Mutual Information [111] measures the amount of information shared between two signals, reflecting the degree of statistical dependency and potential interaction between different brain regions.

6) Connectivity Features (CF) : Beyond analyzing individual brain regions, CF offers a powerful tool for understanding the interplay between different brain areas in EEG signal processing. These features aim to quantify the functional relationships and synchronization patterns between various regions, providing valuable insights into the coordinated activity underlying cognitive processes and brain function. Coherence [92] is a widely utilized CF technique that measures the degree of correlation between the activities of different brain regions in the frequency domain. Phase synchronization [94] goes a step further, evaluating the temporal alignment of neural oscillations across different regions, which can reveal intricate patterns of neural communication pivotal for cognitive and motor functions.

D. Model Building

Many researchers have dedicated their efforts to translating human brain signals into text, exploring various innovative techniques. This field of study encompasses different methodologies for data collection, including EEG, fMRI, and ECoG. Each method offers unique insights and approaches to understanding brain activity. Deep neural networks, particularly RNNs [10], [112] and Long Short-Term Memory (LSTM) [113] networks, have demonstrated significant potential in decoding EEG signals into text. The BART [10] model is also emerging as a noteworthy tool in this domain. Bidirectional Auto-Regressive Transformers (BART), known for its effectiveness in various natural language processing tasks, is being adapted to interpret and convert EEG signals into coherent text. This adaptation signifies a promising convergence of advanced neural network architectures and neuroscientific data, potentially leading to more accurate and efficient EEG-to-text conversion methodologies. In this article, the primary emphasis is on those papers that utilize EEG signals to generate text. This approach presents a fascinating challenge and holds significant potential for advancements in neuroscientific research and assistive technologies.

*1) **DeWave** [112] :* DeWave is a novel end-to-end model for EEG-to-text translation using discrete codex encoding. It uses self-supervised learning and contrastive learning to establish the link between brain activity and language. DeWave eliminates the need for pre-processing steps like feature extraction, potentially simplifying the overall decoding process. Figure 3 describes the architecture of the DeWave model.

Key steps in DeWave:

Discrete Codex Encoding: DeWave utilizes a technique called discrete codex encoding. This involves learning a codebook that maps continuous EEG signals into discrete code tokens. Imagine this codebook as a dictionary, where specific patterns in the EEG signal map to specific words or phrases.

Self-supervised Wave Encoding: The model learns this codebook through a self-supervised learning process. It essentially analyzes vast amounts of unlabeled EEG data and the corresponding text, allowing the model to discover the inherent relationships between brain activity and language.

Contrastive Learning for Alignment: DeWave employs contrastive learning to refine the alignment between the encoded EEG representations and the corresponding text. This process helps the model identify the most relevant EEG patterns that accurately reflect the intended text.

Text Generation: Finally, the model utilizes a decoder, often a Transformer-based architecture, which takes the encoded EEG representations (discrete code tokens) and translates them into the final text output, sentence by sentence.

*2) **MDADenseNet-AM** [114] :* The MDADenseNet-AM model is a complex and sophisticated deep learning architecture designed for converting EEG signals into text. Here's a breakdown of its mathematical and operational structure:

DenseNet Mechanism: The model uses the DenseNet mechanism, where the output from the m^{th} convolutional layer (denoted as xd_m) is determined by the application of 3×3 convolution filters wd_dm on the outputs of all preceding layers. The mathematical representation of this output is given by:

$$xd_m = \psi\left([xd_0, xd_1, \ldots, xd_{m-1}]\right) \otimes wd_m \tag{1}$$

Multiscale Dilated Convolution: To address issues like aliasing, the model incorporates a multiscale dilated convolution operation. This operation allows the network to have a variable and adjustable receptive field, enabling it to capture information at different scales and contexts. The multiscale dilated convolution is characterized by different dilation factors, which determine the spacing of the convolution kernel elements and influence the channel obtained in the network.

Integration with Attention Mechanism: The attention mechanism is integrated into the MDADenseNet to focus on the most relevant information and suppress less important details. This is achieved by performing a weighted summation over the hidden layers of the network. Mathematically, for feature vectors hh_v, the environment vector cc_v is calculated as follows:

$$cc_v = \sum_{v=1}^{V} aa_v hh_v \tag{2}$$

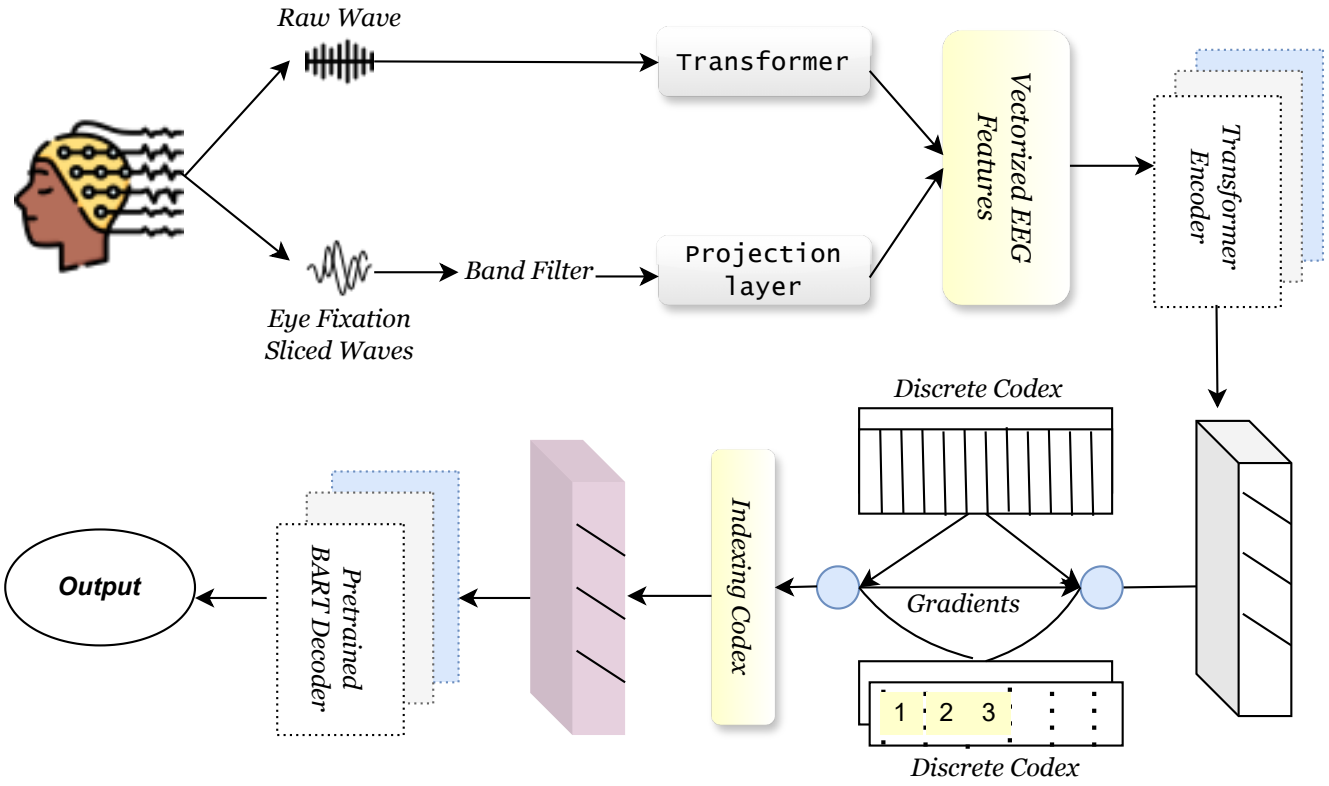


Figure 3: DeWave: A Transformer-Based EEG-to-Text Conversion Model.

Here, aa_v represents the weights combined with the hidden state, indicating the importance of each feature in the context of the task.

Parameter Optimization with EOWGMO: The model is developed with an adaptive strategy for parameter optimization using the Eurasian Oystercatcher Wild Geese Migration Optimization (EOWGMO) algorithm. This approach optimizes parameters like optimizers and epochs in the DenseNet to enhance the accuracy and precision of the text conversion model. The optimization function can be represented as:

$$ON(2) = \arg\min \{OpDnt_{rr} EpDnt_{qr}\} \frac{1}{Ac + Pn} \tag{3}$$

In this equation, $OpDnt_{rr}$ and $EpDnt_{qr}$ likely represent the ranges for the optimizer and epochs, respectively, while Ac and Pn are the accuracy and precision of the model.

3) ***EEG-to-Text*** *[10] :* This model, named EEG-To-Text, proposes a novel approach for open-vocabulary EEG-to-text decoding. It utilizes a combination of CNNs for feature extraction and LSTMs for sequence-to-sequence decoding. The model aims to overcome the limitations of pre-defined vocabularies and enable expressing a wider range of thoughts. The total working flow of the EEG-to-Text conversation model is shown in figure 4.

In the decoding stage of EEG-To-Text, LSTM networks play a crucial role in translating extracted EEG features into natural language sentences. Unlike traditional models, LSTMs excel at handling sequential data like EEG signals. Their internal memory mechanism allows them to retain information from previous time steps, capturing long-range dependencies between brain activity and intended meaning. Through a combination of "gates" that control information flow, the LSTM network analyzes current EEG features alongside previously predicted words, building the sentence word by word while considering the context of the entire sequence. This enables EEG-To-Text to decode diverse vocabulary and potentially capture the nuances of human thought.

4) ***J-CRNN-BCI*** *[113] :* This paper proposed a deep learning framework for brain-computer interface (BCI) typing. This framework utilizes a joint convolutional recurrent neural network (J-CRNN) to decode motor imagery EEG (MI-EEG) signals, translating imagined typing movements into actual text output.

Convolutional Neural Network (CNN): The first stage of the J-CRNN acts as a feature extractor. It processes the raw EEG data, which captures electrical activity across the brain, and identifies spatiotemporal patterns related to specific imagined movements. By applying filters and performing convolutions, the CNN extracts relevant features that differentiate between different typing intentions.

Recurrent Neural Network (RNN): The second stage of the J-CRNN captures the sequential nature of the EEG signals. Unlike the CNN, which analyzes individual data points, the RNN considers the temporal dependencies between these points. This allows the model to understand the evolving patterns within the EEG signal over time, which is crucial for distinguishing between the unique sequences associated with each imagined letter.

By combining the strengths of CNNs and RNNs in the J-

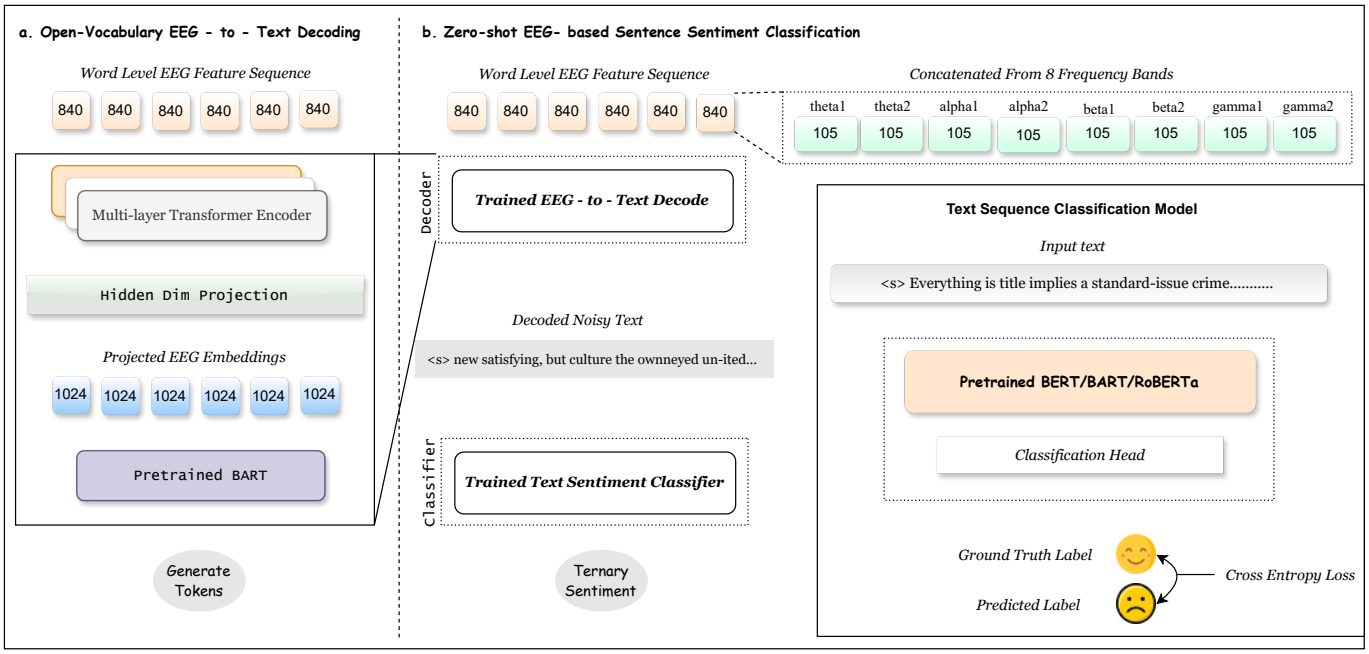


Figure 4: Integrative Framework for EEG Signal Decoding and Sentiment Analysis: Leveraging Pretrained Language Models for Text Generation and Emotion Classification.

CRNN architecture, the model effectively extracts informative features from the EEG data while simultaneously capturing the temporal dynamics of brain activity.

IV. Future Research Direction

A. Decoding complex thoughts and emotions

The recent advancements in EEG-to-text conversion are a significant achievement in neuroscientific research, offering new opportunities for communication, especially for those with speech and movement limitations. However, a notable gap in this domain is the current inability to accurately interpret and convey emotions through EEG signals. While researchers [10], [112] have developed models capable of translating brain activity into text, these models predominantly concentrate on the literal content and fail to consider the emotional context. This limitation underscores the complexity of human emotions and their representation in brain activity, which existing technologies have not fully captured. It requires a deeper understanding of the neural correlates of emotions and the development of more sophisticated algorithms to address these challenges. Enhancing the emotional sensitivity of EEG-to-text systems will improve the fidelity of communication for users and has implications for fields like mental health. In this context, recognizing emotions can facilitate accurate diagnosis and effective therapy. Therefore, the next research phase in this domain focuses on integrating emotional intelligence into EEG-to-text systems, which will bridge the gap between technological capability and the nuanced spectrum of human expression.

B. Data Source Diversification

The advancement of research in EEG-to-text generation is contingent upon the diversification of data sources, a crucial step towards enhancing the accuracy and applicability of these systems. Currently, EEG datasets used in research are often limited in size and diversity, leading to models that may exhibit poor generalizability across diverse populations or settings. To address this, it is important to integrate a broader array of data sources, encompassing datasets from diverse demographics, emotional states, and environmental situations. Diversifying data sources improves the robustness of EEG-to-text models and ensures their inclusivity and adaptability in real-world scenarios. This approach aligns with the growing AI and machine learning trend towards creating more equitable and universally applicable technologies. Future research will likely focus on establishing large, varied datasets and developing models that can effectively learn from such complexity. This will facilitate the creation of EEG-to-text systems that are precise, dependable, and reflective of the wide-ranging human encounter.

C. Improving Accuracy and Reducing Errors

An important area of future research in EEG-to-text generation lies in enhancing accuracy and minimizing errors. This can be tackled through advancements in two key areas: deep learning architectures and signal processing techniques. On the one hand, novel deep learning models specifically designed to handle the inherent noise present in EEG data can be explored. These models could use techniques like residual connections or attention mechanisms to extract more robust features and reduce the impact of noise on text generation. On the other hand, developing advanced filtering and artifact removal methods can significantly improve the quality of EEG signals before they are fed into the text generation model. By integrating these approaches, researchers can significantly reduce errors and pave the way for more reliable and accurate communication through EEG-based text generation.

D. Developed a Multi-model

Future EEG-to-text generation advancements can extend beyond textual outputs. Although present research mostly concentrates on utilizing EEG for precise text generation, there is potential for a broader exploration of multimodality. This might involve using a single EEG signal to generate text and create complementary representations of the user's intent. However, it's important to distinguish this from directly generating images or voices from the EEG data. Instead, the focus would be on translating brain activity into additional modalities like simplified auditory representations of speech or basic visual icons that complement the textual output. This would require significant breakthroughs in deciphering the complex neural correlates not just of language processing but also of sensory information processing within the brain. Overall, multimodal EEG-to-text generation presents a fascinating future direction, offering the potential for a richer and more nuanced communication experience.

E. Cross-Domain Application

The expansion of EEG-to-text generation research into cross-domain applications represents a significant future direction, highlighting its potential beyond traditional boundaries. This approach involves applying EEG-to-text technologies across diverse domains such as healthcare, neuromarketing, and even the realm of creative arts, transcending its initial scope of aiding communication for those with disabilities. In healthcare, EEG-to-text systems could revolutionize patient care by providing non-verbal patients with a means to communicate their needs and symptoms. Moreover, in creative arts, this technology could offer a new medium for artists to translate their thoughts directly into textual form, pushing the boundaries of artistic expression. These cross-domain applications broaden the impact of EEG-to-text systems and encourage interdisciplinary collaboration, driving innovation and enhancing the depth of research in this field. To achieve successful cross-domain applications, it is important to modify the technology to align with the distinct demands and intricacies of each discipline. This task necessitates ongoing improvement and advancement of EEG-to-text systems.

V. Conclusion

The study of EEG-to-text translation represents a cutting-edge field in neuroscience and assistive technology that has demonstrated significant advancements in recent years. The progress in machine learning, namely in deep learning structures like RNNs, LSTMs, and transformer models, has established the foundation for advanced EEG decoding. DeWave and similar innovations provide a glimpse of the possibility of creating more direct and efficient communication pathways for individuals with speech or physical difficulties. The non-invasive nature of EEG makes it a prospective option for wide-scale application, given its potential to provide a means of communication for individuals who are unable to speak or write traditionally.

Future research directions indicate the need to incorporate emotional intelligence into EEG-to-text systems, broaden the range of data sources, and improve the precision and fluency of these models. The potential to use this technology in multimodal communication and other fields indicates a significant impact that may exceed its existing limitations. This extension into cross-domain applications demonstrates the versatility of EEG-to-text systems and their capacity to enhance different facets of human existence, from healthcare to creative expression.

References

[1] F. R. Willett, D. T. Avansino, L. R. Hochberg, J. M. Henderson, and K. V. Shenoy, "High-performance brain-to-text communication via handwriting," *Nature*, vol. 593, no. 7858, pp. 249–254, 2021.

[2] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for eeg-based brain–computer interfaces," *Journal of neural engineering*, vol. 4, no. 2, p. R1, 2007.

[3] C. Guger, N. F. Ince, M. Korostenskaja, and B. Z. Allison, "Brain-computer interface research: A state-of-the-art summary 11," in *Brain-Computer Interface Research: A State-of-the-Art Summary 11*. Springer, 2024, pp. 1–11.

[4] C. Chatelle, S. Laureys *et al.*, *Assessing Pain and Communication in Disorders of consciousness*. Psychology Press, 2016.

[5] M. M. Rahman, A. K. Sarkar, M. A. Hossain, M. S. Hossain, M. R. Islam, M. B. Hossain, J. M. Quinn, and M. A. Moni, "Recognition of human emotions using eeg signals: A review," *Computers in Biology and Medicine*, vol. 136, p. 104696, 2021.

[6] C. Yu and M. Wang, "Survey of emotion recognition methods using eeg information," *Cognitive Robotics*, vol. 2, pp. 132–146, 2022.

[7] M. Ouchani, S. Gharibzadeh, M. Jamshidi, and M. Amini, "A review of methods of diagnosis and complexity analysis of alzheimer's disease using eeg signals," *BioMed Research International*, vol. 2021, pp. 1–15, 2021.

[8] A. Chaddad, Y. Wu, R. Kateb, and A. Bouridane, "Electroencephalography signal processing: A comprehensive review and analysis of methods and techniques," *Sensors*, vol. 23, no. 14, p. 6434, 2023.

[9] W. O. Tatum, B. A. Dworetzky, and D. L. Schomer, "Artifact and recording concepts in eeg," *Journal of clinical neurophysiology*, vol. 28, no. 3, pp. 252–263, 2011.

[10] Z. Wang and H. Ji, "Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 5, 2022, pp. 5350–5358.

[11] M. Trigka, E. Dritsas, and C. Fidas, "A survey on signal processing methods for eeg-based brain computer interface systems," in *Proceedings of the 26th Pan-Hellenic Conference on Informatics*, 2022, pp. 213–218.

[12] M. Rashid, N. Sulaiman, A. PP Abdul Majeed, R. M. Musa, A. F. Ab Nasir, B. S. Bari, and S. Khatun, "Current status, challenges, and possible solutions of eeg-based brain-computer interface: a comprehensive review," *Frontiers in neurorobotics*, p. 25, 2020.

[13] M. Orban, M. Elsamanty, K. Guo, S. Zhang, and H. Yang, "A review of brain activity and eeg-based brain–computer interfaces for rehabilitation application," *Bioengineering*, vol. 9, no. 12, p. 768, 2022.

[14] X. Xu, M. Lin, and T. Xu, "Epilepsy seizures prediction based on nonlinear features of eeg signal and gradient boosting decision tree," *International Journal of Environmental Research and Public Health*, vol. 19, no. 18, p. 11326, 2022.

[15] L. Hu and Z. Zhang, "Eeg signal processing and feature extraction," 2019.

[16] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Muller, "Optimizing spatial filters for robust eeg single-trial analysis," *IEEE Signal processing magazine*, vol. 25, no. 1, pp. 41–56, 2007.

[17] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update," *Journal of neural engineering*, vol. 15, no. 3, p. 031005, 2018.

[18] J. Van Erp, F. Lotte, and M. Tangermann, "Brain-computer interfaces: beyond medical applications," *Computer*, vol. 45, no. 4, pp. 26–34, 2012.

[19] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update," *Journal of neural engineering*, vol. 15, no. 3, p. 031005, 2018.

[20] K. G. Hartmann, R. T. Schirrmeister, and T. Ball, "Eeg-gan: Generative adversarial networks for electroencephalograhic (eeg) brain signals," *arXiv preprint arXiv:1806.01875*, 2018.

[21] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[22] J. Van Erp, F. Lotte, and M. Tangermann, "Brain-computer interfaces: beyond medical applications," *Computer*, vol. 45, no. 4, pp. 26–34, 2012.

[23] U. Chaudhary, N. Birbaumer, and A. Ramos-Murguialday, "Brain–computer interfaces for communication and rehabilitation," *Nature Reviews Neurology*, vol. 12, no. 9, pp. 513–525, 2016.

[24] M. A. Lopez-Gordo, D. Sanchez-Morillo, and F. P. Valle, "Dry eeg electrodes," *Sensors*, vol. 14, no. 7, pp. 12 847–12 870, 2014.

[25] J. Minguillon, M. A. Lopez-Gordo, and F. Pelayo, "Trends in eeg-bci for daily-life: Requirements for artifact removal," *Biomedical Signal Processing and Control*, vol. 31, pp. 407–418, 2017.

[26] J. Van Erp, F. Lotte, and M. Tangermann, "Brain-computer interfaces: beyond medical applications," *Computer*, vol. 45, no. 4, pp. 26–34, 2012.

[27] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update," *Journal of neural engineering*, vol. 15, no. 3, p. 031005, 2018.

[28] C. Vidaurre and B. Blankertz, "Towards a cure for bci illiteracy," *Brain topography*, vol. 23, pp. 194–198, 2010.

[29] V. Jayaram, M. Alamgir, Y. Altun, B. Scholkopf, and M. Grosse-Wentrup, "Transfer learning in brain-computer interfaces," *IEEE Computational Intelligence Magazine*, vol. 11, no. 1, pp. 20–31, 2016.

[30] S. Makeig, M. Westerfield, T.-P. Jung, S. Enghoff, J. Townsend, E. Courchesne, and T. J. Sejnowski, "Dynamic brain sources of visual evoked responses," *Science*, vol. 295, no. 5555, pp. 690–694, 2002.

[31] D. J. Krusienski, M. Grosse-Wentrup, F. Galán, D. Coyle, K. J. Miller, E. Forney, and C. W. Anderson, "Critical issues in state-of-the-art brain–computer interface signal processing," *Journal of neural engineering*, vol. 8, no. 2, p. 025002, 2011.

[32] J. Mak, Y. Arbel, J. W. Minett, L. M. McCane, B. Yuksel, D. Ryan, D. Thompson, L. Bianchi, and D. Erdogmus, "Optimizing the p300-based brain–computer interface: current status, limitations and future directions," *Journal of neural engineering*, vol. 8, no. 2, p. 025003, 2011.

[33] G. Mecacci and P. Haselager, "Identifying criteria for the evaluation of the implications of brain reading for mental privacy," *Science and Engineering Ethics*, vol. 25, pp. 443–461, 2019.

[34] T. Bonaci, J. Herron, C. Matlack, and H. J. Chizeck, "Securing the exocortex: A twenty-first century cybernetics challenge," in *2014 IEEE conference on Norbert Wiener in the 21st century (21CW)*. IEEE, 2014, pp. 1–8.

[35] S. Wachter, B. Mittelstadt, and L. Floridi, "Transparent, explainable, and accountable ai for robotics," *Science robotics*, vol. 2, no. 6, p. eaan6080, 2017.

[36] M. Ienca and R. Andorno, "Towards new human rights in the age of neuroscience and neurotechnology," *Life sciences, society and policy*, vol. 13, pp. 1–27, 2017.

[37] P. Kellmeyer, T. Cochrane, O. Müller, C. Mitchell, T. Ball, J. J. Fins, and N. Biller-Andorno, "The effects of closed-loop medical devices on the autonomy and accountability of persons and systems," *Cambridge Quarterly of Healthcare Ethics*, vol. 25, no. 4, pp. 623–633, 2016.

[38] S. Schicktanz, T. Amelung, and J. W. Rieger, "Qualitative assessment of patients' attitudes and expectations toward bcis and implications for future technology development," *Frontiers in systems neuroscience*, vol. 9, p. 64, 2015.

[39] G. Wolbring, L. Diep, S. Yumakulov, N. Ball, and D. Yergens, "Social robots, brain machine interfaces and neuro/cognitive enhancers: Three emerging science and technology products through the lens of technology acceptance theories, models and frameworks," *Technologies*, vol. 1, no. 1, pp. 3–25, 2013.

[40] N. Hollenstein, J. Rotsztejn, M. Troendle, A. Pedroni, C. Zhang, and N. Langer, "Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading," *Scientific data*, vol. 5, no. 1, pp. 1–13, 2018.

[41] N. Hollenstein, M. Troendle, C. Zhang, and N. Langer, "Zuco 2.0: A dataset of physiological recordings during natural reading and annotation," *arXiv preprint arXiv:1912.00903*, 2019.

[42] F. Pereira, B. Lou, B. Pritchett, S. Ritter, S. J. Gershman, N. Kanwisher, M. Botvinick, and E. Fedorenko, "Toward a universal decoder of linguistic meaning from brain activation," *Nature communications*, vol. 9, no. 1, p. 963, 2018.

[43] N. Affolter, B. Egressy, D. Pascual, and R. Wattenhofer, "Brain2word: decoding brain activity for language generation," *arXiv preprint arXiv:2009.04765*, 2020.

[44] C. Herff, D. Heger, A. De Pesters, D. Telaar, P. Brunner, G. Schalk, and T. Schultz, "Brain-to-text: decoding spoken phrases from phone representations in the brain," *Frontiers in neuroscience*, vol. 9, p. 217, 2015.

[45] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, 2019.

[46] P. Wang, R. Zhou, S. Wang, L. Li, W. Bai, J. Fan, C. Li, P. Childs, and Y. Guo, "A general framework for revealing human mind with auto-encoding gans," *arXiv preprint arXiv:2102.05236*, 2021.

[47] P. Kumar, R. Saini, P. P. Roy, P. K. Sahu, and D. P. Dogra, "Envisioned speech recognition using eeg sensors," *Personal and Ubiquitous Computing*, vol. 22, pp. 185–199, 2018.

[48] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, and M. Shah, "Deep learning human mind for automated visual classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6809–6817.

[49] G. Shen, T. Horikawa, K. Majima, and Y. Kamitani, "Deep image reconstruction from human brain activity," *PLoS computational biology*, vol. 15, no. 1, p. e1006633, 2019.

[50] K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant, "Identifying natural images from human brain activity," *Nature*, vol. 452, no. 7185, pp. 352–355, 2008.

[51] S. Lin, T. Sprague, and A. K. Singh, "Mind reader: Reconstructing complex images from brain activities," *Advances in Neural Information Processing Systems*, vol. 35, pp. 29 624–29 636, 2022.

[52] "Compumedics Neuroscan – World Leader in Functional Neuroimaging." [Online]. Available: https://compumedicsneuroscan.com/

[53] S. Sauer, "Brain Products GmbH | Solutions for neurophysiological research." [Online]. Available: https://www.brainproducts.com/

[54] "Biosemi EEG ECG EMG BSPM NEURO amplifier electrodes." [Online]. Available: https://www.biosemi.com/

[55] "Homepage." [Online]. Available: https://www.emotiv.com/

[56] "EEG - ECG - Biosensors." [Online]. Available: https://neurosky.com/

[57] "ANT Neuro | inspiring technology for the human brain." [Online]. Available: https://www.ant-neuro.com/

[58] "Advanced Brain Monitoring." [Online]. Available: https://www.advancedbrainmonitoring.com/

[59] "OpenBCI Featured Products." [Online]. Available: https://shop.openbci.com/collections/frontpage

[60] G. Di Flumeri, P. Aricò, G. Borghini, A. Colosimo, and F. Babiloni, "A new regression-based method for the eye blinks artifacts correction in the eeg signal, without using any eog channel," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 3187–3190.

[61] N. Ille, Y. Nakao, Y. Shumpei, T. Taura, A. Ebert, H. Bornfleth, S. Asagi, K. Kozawa, I. Itabashi, T. Sato *et al.*, "Ongoing eeg artifact correction using blind source separation," *Clinical Neurophysiology*, 2024.

[62] M. Dora and D. Holcman, "Adaptive single-channel eeg artifact removal with applications to clinical monitoring," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 286–295, 2022.

[63] M. K. Islam, P. Ghorbanzadeh, and A. Rastegarnia, "Probability mapping based artifact detection and removal from single-channel eeg signals for brain–computer interface applications," *Journal of Neuroscience Methods*, vol. 360, p. 109249, 2021.

[64] Y. Dai, F. Duan, F. Feng, Z. Sun, Y. Zhang, C. F. Caiafa, P. Marti-Puig, and J. Solé-Casals, "A fast approach to removing muscle artifacts for eeg with signal serialization based ensemble empirical mode decomposition," *Entropy*, vol. 23, no. 9, p. 1170, 2021.

[65] C. Dora and P. K. Biswal, "An improved algorithm for efficient ocular artifact suppression from frontal eeg electrodes using vmd," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 1, pp. 148–161, 2020.

[66] H. Massar, M. Miyara, T. Belhoussine Drissi, and B. Nsiri, "An integrated approach for artifact elimination in eeg signals: Combining variational mode decomposition with blind source separation (vmd-bss)," in *The International Conference on Artificial Intelligence and Smart Environment*. Springer, 2023, pp. 84–90.

[67] L. Shoker, S. Sanei, and J. Chambers, "Artifact removal from electroencephalograms using a hybrid bss-svm algorithm," *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 721–724, 2005.

[68] M. B. Hamaneh, N. Chitravas, K. Kaiboriboon, S. D. Lhatoo, and K. A. Loparo, "Automated removal of ekg artifact from eeg data using independent component analysis and continuous wavelet transformation," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1634–1641, 2013.

[69] M. A. Klados, C. Papadelis, C. Braun, and P. D. Bamidis, "Reg-ica: a hybrid methodology combining blind source separation and regression techniques for the rejection of ocular artifacts," *Biomedical Signal Processing and Control*, vol. 6, no. 3, pp. 291–300, 2011.

[70] G. K. Soni, H. Singh, H. Arora, and A. Soni, "Ultra low power cmos low pass filter for biomedical ecg/eeg application," in *2020 Fourth International Conference on Inventive Systems and Control (ICISC)*. IEEE, 2020, pp. 558–561.

[71] I. Winkler, S. Debener, K.-R. Müller, and M. Tangermann, "On the influence of high-pass filtering on ica-based artifact reduction in eeg-erp," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 4101–4105.

[72] J. Baranowski and P. Piątek, "Fractional band-pass filters: design, implementation and application to eeg signal processing," *Journal of Circuits, Systems and Computers*, vol. 26, no. 11, p. 1750170, 2017.

[73] W. Wang, G. Zhang, L. Yang, V. Balaji, V. Elamaran, and N. Arunkumar, "Revisiting signal processing with spectrogram analysis on eeg, ecg and speech signals," *Future Generation Computer Systems*, vol. 98, pp. 227–232, 2019.

[74] S. Seifzadeh, K. Faez, and M. Amiri, "Comparison of different linear filter design methods for handling ocular artifacts in brain computer interface system," *Journal of Computer & Robotics*, vol. 7, no. 1, pp. 51–56, 2014.

[75] M. K. Ahirwal, A. Kumar, G. K. Singh, and N. D. Londhe, "Performance prediction of adaptive filters for eeg signal," *IET Science, Measurement & Technology*, vol. 11, no. 5, pp. 525–531, 2017.

[76] M. Z. Baig, Y. Mehmood, and Y. Ayaz, "A bci system classification technique using median filtering and wavelet transform," in *Dynamics in Logistics: Proceedings of the 4th International Conference LDIC, 2014 Bremen, Germany*. Springer, 2016, pp. 355–364.

[77] T. A. Camilleri, K. P. Camilleri, and S. G. Fabri, "Segmentation and labelling of eeg for brain computer interfaces," in *Computer Analysis of Images and Patterns: 16th International Conference, CAIP 2015, Valletta, Malta, September 2-4, 2015 Proceedings, Part I 16*. Springer, 2015, pp. 288–299.

[78] P. C. Sharma, R. Raja, S. K. Vishwakarma, S. Sharma, P. K. Mishra, and V. S. Kushwah, "Analysis of brain signal processing and real-time eeg signal enhancement," *Multimedia Tools and Applications*, vol. 81, no. 28, pp. 41 013–41 033, 2022.

[79] T. Parviainen and J. Kujala, "Event-related potentials (erps) and event-related fields (erfs)," in *Language Electrified: Principles, Methods, and Future Perspectives of Investigation*. Springer, 2023, pp. 195–239.

[80] Y. Tran, "Eeg signal processing for biomedical applications," p. 9754, 2022.

[81] S. M. Qaisar, "A computationally efficient eeg signals segmentation and de-noising based on an adaptive rate acquisition and processing," in *2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP)*. IEEE, 2018, pp. 182–186.

[82] A. Rastogi and V. Bhateja, "Pre-processing of electroencephalography signals using stationary wavelet transform-enhanced fixed-point fast-ica," in *Data Engineering and Intelligent Computing: Proceedings of ICICC 2020*. Springer, 2021, pp. 387–396.

[83] C. M. Segning, J. Harvey, H. Ezzaidi, K. B. P. Fernandes, R. A. da Silva, and S. Ngomo, "Towards the objective identification of the presence of pain based on electroencephalography signals' analysis: A proof-of-concept," *Sensors*, vol. 22, no. 16, p. 6272, 2022.

[84] R. Zhang, P. Xu, L. Guo, Y. Zhang, P. Li, and D. Yao, "Z-score linear discriminant analysis for eeg based brain-computer interfaces," *PloS one*, vol. 8, no. 9, p. e74433, 2013.

[85] A. Apicella, F. Isgrò, A. Pollastro, and R. Prevete, "On the effects of data normalization for domain adaptation on eeg data," *Engineering Applications of Artificial Intelligence*, vol. 123, p. 106205, 2023.

[86] Garima, N. Goel, and N. Rathee, "Modified multidimensional scaling on eeg signals for emotion classification," *Multimedia Tools and Applications*, pp. 1–22, 2023.

[87] H. Zhang, Q.-Q. Zhou, H. Chen, X.-Q. Hu, W.-G. Li, Y. Bai, J.-X. Han, Y. Wang, Z.-H. Liang, D. Chen *et al.*, "The applied principles of eeg analysis methods in neuroscience and clinical neurology," *Military Medical Research*, vol. 10, no. 1, p. 67, 2023.

[88] D. Hernández, L. Trujillo, E. Z-Flores, O. Villanueva, and O. Romo-Fewell, "Detecting epilepsy in eeg signals using time, frequency and time-frequency domain features," *Computer science and engineering—theory and applications*, pp. 167–182, 2018.

[89] C. Wang, A. K. Verma, B. Guragain, X. Xiong, and C. Liu, "Classification of bruxism based on time-frequency and nonlinear features of single channel eeg," *BMC Oral Health*, vol. 24, no. 1, p. 81, 2024.

[90] A. S. Al-Fahoum and A. A. Al-Fraihat, "Methods of eeg signal features extraction using linear analysis in frequency and time-frequency domains," *International Scholarly Research Notices*, vol. 2014, 2014.

[91] T. Wen and Z. Zhang, "Effective and extensible feature extraction method using genetic algorithm-based frequency-domain feature search for epileptic eeg multiclassification," *Medicine*, vol. 96, no. 19, 2017.

[92] H. Li, M. Liu, X. Yu, J. Zhu, C. Wang, X. Chen, C. Feng, J. Leng, Y. Zhang, and F. Xu, "Coherence based graph convolution network for motor imagery-induced eeg after spinal cord injury," *Frontiers in Neuroscience*, vol. 16, p. 1097660, 2023.

[93] S. Chen, Z. Luo, and H. Gan, "An entropy fusion method for feature extraction of eeg," *Neural Computing and Applications*, vol. 29, pp. 857–863, 2018.

[94] S. Ansarinasab, F. Ghassemi, Z. Tabanfar, and S. Jafari, "Investigation of phase synchronization in functional brain networks of children with adhd using nonlinear recurrence measure," *Journal of Theoretical Biology*, vol. 560, p. 111381, 2023.

[95] K. Makkar and A. Bisen, "Eeg signal processing and feature extraction," *International Journal for Modern Trends in Science and Technology*, vol. 9, no. 08, pp. 45–50, 2023.

[96] Q. Wei, Y. Wang, X. Gao, and S. Gao, "Amplitude and phase coupling measures for feature extraction in an eeg-based brain–computer interface," *Journal of neural engineering*, vol. 4, no. 2, p. 120, 2007.

[97] M. N. Alam, M. I. Ibrahimy, and S. Motakabber, "Feature extraction of eeg signal by power spectral density for motor imagery based bci," in *2021 8th International Conference on Computer and Communication Engineering (ICCCE)*. IEEE, 2021, pp. 234–237.

[98] M. Singh and R. Goyat, "Feature extraction for the analysis of multi-channel eeg signals using hilbert-huang technique," *International Journal of Engineering and Technology*, vol. 8, no. 1, pp. 17–27, 2016.

[99] A. S. Al-Fahoum and A. A. Al-Fraihat, "Methods of eeg signal features extraction using linear analysis in frequency and time-frequency domains," *International Scholarly Research Notices*, vol. 2014, 2014.

[100] S.-H. Oh, Y.-R. Lee, and H.-N. Kim, "A novel eeg feature extraction method using hjorth parameter," *International Journal of Electronics and Electrical Engineering*, vol. 2, no. 2, pp. 106–110, 2014.

[101] A. Patil, C. Deshmukh, and A. Panat, "Feature extraction of eeg for emotion recognition using hjorth features and higher order crossings," in *2016 Conference on Advances in Signal Processing (CASP)*. IEEE, 2016, pp. 429–434.

[102] L. I. Kuncheva and J. J. Rodríguez, "Interval feature extraction for classification of event-related potentials (erp) in eeg data analysis," *Progress in Artificial Intelligence*, vol. 2, pp. 65–72, 2013.

[103] M. M. Ali, M. Taib, N. M. Tahir, and A. Jahidin, "Eeg spectral centroid amplitude and band power features: A correlation analysis," in *2014 IEEE 5th Control and System Graduate Research Colloquium*. IEEE, 2014, pp. 223–226.

[104] E. Wang, L. Wang, C. Ye, N. Luo, Y. Zhang, Y. Zhong, M. Zhu, Y. Zou, Q. Li, L. Li *et al.*, "Effect of electroencephalography spectral edge frequency (sef) and patient state index (psi)-guided propofol-remifentanil anesthesia on delirium after laparoscopic surgery: the emodipod randomized controlled trial," *Journal of Neurosurgical Anesthesiology*, vol. 34, no. 2, pp. 183–192, 2022.

[105] T.-j. Luo, "Parallel genetic algorithm based common spatial patterns selection on time–frequency decomposed eeg signals for motor imagery brain-computer interface," *Biomedical Signal Processing and Control*, vol. 80, p. 104397, 2023.

[106] K. Saranya and M. Paulraj, "Certain investigation on eeg signal processing using auto regression feature for various colour stimuli," in *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*. IEEE, 2023, pp. 247–252.

[107] J.-H. Kang, C. H. Lee, and S.-P. Kim, "Eeg feature selection and the use of lyapunov exponents for eeg-based biometrics," in *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2016, pp. 228–231.

[108] S. Geng, W. Zhou, Q. Yuan, D. Cai, and Y. Zeng, "Eeg non-linear feature extraction using correlation dimension and hurst exponent," *Neurological research*, vol. 33, no. 9, pp. 908–912, 2011.

[109] M. Aboy, R. Hornero, D. Abásolo, and D. Álvarez, "Interpretation of the lempel-ziv complexity measure in the context of biomedical signal analysis," *IEEE transactions on biomedical engineering*, vol. 53, no. 11, pp. 2282–2288, 2006.

[110] A. Goshvarpour, A. Abbasi, and A. Goshvarpour, "Recurrence quantification analysis and neural networks for emotional eeg classification," *Applied Medical Informatics*, vol. 38, no. 1, pp. 13–24, 2016.

[111] C. Guerrero-Mosquera, M. Verleysen, and A. N. Vazquez, "Eeg feature selection using mutual information and support vector machine: A comparative analysis," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 2010, pp. 4946–4949.

[112] Y. Duan, J. Zhou, Z. Wang, Y.-K. Wang, and C.-T. Lin, "Dewave: Discrete eeg waves encoding for brain dynamics to text translation," *arXiv preprint arXiv:2309.14030*, 2023.

[113] X. Zhang, L. Yao, Q. Z. Sheng, S. S. Kanhere, T. Gu, and D. Zhang, "Converting your thoughts to texts: Enabling brain typing via deep feature learning of eeg signals," in *2018 IEEE international conference on pervasive computing and communications (PerCom)*. IEEE, 2018, pp. 1–10.

[114] J. Yang, M. Awais, A. Hossain, L. Yee, M. Haowei, I. M. Mehedi, and A. Iskanderani, "Thoughts of brain eeg signal-to-text conversion using weighted feature fusion-based multiscale dilated adaptive densenet with attention mechanism," *Biomedical Signal Processing and Control*, vol. 86, p. 105120, 2023.

MAD: Multi-Alignment MEG-to-Text Decoding

Yiqian Yang[1*] **Hyejeong Jo** [2*] **Yiqun Duan** [3*] **Qiang Zhang**[1] **Jinni Zhou**[1]
Won Hee Lee[2†] **Renjing Xu**[1†] **Hui Xiong**[1†]

Abstract

Deciphering language from brain activity is a crucial task in brain-computer interface (BCI) research. Non-invasive cerebral signaling techniques including electroencephalography (EEG) and magnetoencephalography (MEG) are becoming increasingly popular due to their safety and practicality, avoiding invasive electrode implantation. However, current works under-investigated three points: 1) a predominant focus on EEG with limited exploration of MEG, which provides superior signal quality; 2) poor performance on unseen text, indicating the need for models that can better generalize to diverse linguistic contexts; 3) insufficient integration of information from other modalities, which could potentially constrain our capacity to comprehensively understand the intricate dynamics of brain activity.

This study presents a novel approach for translating MEG signals into text using a speech-decoding framework with multiple alignments. Our method is the first to introduce an end-to-end multi-alignment framework for totally unseen text generation directly from MEG signals. We achieve an impressive BLEU-1 score on the *GWilliams* dataset, significantly outperforming the baseline from 5.49 to 10.44 on the BLEU-1 metric. This improvement demonstrates the advancement of our model towards real-world applications and underscores its potential in advancing BCI research. Code is available at https://github.com/NeuSpeech/MAD-MEG2text.

1 Introduction

Decoding brain to language has emerged as a rapidly developing area of neurotechnology, offering semantic communication and control for general Brain-Computer-Interface (BCI) tasks. This region has garnered growing focus as it may profoundly impact individuals with verbal and movement disabilities resulting from conditions such as severe spinal cord trauma or end-stage amyotrophic lateral sclerosis (ALS). Moreover, the scope of brain-to-text technology extends to pioneer novel human-machine interfaces, allowing seamless control of prosthetic limbs, software, and virtual environments, shifting the paradigm of interaction for both able-bodied individuals and those with disabilities, and re-defining what is achievable in both everyday life and professional spheres.

[*]These authors contributed equally to this work

[†]These are corresponding authors

[1]The Hong Kong University of Science and Technology (Guangzhou), People's Republic of China,
Yiqian Yang and Hui Xiong are with AI Thrust, HKUST(GZ).
Qiang Zhang and Renjing Xu are with MICS Thrust, HKUST(GZ),
Jinni Zhou is with RBM Base, College of Future Technology,
Email: yyang937@connect.hkust-gz.edu.cn, xionghui@hkust-gz.edu.cn, qzhang749@connect.hkust-gz.edu.cn, renjingxu@hkust-gz.edu.cn, eejinni@hkust-gz.edu.cn

[2]Department of Software Convergence, Kyung Hee University, Republic of Korea,
Email:girlsending0@khu.ac.kr, whlee@khu.ac.kr

[3]GrapheneX-UTS HAI Centre, Australia Artificial Intelligence Institute, University of Technology Sydney, Australia
Email:duanyiquncc@gmail.com

Preprint. Under review.

Under this scope, various previous works have explored this area in multiple ways. Pioneer researchers first verify this idea by using invasive signals such as Electrocorticography (ECoG) [1, 2, 3, 4]. Recently, these invasive methods [5, 6] concentrate on decoding speech, phonemes or letter from ECoG signals and have achieved remarkably high accuracy using limited word sets for real-time brain-to-text translation. However, these invasive-signal-based approaches pose significant medical risks and challenges for long-term use.

Non-invasive techniques, therefore, present a safer and more sustainable alternative, albeit with their own set of challenges. Wang et al. [7] showcased a method for translating EEG signals into text with an extensive lexicon, utilizing language models that had been pre-trained on EEG data features at word-level. Duan et al. [8] progressed this methodology by interpreting raw EEG signals directly, devoid of reliance on temporal indicators, but their models still relied heavily on teacher forcing for evaluation, limiting their ability to generate meaningful sentences autonomously in real-life scenarios. At the same time, although Magnetoencephalography (MEG) provides better signal quality, previous works [9, 10, 11] on MEG have primarily focused on decoding limited classes or short phrases from MEG signals, showing limited success in generating whole sentences and complete semantic segments.

Furthermore, as pointed out by Jo et al. [12], all previous works in EEG-to-Text translation following Wang's method [7] meets the "decoder dominated" problem. It means that given a strong decoder and noisy EEG input, these models are more likely to memorize the text distribution corresponding to certain statistical features rather than mapping EEG to semantic texts. Thus, these models have similar performances even when we replace EEG input with random noise. Besides, due to the nature of limited data and the non-understandability of the neural signal, it is difficult to train and evaluate the model. Yang et al. [13] proposed NeuSpeech model on MEG to text task, however, their model is evaluated on the text that is seen in the training set, which does not meet the need for open-vocabulary translation. Defossez et al. [14] highlighted the potential to decode speech perception from MEG signals, where they matched MEG signals with corresponding speech segments. However, their approach was limited to classification tasks and could not generate sentences directly from MEG signals. This underscores a significant gap in the current state of MEG-based brain-to-text decoding.

In this paper, our motivation is to establish an end-to-end framework for open-vocabulary MEG-to-Text translation capable of processing unseen text without relying on biomarkers, while ensuring that the encoder captures brain dynamics effectively. We propose Multi-Alignment MEG-to-Text Decoding (MAD) with the aim of guiding the brain encoders towards learning salient representations. To achieve this, we incorporate audio as an auxiliary modality to facilitate alignment. Here, we make a bold assumption that directly formatting noise brain signals into discrete text is difficult due to limited data. Hence, we utilize brain module [14] and an extra whisper model [15] to align brain representation in three aspects as shown in Figure 1, the Mel spectrogram, hidden state, and text. 1) We first align the Brain module with audio in the Mel spectrogram feature space to learn low-level features, such as acoustic features. 2) Additionally, we align the hidden state output from both the whisper encoder and the brain module in latent space, enhancing the model's ability to extract high-level semantic features. 3) Lastly, we align the text representation from both streams within the framework.

Our objective in incorporating textual data is to assess whether it can furnish supplementary contextual cues that enhance the correspondence between neural activity and the resulting linguistic output.

Comprehensive experiments are conducted by utilizing non-invasive public MEG data from *GWilliams* [16] dataset, which captured MEG signals during a speech listening task. Remarkably, MAD **is capable of generalizing to unseen text**. Performance is evaluated using translation text relevancy metrics [17, 18]. On raw MEG waves, MAD achieves 10.44 BLEU-1 on *GWilliams* **without teacher-forcing** evaluation on **entirely unseen text** which largely exceeds the current SOTA performance. This paper also provides insights though numerous ablation studies to help people understand the impact of each component on aligning the MEG signal with texts. The contributions of this research could be summarized as follows:

- MAD presents an end-to-end neural network design for the direct conversion of MEG signals into text in open-vocabulary, obviating the dependence on markers, teacher forcing, or pre-training, representing the initial implementation of translating raw MEG waves into text for unseen content.

- We are the first to investigate various alignments and demonstrate the benefits of aligning with speech modality rather than text modality in the MEG-to-text transcription task, offering significant insights for network improvement.
- Our extensive experimentation and thorough analysis of the proposed model showcase its effectiveness and highlight its superiority over existing methods in terms of translation accuracy, efficiency ,and reliability.

2 Related Works

The discipline of converting brain signals into textual output has undergone considerable development in the contemporary era. In 2019, Anumanchipalli et al. [1] introduced a pioneering model capable of translating ECoG patterns into the articulatory movements necessary for speech production, subsequently generating acoustic properties such as MFCCs, leading to the production of intelligible speech. This landmark study ignited further exploration within the field. In the subsequent year, Wang et al. [2] leveraged the capabilities of generative adversarial networks (GANs) to decipher ECoG data and synthesize speech. The year following, Willett et al. [3] engineered a system that utilized a recurrent neural network (RNN) alongside a probabilistic language model to decode letters from neural activity during the act of handwriting. Most recently, Metzger et al. [19] constructed a sequence of processes that converted ECoG signals into textual information using an RNN, enhancing the results with the GPT-2 language model.

Within the domain of open-vocabulary interpretation, Metzger et al. [6] unveiled an RNN architecture capable of real-time decoding of speech, text, sentiment, and facial expressions from ECoG data. Simultaneously, Willett et al. [5] managed to interpret text directly from neural activity. Liu et al. [20] introduced a tripartite model designed to decode logo-syllabic languages, such as Chinese, by transforming ECoG signals into Chinese pinyin inclusive of tones and syllables, followed by speech synthesis. In a related development, Feng et al. [21] achieved text interpretation from SEEG recordings. It is essential to highlight that these functional systems are predominantly reliant on invasive neural recordings.

In the domain of non-invasive neural recording, Meta unveiled a brain-to-speech system that leverages contrastive learning with MEG and EEG data [14]. While this system is proficient in categorizing a constrained set of sentences, it is not conducive to open-vocabulary textual interpretation. Ghazaryan et al. [11] explored the decoding of a restricted vocabulary from MEG responses. Wang et al. [7] crafted a mechanism for translating EEG features at the word level into text, employing a pre-trained BART model [22]. Subsequent investigations, including Dewave [8], adopted the methodology established by Wang et al. [7], proposing a schema that incorporates wave2vec [23] and discrete codex for robust representations, which are subsequently funneled into a BART [22] model for text synthesis. These approaches, however, are dependent on teacher-forcing and disregard the necessity of comparing results with noise-injected inputs, potentially resulting in an inflated assessment of system efficacy. Recent scholarship [12] has revealed the limitations of these methods.

Yang et al. [13] proposed an end-to-end paradigm for converting MEG signals to text, demonstrating high performance when training and evaluation sets were fully overlapped. However, it does not show good performance on unseen text. Our approach diverges from these methods by employing transfer learning with assistance of extra modality(Mel spectrogram) to align the model through multiple stages with low-level and high-level features of the ground truth. This enables our model to learn more effectively and generalize better to unseen text.

3 Method

3.1 Task Definition

Given a sequence of raw segment-level MEG signals ε, the goal is to decode the associated open-vocabulary text tokens T. This task also incorporates additional information in the form of speech Ξ. The MEG-Speech-Text pairs $\langle \varepsilon, \Xi, T \rangle$ are collected during speech perception. Our approach focuses on decoding T using only the raw MEG signal ε, with the support of Ξ. MAD represents the first attempt at tackling this MEG to unseen text translation challenge.

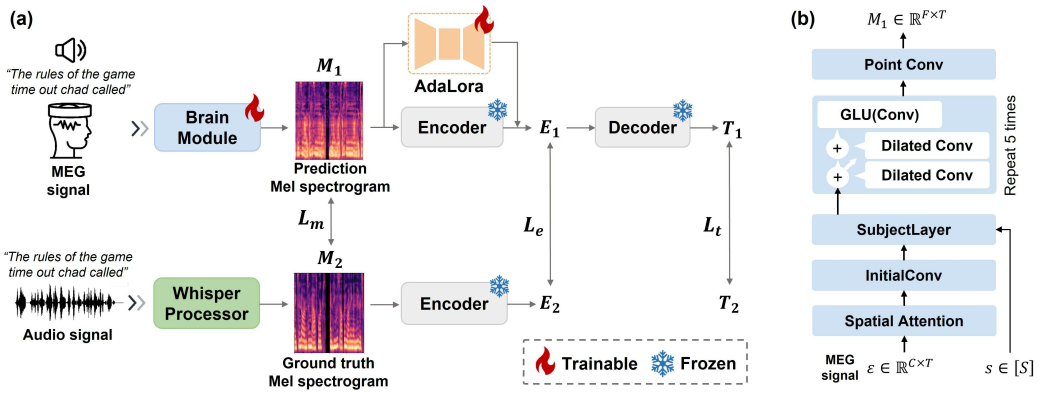


Figure 1. (a) Overview of model architecture. We added alignments on the Mel spectrogram, the hidden states, and the text. There are three types of alignment, which are either based on our physics world (text and speech) or a largely pre-trained model. M_1, M_2 is predicted and ground truth Mel spectrogram, E_1, E_2 is the hidden state of meg-input and speech-input encoder respectively. T_1, and T_2 are predicted and ground truth text respectively. (b) Details about the brain module.

3.2 Model

Figure. 1 shows the overview of our work. Our model uses some transfer learning techniques to facilitate better performance on unseen text. The encoder and decoder models are from the Whisper model [15], a transformer-based encoder-decoder architecture tailored for robust speech recognition in challenging environments such as noisy conditions. The brain module [14] first takes the MEG signal ε of C channels in the Spatial Attention layer, it adds position embedding of physical sensors to the MEG, then Initial Conv maps the MEG channel number to hidden model dimensions. After that, the Subject Layer takes the MEG feature and subject index and applies subject embedding on the MEG feature. Next, the MEG feature is input into the residual-designed module which is repeated 5 times. Finally, after Point Conv of which kernel size is 1, it maps to the Mel spectrogram M_1.

L_m is the loss to align the Mel spectrogram, which is Clip loss [24] in this situation. Then we want to make sure the encoder model can learn high-level features, so we designed to align the encoder output with L_e, which is Maximum Mean Discrepancy (MMD) loss [25]. We used LoRA [26] module to train our architecture for saving memory. Last but not least, we have the cross entropy loss L_t for predicted text and ground truth text. The overall loss L is below:

$$L = \lambda_m \cdot L_m + \lambda_e \cdot L_e + \lambda_t \cdot L_t \tag{1}$$

Recall the clip loss [24] function, it takes two feature representations from each modality. These features are then used to calculate the similarity scores between the representations of the image and text modalities. The Clip loss function aims to minimize the distance between matching pairs of image and text representations while maximizing the distance between non-matching pairs. This approach allows the CLIP model to learn a joint embedding space where semantically similar image-text pairs are close together, enabling tasks like zero-shot image classification and text-based image retrieval. Here the clip loss is applied on the Mel spectrogram, which is of 3 dimensions, so we flattened the batch size and time length dimensions as the first dimension, and then the loss is calculated as follows:

The MMD loss (Maximum Mean Discrepancy loss) is a measure of the discrepancy between two probability distributions. It is commonly used in domain adaptation and generative modeling to encourage the distributions of source and target data to be similar. If we flatten the hidden state E of the batch size n, time dimension t_d and feature dimension d_e, it will run out of memory if we input full length into the model, so we randomly select features time wise t_r, therefore the selected features is E_r shape is $[n, t_r, d_e]$ The formula for the MMD loss is:

$$\mathrm{MMD}^2(E_1, E_2) = \frac{1}{n} \left\| \sum_{i=1}^{n} \phi(E_{1r}(i)) - \sum_{i=1}^{n} \phi(E_{2r}(i)) \right\|_{\mathcal{H}}^2 \tag{2}$$

Algorithm 1: CLIP-like Loss Calculation

Input: M_1 $[n, d_m]$ Predicted Mel spectrogram ,
M_2 $[n, d_m]$ Ground truth Mel spectrogram ,
d_m Dimensionality of multimodal embedding,
t Learned temperature parameter,
n Batch size.
Output: CLIP loss

$logits \leftarrow M_1 \cdot M_2^T \cdot e^t$; `// Scaled pairwise cosine similarities, [n,n]`

$labels \leftarrow$ Range(n) ; `// Labels for each example`

$loss_1 \leftarrow$ CrossEntropyLoss$(logits, labels, \text{axis} = 0)$;

$loss_2 \leftarrow$ CrossEntropyLoss$(logits, labels, \text{axis} = 1)$;

$L_m \leftarrow$ Mean$(loss_1, loss_2)$;

return L_m;

For an Automatic Speech Recognition (ASR) system, the cross-entropy loss is commonly used as a loss function to train the model. The basic idea is similar to the general cross-entropy loss but adapted for the ASR context where the inputs are speech features and the outputs are text transcriptions. The cross-entropy loss in the context of ASR can be defined as follows:

$$\text{CrossEntropyLoss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{c=1}^{C} T_{1,i,t,c}\log(T_{2,i,t,c}) \tag{3}$$

3.3 Evaluation

We evaluate transcribing performance on *GWilliams* dataset [16] using NLP metrics, BLEU [17] is used to evaluate the accuracy of machine-translated text, ROUGE-1-F [18] is to measure the quality of automatic summarization, BertScore [27] is a measurement of semantic similarity, CER [28] is used to evaluate the accuracy of speech recognition and self-BLEU [29] is used to assess the diversity of generated text.

1. BLEU is used to evaluate the accuracy of machine-translated text.
2. ROUGE-1-F is to measure the quality of automatic summarization.
3. BertScore is a measurement of semantic similarity.
4. CER is used to evaluate the accuracy of speech recognition.
5. Self-BLEU is used to assess the diversity of generated text.

4 Experiments

4.1 Dataset

The GWilliams dataset [16] is a magnetoencephalography (MEG) dataset designed for assessing natural speech comprehension. It features authentic MEG recordings from 27 participants proficient in English. These participants engaged in two separate sessions, each involving two hours of listening to four stories, which are “cable spool fort",“easy money",“lw1",“the black willow". To get a fair evaluation, we split our dataset directly on stories, we test on “cable spool fort", validate on “lw1" and train on other stories. Details are in Table. 1. For more details about the dataset, please refer to Supp. A.

For preprocessing, we used first band pass filter the MEG signal ε between 1 Hz and 40 Hz, then it is resampled to 100Hz to reduce computing. We ensure that we separated training, evaluation, testing set totally since we used one story for testing, another story for evaluation, last two ones for training. We extract 4-second windows from the MEG-speech-text pairs, sliding every second and randomly shifting the window by ±0.5 seconds to generate samples. Speech Ξ is then transformed to Mel M with window length of 400, hop length of 160, which is the original configuration in Whisper

Table 1. Details about the dataset splits, we ensured the three splits are totally separated. Unique sentences means the sentences that are different with other sentences, same meaning for unique words. There is no overlap sentence between train and test set. 371(46%) means 371 words in test set is also in train set, accounting for 46 percentage.

Split	Segments	Unique sentences	Words	Unique words	Overlap sentence	Overlap words
train	133966	13266	150497	2776	-	-
validation	14896	1387	156027	478	-	-
test	31115	3151	355654	805	0	371(46%)

Table 2. Comparison with other models. Lo is LoRA, B is brain module. Bert here means Bertscore. Results is obtained without teacher forcing in evaluation. Here, Tr stands for trainable modules. B-1 stands for BLEU-1. R-1 stands for ROUGE-1-F. SB stands for Self-BLEU. RS means randomly selecting sentences from test set as predictions. As we can see, only MAD is much higher than RS on BLEU-1 score.

Modality	Method	Tr	Loss	B-1(%)↑	R-1 (%)↑	Bert(%)↑	CER(%)↓	SB(%)↓
-	RS	-	-	5.86	7.20	83.73	87.30	96.12
MEG	NeuSpeech [13]	Lo	L_t	5.49	8.43	83.98	77.02	99.7
MEG	Wav2vec2CTC [14]	B	L_m	0.55	1.44	76.02	152.23	92.67
MEG	MAD	B	$L_m + L_e$	10.44	6.93	83.39	89.82	85.66
Noise	MAD	B	$L_m + L_e$	3.87	3.16	83.20	126.95	87.54
MEG	MAD w/tf	B	$L_m + L_e$	12.93	18.28	82.87	74.31	83.35
Noise	MAD w/tf	B	$L_m + L_e$	0.19	6.68	59.92	87.57	68.63

model [15], since the setted speech sampling rate is 16kHz, after conversion, M is of shape [400, 80] time and feature wise for 4 second speech, then it is matched with ε of time length 400.

4.2 Implementation details

All models were trained using Nvidia 4090 (24GB) GPUs. Training was conducted with a learning rate of 3e-4 and a batch size of 32 over 5 epochs, selecting the best-performing model based on evaluation loss. AdamW was employed as the optimizer across all models. Each experiment takes about 18 hours on signal GPU with 8 workers to finish. Lambda value in all experiment on MAD model set as follows: λ_m = 1, λ_e = 0.01, λ_t = 1.

4.3 Evaluation Metrics

The performance comparison of our proposed MAD model with other state-of-the-art models is summarized in Table 2. The table highlights various configurations and the corresponding evaluation metrics: BLEU-1, ROUGE-1, BertScore, CER and self-BLEU. Each model's performance is evaluated on MEG data, with results illustrating the impact of different loss functions and modules on decoding accuracy.

We compare the performance of our proposed model, MAD, against existing state-of-the-art methods, NeuSpeech [13] and Wav2vec2CTC [14], for decoding MEG signals into text. The performance metrics used for evaluation include BLEU-1, ROUGE-1-F, Bertscore, and Character Error Rate (CER). The results are summarized in Table 2. We find out BLEU-1 seems to be the most effective measurement in this situation.

NeuSpeech [13] is a encoder-decoder framework model used for MEG, utilizing the Low-Rank Adaptation (LoRA) method with a text-based loss (L_t), achieves best scores on ROUGE-1-F, BertScore, and CER. However, the self-bleu score is almost 100%, which means the generation always repeat same thing. Besides, the BLEU-1 score is lower than RS, which means these three metrics are not reliable, which is further discussed in Supp. B.

Wav2vec2CTC [14]: The original model predicts the output of the Wav2vec2 [23] encoder with brain module. We add the pretrained language model head in the Wav2vec2CTC [23] model as another baseline. This model shows significantly lower performance across all metrics, which is not effective.

Our MAD model, which integrates the brain module with a combined loss ($L_m + L_e$), demonstrates superior performance with a BLEU-1 score of 10.44% which is about 5 points higher than NeuSpeech [13] and RS. Besides, we compared the performance of our model when it receives pure Gaussian noise which is the shape of the MEG signal to show that our model is generating text based on MEG signal. For noise input, MAD's performance BLEU-1 dropped to 3.87%, indicating that MAD model has learned from the MEG signal rather than just noise. Additionally, we evaluated MAD with teacher forcing. When teacher forcing was applied (MAD w/tf), the model's performance significantly improved, achieving a BLEU-1 score of 12.93% and a ROUGE-1-F score of 18.28%, confirming the effectiveness of teacher forcing in enhancing model performance. Similarly, the BLEU-1 score for noise w/tf is low too (0.19%), further indicating our model can distinguish noise and MEG. In addition, our model has low Self-BLEU which means our model is generate diverse sentences according to MEG signal rather than simply repeating.

Overall, our MAD model achieved state-of-the-art (SOTA) performance for MEG-to-text decoding compared to previous SOTA models, demonstrating significant progress in MEG-to-text translation. Additionally, we performed a fair comparison with noise and RS, which served as two error bars to validate the robustness and reliability of our model's performance. Furthermore, the self-BLEU scores indicated the diversity of our model's generated text, demonstrating its ability to truly learn and generalize from the data. Next section, we will show the generated sample along with the Mel spectrogram to further show the effectiveness of our MAD model.

4.4 Generated Samples

4.4.1 Text

Table 3. Transcription results. These are some results obtained with teacher forcing evaluation. **Bold** for exact matched words, underline for similar semantic words.

Decoding Results on *GWilliams* [16]
Ground Truth: corner **of his** eyes two forts stood on the playground and a **hot** Prediction: the own **of his** in the ground for and the few spot **hot**
Ground Truth: **of the top** of the black fort like a gold headed monster tucker Prediction: **of the top** hole a giant medal to is
Ground Truth: **he** knew it wasn't going to be as relax easy as just pretending he was too Prediction: **he** wast a to be a bad as as as it a to was a lazy

We showed the text result in table 3. It presents the transcription results obtained using the teacher forcing evaluation method. The transcription results indicate that while the model can generate segments of text that partially match the ground truth, there are significant gaps in overall accuracy and coherence. Specifically, for the first example, the ground truth is "corner of his eyes two forts stood on the playground and a hot" and the model's prediction is "the own of his in the ground for and the few spot hot". Although the model captures some keywords like "ground" and "hot," the overall sentence diverges significantly from the ground truth, exhibiting repetition and grammatical errors. This outcome highlights the model's struggle with complex sentence structures and semantic relationships.

In the second example, the ground truth is "of the top of the black fort like a gold headed monster tucker," while the model predicts "of the top hole a giant medal to is." Here, the model successfully identifies "of the top," and the subsequent key word matches the ground truth in semantics, particularly "medal" which is semantically similar to "gold" in the ground truth sentence. Though it ignores the "monster tucker", this suggests that, despite the MAD model's failure in maintaining coherence and context understanding between words, it can yield some keywords which are semantically similar to the keywords in the original sentence.

For the third example, the ground truth is "he knew it wasn't going to be as relax easy as just pretending he was too," whereas the model predicts "he wast a to be a bad as as as it a to was a lazy." Although the initial word "he" matches the ground truth, the following prediction includes repeated words and grammatical errors, making it difficult to form a meaningful sentence. However, we should notice that "lazy" may be similar to "relax" in meaning.

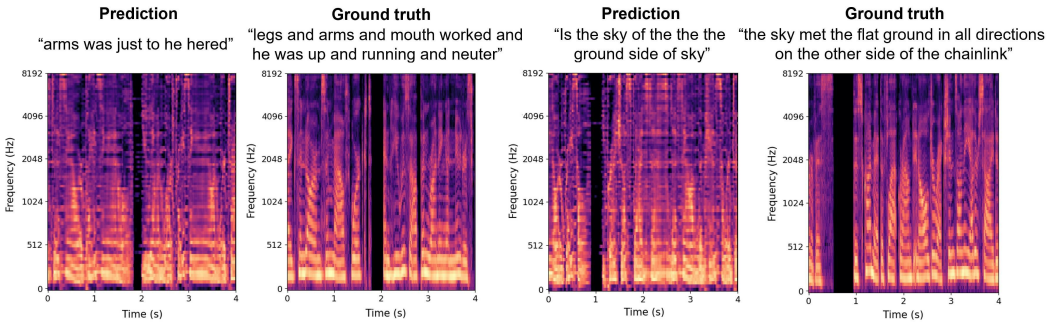


Figure 2. Two sample examples from the test set. Predictions refer to Mel spectrograms generated by the brain module. Ground truth refers to Mel spectrograms of the audio signal processed by the whisper processor. The predicted text was generated using teacher forcing.

For these three texts, we can observe that MAD can generate semantically similar words though they may not be coherent. Besides, the noise inputs generated only blanks. This demonstrates that our MAD model can capture the semantic meaning in the MEG signals, rather than only relying on the decoder.

Overall, these results reflect the inherent difficulties in directly decoding MEG signals into natural language text. While the model demonstrates some capability in recognizing individual words, there is substantial room for improvement in generating coherent and accurate sentences. Particularly, the model struggles with complex grammatical structures and longer sentences. These findings underscore the necessity of further optimizing the decoding model, especially in enhancing contextual awareness and semantic coherence. Future work should focus on improving the model's ability to understand context and the relationships between words to enhance the overall accuracy and readability of the transcriptions.

4.4.2 Mel spectrogram

More than text, we showed the Mel spectrogram in Figure 2. It presents the Mel spectrogram of the two sample sentences in the test set. In this context, it is employed to compare the predicted audio signal generated by the model with the actual ground truth audio signal in the form of Mel spectrogram.

Upon examining the spectrograms of two samples, several observations can be made regarding the model's capabilities and performance. 1) There is a general similarity between prediction and ground truth in the overall structure, 2) the model learns some fine-grained details such as temporal variations in the low-frequency regions which have bigger energy than the high-frequency region, 3) the model can predict the speech signal's temporal blanks, proving it understands the MEG features associated with the absence of speech. However, significant discrepancies are apparent. While the ground truth spectrogram displays a more complex and detailed pattern with distinct frequency bands and variations over time, the predicted spectrogram seems less detailed and exhibits more uniform and repetitive patterns.

These discrepancies highlight the current limitations of the model in producing high-quality, accurate, natural audio signals from MEG data. Future work can introduce pre-trained generative models in speech modality to improve the model's ability to learn and represent these fine-grained details, which is important for accurate speech recognition.

4.5 Model Ablation

Table 4 presents a comparison of various configurations of trainable modules and loss functions in the brain-to-text decoding model, evaluated under teacher forcing conditions. The configurations include different combinations of the brain module (B), LoRA applied to the encoder (Lo), the encoder (E), and the decoder (D). The evaluation metrics used are BLEU-1, ROUGE-1, Bert score, and CER (Character Error Rate), Self-BLEU.

Table 4. Here shows the comparison of using different modules and loss. B means brain module, Lo means LoRA applied on encoder. E means encoder, D means decoder. These results are obtained without teacher forcing in evaluation. $L_m(mmd)$ is the mmd loss for aligning mel spectrogram instead of Clip loss. B-1 is the abbreviation of BLEU-1. R-1 is the ROUGE-1-F. SB is self-BLEU.

Loss	Trainables	Architect.	B-1 (%)↑	R-1 (%)↑	Bert (%)↑	CER (%)↓	SB (%)↓
L_e	B	B+D	10.09	6.29	82.74	88.84	83.62
$L_e + L_t$	B	B+D	6.15	4.81	84.43	80.33	95.32
L_m	B	B+E+D	1.88	2.24	79.83	83.65	99.03
$L_m + L_e$	B	B+E+D	10.44	6.93	83.39	89.82	85.28
$L_m(mmd) + L_e$	B	B+E+D	9.64	5.71	81.62	87.95	80.55
$L_m + L_e + L_t$	B	B+E+D	7.14	4.37	82.29	88.40	83.95
$L_m + L_e$	B+Lo	B+E+D	1.13	0.79	81.17	87.65	99.98
$L_m + L_e + L_t$	B+Lo	B+E+D	8.33	6.40	83.14	91.43	99.11

The baseline configuration, using only the brain module with the loss function L_m, achieves a BLEU-1 score of 1.88 and a ROUGE-1 score of 2.24, with Bert and CER scores of 79.83 and 83.65, respectively. Self-BLEU scores of 99.03 indicate the model generated almost identical sentences, showing that using only the brain module results in significant errors and inaccurate content.

Adding the encoder loss L_e to L_m while maintaining the same modules (B+E+D) significantly improves performance, yielding a BLEU-1 score of 10.44 and a ROUGE-1 score of 6.93. The Bert and CER scores were 83.39 and 89.82, respectively. In this configuration, we changed the alignment loss from clip to MMD($L_m(mmd) + L_e$), resulting in BLEU-1 scores of 9.64 and ROUGE-1 scores of 5.71. Similarly, the configuration using the brain module with L_e (B+D) achieves the following scores: a BLEU-1 score of 10.09, a ROUGE-1 score of 6.29, a BERT score of 82.74, and a CER of 88.84. This indicates enhanced decoding accuracy when using encoder loss. Adding the triplet loss L_t to this configuration decreases the BLEU-1 and ROUGE-1 scores to 6.15 and 4.81, respectively.

Using LoRA with the combination of L_m and L_e results in significantly poor performance, with BLEU-1 and ROUGE-1 scores of 1.13 and 0.79, and Bert scores of 81.17. The Self-BLEU score of 99.98 indicates that this configuration is highly ineffective, likely due to an incompatibility between LoRA and the task requirements. Incorporating the triplet loss L_t along with L_m and L_e for the same architecture (B+E+D) resulted in a Self-BLEU score of 99.11, indicating that the model generated almost identical sentences.

The results indicate that the brain module (B) is crucial for effective brain-to-text decoding, and the combination of multiple loss functions, particularly with the inclusion of the encoder loss (L_e), enhances the model's performance. Configurations involving LoRA applied to the encoder are generally less effective unless complemented with the L_t, highlighting the need for carefully designed adaptation strategies for optimal performance in this context.

5 Limitation

Although our MAD model outperforms previous SOTA models, we have to point out that this model's generation is far from practical utilization in reality since the performance is much lower than speech recognition models. Besides, this work is implemented on listening datasets, which is different from silent speech.

6 Conclusion

In this paper, we presented MAD, a novel end-to-end training framework for MEG-to-Text translation. Our model leverages a multi-stage alignment utilizing auxiliary audio modality, which aligns brain activity data more effectively with corresponding textual outputs. Experimental results suggest that the newly proposed MAD framework achieves 10.44 BLEU-1 on *GWilliams* **without teacher-forcing** evaluation on **entirely unseen text** which largely exceeds the current SOTA performance. Through comprehensive ablation studies, we demonstrated the performance of our approach in various situations. Our results indicate that the brain module, in conjunction with appropriate loss functions, substantially enhances decoding performance. The inclusion of encoder and decoder

modules further refines the text generation process, with the triplet loss playing a crucial role in improving the model's robustness and accuracy. Particularly, the combination of the brain module with both the encoder and decoder, enhanced by multiple loss functions, shows marked improvements in BLEU-1 and ROUGE-1 scores, while reducing word and character error rates. The insights gained from this research underline the potential of the MAD framework in the realm of neural decoding. By effectively capturing the complex patterns in MEG signals and translating them into coherent text, our approach offers a promising solution for brain-to-text applications. This work sets the stage for further exploration into multi-modal alignments and their impact on neural decoding systems.

In conclusion, our proposed MAD framework significantly advances the state-of-the-art in brain-to-text decoding, offering new avenues for enhancing communication tools for individuals with severe speech and motor impairments. Future work will focus on refining alignment mechanisms and extending the application of our model to more diverse linguistic tasks.

References

[1] Gopala K. Anumanchipalli, Josh Chartier, and Edward F. Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, April 2019.

[2] Ran Wang, Xupeng Chen, Amirhossein Khalilian-Gourtani, Zhaoxi Chen, Leyao Yu, Adeen Flinker, and Yao Wang. Stimulus speech decoding from human cortex with generative adversarial network transfer learning. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, April 2020.

[3] Francis R. Willett, Donald T. Avansino, Leigh R. Hochberg, Jaimie M. Henderson, and Krishna V. Shenoy. High-performance brain-to-text communication via handwriting. *Nature*, 593(7858):249–254, May 2021.

[4] Ran Wang, Xupeng Chen, Amirhossein Khalilian-Gourtani, Leyao Yu, Patricia Dugan, Daniel Friedman, Werner Doyle, Orrin Devinsky, Yao Wang, and Adeen Flinker. Distributed feedforward and feedback processing across perisylvian cortex supports human speech. December 2021.

[5] Francis R. Willett, Erin M. Kunz, Chaofei Fan, Donald T. Avansino, Guy H. Wilson, Eun Young Choi, Foram Kamdar, Leigh R. Hochberg, Shaul Druckmann, Krishna V. Shenoy, and Jaimie M. Henderson. A high-performance speech neuroprosthesis. January 2023.

[6] Sean L. Metzger, Kaylo T. Littlejohn, Alexander B. Silva, David A. Moses, Margaret P. Seaton, Ran Wang, Maximilian E. Dougherty, Jessie R. Liu, Peter Wu, Michael A. Berger, Inga Zhuravleva, Adelyn Tu-Chan, Karunesh Ganguly, Gopala K. Anumanchipalli, and Edward F. Chang. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620(7976):1037–1046, August 2023.

[7] Zhenhailong Wang and Heng Ji. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5350–5358, 2022.

[8] Yiqun Duan, Charles Zhou, Zhen Wang, Yu-Kai Wang, and Chin teng Lin. Dewave: Discrete encoding of eeg waves for eeg to text translation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[9] Debadatta Dash, Paul Ferrari, and Jun Wang. Decoding imagined and spoken phrases from non-invasive neural (meg) signals. *Frontiers in neuroscience*, 14:290, 2020.

[10] Richard Csaky, Mats WJ van Es, Oiwi Parker Jones, and Mark Woolrich. Interpretable many-class decoding for meg. *NeuroImage*, 282:120396, 2023.

[11] Gayane Ghazaryan, Marijn van Vliet, Aino Saranpää, Lotta Lammi, Tiina Lindh-Knuutila, Annika Hultén, Sasa Kivisaari, and Riitta Salmelin. Trials and tribulations when attempting to decode semantic representations from meg responses to written text. *Language, Cognition and Neuroscience*, pages 1–12, 2023.

[12] Jo Hyejeong, Yang Yiqian, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hee Lee. Are eeg-to-text models working? *arXiv preprint arXiv:2405.06459*, 2024.

[13] Yiqian Yang, Yiqun Duan, Qiang Zhang, Renjing Xu, and Hui Xiong. Decode neural signal as speech. *arXiv preprint arXiv:2403.01748*, 2024.

[14] Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5(10):1097–1107, October 2023.

[15] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[16] Laura Gwilliams, Graham Flick, Alec Marantz, Liina Pylkkänen, David Poeppel, and Jean-Rémi King. Introducing meg-masc a high-quality magneto-encephalography dataset for evaluating natural speech processing. *Scientific Data*, 10(1), December 2023.

[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[18] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[19] Sean L. Metzger, Jessie R. Liu, David A. Moses, Maximilian E. Dougherty, Margaret P. Seaton, Kaylo T. Littlejohn, Josh Chartier, Gopala K. Anumanchipalli, Adelyn Tu-Chan, Karunesh Ganguly, and Edward F. Chang. Generalizable spelling using a speech neuroprosthesis in an individual with severe limb and vocal paralysis. *Nature Communications*, 13(1), November 2022.

[20] Yan Liu, Zehao Zhao, Minpeng Xu, Haiqing Yu, Yanming Zhu, Jie Zhang, Linghao Bu, Xiaoluo Zhang, Junfeng Lu, Yuanning Li, Dong Ming, and Jinsong Wu. Decoding and synthesizing tonal language speech from brain activity. *Science Advances*, 9(23), June 2023.

[21] C Feng, L Cao, D Wu, E Zhang, T Wang, X Jiang, H Ding, C Zhou, J Chen, H Wu, et al. A high-performance brain-to-sentence decoder for logosyllabic language. 2023.

[22] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.

[23] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[25] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

[26] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023.

[27] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[28] Emilia P Martins and Theodore Garland Jr. Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. *Evolution*, 45(3):534–557, 1991.

[29] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100, 2018.

Supplementary Material for MAD: Multi-alignment MEG-text decoding

A Dataset

The Gwilliams [16] dataset is described below:

A.1 Participants

- **Total Participants:** 27 English-speaking adults (15 females)
- **Age:** Mean = 24.8 years, SD = 6.4 years
- **Recruitment:** Subject pool of NYU Abu Dhabi
- **Consent and Compensation:** All provided written informed consent and were compensated
- **Health:** Reported normal hearing and no history of neurological disorders
- **Language:** All but one participant (S20) were native English speakers
- **Sessions:**
  - Majority (22 participants) performed two identical one-hour-long sessions
  - Sessions were separated by 1 day to 2 months
- **Ethics Approval:** Approved by the IRB ethics committee of NYU Abu Dhabi

A.2 Procedure

- **Recording Sessions:**
  - Duration: Each session lasted approximately 1 hour.
  - Equipment: Recorded with a 208 axial-gradiometer MEG scanner (Kanazawa Institute of Technology).
  - Sampling Rate: 1,000 Hz.
  - Filtering: Online band-pass filtered between 0.01 and 200 Hz.
  - Task: Participants listened to four distinct stories through binaural tube earphones (Aero Technologies) at a mean level of 70 dB sound pressure level.
- **Pre-Experiment Exposure:**
  - Participants were exposed to 20 seconds of each distinct speaker voice.
  - Purpose: To clarify session structure and familiarize participants with the voices.
- **Story Presentation Order:**
  - Assigned pseudo-randomly using a "Latin-square design."
  - Same order used for both recording sessions for each participant.
- **Attention Check:**
  - Participants answered a two-alternative forced-choice question every 3 minutes.
  - Example Question: "What precious material had Chuck found? Diamonds or Gold."
  - Average Accuracy: 98%, confirming engagement and comprehension.
- **MRI Scans:**
  - T1-weighted anatomical scans were performed after MEG recording if not already available.
  - Six participants did not return for their T1 scan.
- **Head Shape Digitization:**
  - Head shape digitized with a hand-held FastSCAN laser scanner (Polhemus).
  - Co-registered with five head-position coils.
  - Coil positions collected before and after each recording, stored in the 'marker' file.
  - Experimenter continuously monitored head position to minimize movement.

A.3 Stimuli

- **Stories:** Four English fictional stories selected from the Open American National Corpus:
  - **'Cable spool boy':** 1,948 words, narrating two young brothers playing in the woods.
  - **'LW1':** 861 words, narrating an alien spaceship trying to find its way home.
  - **'Black willow':** 4,652 words, narrating the difficulties an author encounters during writing.
  - **'Easy money':** 3,541 words, narrating two friends using a magical trick to make money.
- **Audio Tracks:**
  - Synthesized using Mac OS Mojave's (c) text-to-speech.
  - Voices and speech rates varied every 5-20 sentences to decorrelate language from acoustic representations.
  - Voices used: 'Ava', 'Samantha', and 'Allison'.
  - Speech rate: Between 145 and 205 words per minute.
  - Silence between sentences: Varied between 0 and 1,000 ms.
- **Story Segments:**
  - Each story divided into 5-minute sound files.
  - Random word list played approximately every 30 seconds, generated from unique content words of the preceding segment.
  - Very small fraction (<1%) of non-words introduced in natural sentences.
- **Task Definition:**
  - Each "task" corresponds to the concatenation of sentences and word lists.
  - All subjects listened to the same set of four tasks, in different block orders.

B Discussion about the main table

We used BLEU-1 [17],ROUGE-1-F [18],BertScore [27],CER [28],Self-BLEU [29] as metrics in the main table to show the capability of previous models and our models. However, as observed, NeuSpeech [13] model has the best score for ROUGE-1,Bert,CER, which is incredible, therefore we measured the Self-BLEU of this model, which is almost 100%, and found out NeuSpeech predicts almost the same sentence "He looked at me and said to me" all the time for different sentences in Supp. 1. Generation of this bad quality has best score on these three metrics, which means these three metrics are not effective in measuring the generation quality. Therefore, we think BLEU-1 the most reliable metric in this task for now. Besides, we randomly selected sentences, which is RS in the table, from the test set as another baseline, we found out that the BLEU-1 score is higher than NeuSpeech, which means the NeuSpeech model is not effective, which is very reasonable. After all, it seems that using BLEU score is the only reasonable metric of evaluating the quality of generated text.

As observed in the table, it is very clear that our MAD model is significantly higher than RS and NeuSpeech and Wav2vec2CTC on BLEU-1, which means our MAD model is effective on unseen text.

C More generated samples

We showed more generate samples here to show that we are not cherry-picking.

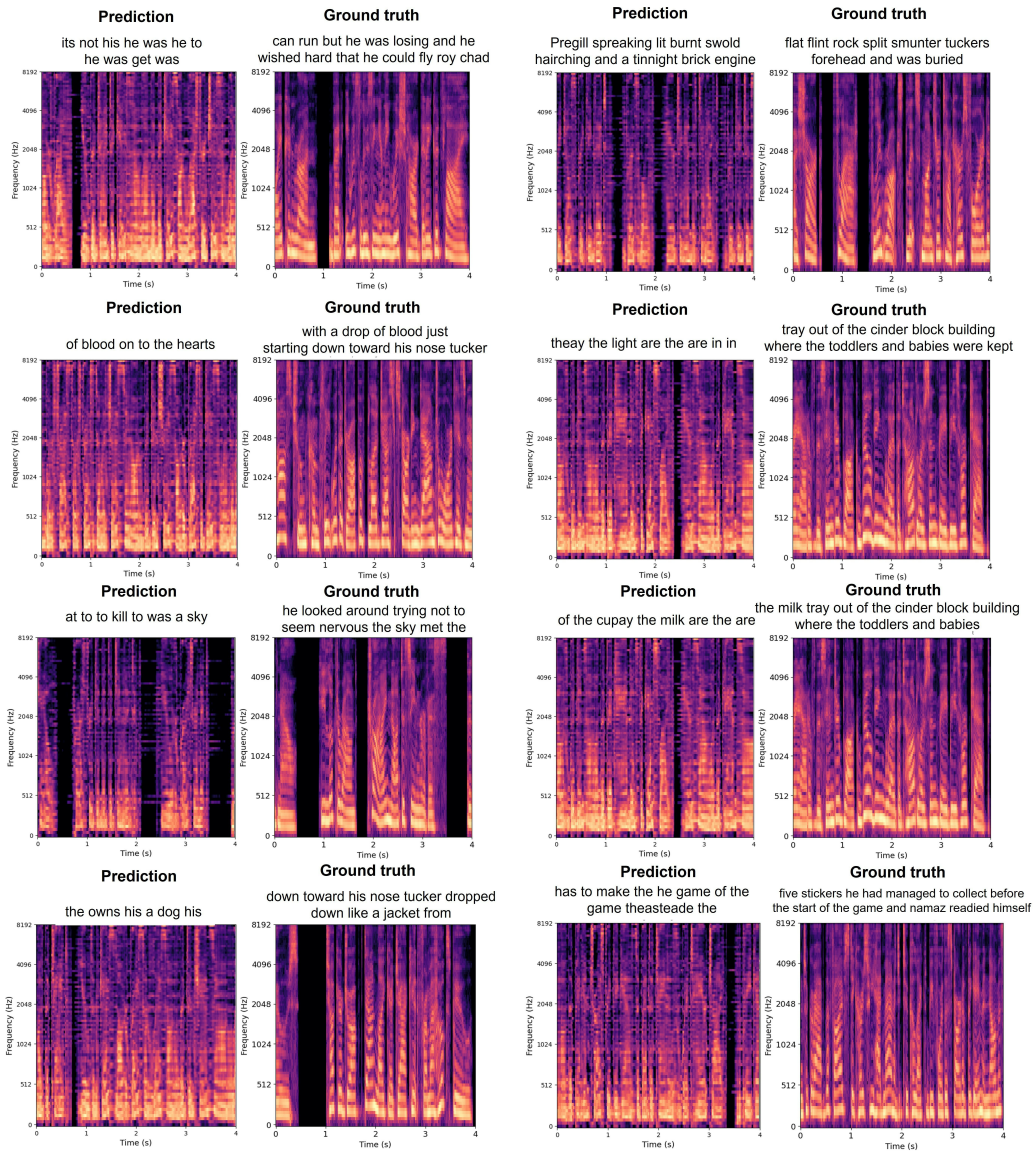


Figure 3. Eight sample examples of the test set. Prediction refers to Mel spectrograms generated by the brain module. Ground truth refers to Mel spectrograms of the audio signal processed by the whisper processor. The predicted text was generated using teacher forcing. These examples were produced using $L_m(mmd)$ with only a trainable brain module.

Listing 1. NeuSpeech [13] generation without teacher forcing.

```
start********************************
Predicted:  He looked at me and said to me,
True: were smelly thistles or cocklebur stems covered with spiked
end==================================

start********************************
Predicted:  He looked at me and said to me,
True: or ordering Chad around or something. But since his fall the
    year before,
end==================================

start********************************
Predicted:  I'm not sure how to do it. It's just a little bit more
True: oldest boy in the playground, and the one who decided the rules
end==================================

start********************************
Predicted:  He looked at me and said to me,
True: Spauw for fear of what was coming next. I'll make you fight.
    Tucker
end==================================

start********************************
Predicted:  he looked at me and said, I don't know what to do.
True: before, Roy had been shuffling and doing what he was told. Chad
end==================================

start********************************
Predicted:  He looked at me and said to me,
True: for the tumbleweed to prove he wasn't a baby to Tucker. But as
    much
end==================================

start********************************
Predicted:  He looked at me and said to me,
True: walk really every something great blade over. Mama
end==================================

start********************************
Predicted:  He looked at me and said to me,
True: other ready to step down into Chad's back. A sharp, Flat,
end==================================

start********************************
Predicted:  He looked at me and said to me,
True: about gathering stickers himself. Roy was too
end==================================

start********************************
Predicted:  He looked at me and said to me,
True: in shade and napped inside the walls. Then could wild and blink-
    breath corner-hard
end==================================
```

Listing 2. Wav2vec2CTC [14] generation.

```
1
2  start********************************
3  Predicted: THLE'S HOAN BSFBHLAG'DS HON CITES HAG THOEANGLEN S QJRANGD
       HOAND'S SORUESTHO E MRERLWOAINS HOAX TH
4  True: AND NAPPED INSIDE THE WALLS THEN COULD WILD AND BLINKBREATH
       CORNERHARD
5  end==================================
6
7  start********************************
8  Predicted: SHROE BHOING TSEDTRAINS BBB
9  True: OF TIRES TWO BIG TRACTOR TIRES CAPPED OUT WITH ONE FROM A TRUCK
       AND TWO SMALLER
10 end==================================
11
12 start********************************
13 Predicted: IES HO BHE HRORA SCIRCIND FBW
14 True: THAT OUT EITHER IT WAS ROY'S FAVORITE GAME NO
15 end==================================
16
17 start********************************
18 Predicted: AGSCHRONDSOUNE HIRS ON HOIN PHRORLI'S HEXSHIS B
19 True: ABOUT GATHERING STICKERS HIMSELF ROY WAS
20 end==================================
21
22 start********************************
23 Predicted: D JABWUISD BHOEND TE AUST THORE MLADS BHAXTS BMOIST OND F
24 True: TWO SMALLER ONES FROM CARS THE OLDER BOYS LAY AROUND IN
25 end==================================
26
27 start********************************
28 Predicted: CHORWALDES OE CSCRER BXSCOUE WONSTFBHE HOITS PR ENS
29 True: WASN'T CHICKEN YOU WANNA PLAY ROBOTS ROY ASKED CHAD
30 end==================================
31
32 start********************************
33 Predicted: BHI'S JMA
34 True: WHAT YOU SUCK CHAD SAID HE WISHED ROY
35 end==================================
36
37 start********************************
38 Predicted: SHOUDTIES BVIEN HOAS S
39 True: MAKING A DOOR TO THE SMALL ROOM INSIDE THE TALL TUMBLEWEED FLAG
40 end==================================
41
42 start********************************
43 Predicted: IDH HOASTD HIE' SCHORK SPHRERG 'S THOANS OABLWSDT'T XSCIED
       HRIE HOER SPTHRALNINDSFOFTHES PHE CHOR HIER
44 True: WEAPONS ALLOWED ACCORDING TO HUMPTY DUMPTY NURSERY RULES
45 end==================================
46
47 start********************************
48 Predicted: SHOURX PHRERLNGDS FHOANS OMBLWSDT'T ESCED RIE HORN
       SFTHRANINDSFOTS FHE CHOR CHIRE HINS HIND HOURXS TH
49 True: ALLOWED ACCORDING TO HUMPTY DUMPTY NURSERY RULES OF ENGAGEMENT
50 end==================================
```

Listing 3. MAD generation with teacher forcing.

```
start**********************************
Predicted: orus said wast be a day but out
True: chad said he wished roy wouldnt fall for that gag every time get
end==================================

start**********************************
Predicted: name is from his head on his head ofs
True: down his head rose and his eyes focused over chads shoulder out
    roy
end==================================

start**********************************
Predicted: be the smell times have at
True: until he could smell the dust several hated must staring brother
end==================================

start**********************************
Predicted: he had not but though he was not a be down to
True: he wished he were there now even if he did have to sit next
end==================================

start**********************************
Predicted: is sky the the the ground side of the sky
True: the sky met the flat ground in all directions on the other side
    of the chainlink fence
end==================================

start**********************************
Predicted: the the up lift him know the the rest the that ist the fool
     the he
True: to lift him and let him reach for the tumbleweed to prove he
    wasnt a baby to tucker but
end==================================

start**********************************
Predicted: sound is the mouth ist been
True: a sick sound but the thing in his head hadnt worked
end==================================

start**********************************
Predicted: the of the top a red medal
True: out of the top of the black fort like a gold headed monster
end==================================

start**********************************
Predicted: the roy him  he name was be
True: out after him roy chad called but his voice would
end==================================

start**********************************
Predicted: soldiers astronautss a on be us and
True: for soldiers and astronauts and its vote going to help roy
end==================================
```

D Ethics

D.1 Safety

Our MEG-to-text translation technology is designed to assist individuals with severe speech and motor impairments by translating brain signals into text. While this technology has the potential to greatly improve quality of life, we acknowledge the possibility of misuse. However, there are no foreseeable situations where the direct application of our technology could harm, injure, or kill people. We do not develop or intend to develop applications that increase the lethality of weapons systems.

D.2 Security

We recognize the importance of securing brain-computer interface systems. Future research should include thorough risk assessments to identify and mitigate potential security vulnerabilities. We recommend employing robust encryption methods and secure data transmission protocols to protect against unauthorized access and ensure the safety of users' neural data.

D.3 Discrimination

Our technology aims to provide equal accessibility to communication for individuals with speech and motor impairments. We are committed to ensuring that our MEG-to-text translation system does not facilitate discrimination or exclusion. We will continuously monitor and test our models to prevent biases that could negatively impact service provision in healthcare, education, or financial sectors.

D.4 Surveillance

We adhere strictly to local laws and ethical guidelines regarding data collection and analysis. Our research does not involve bulk surveillance data. We obtained data from public dataset [16], and we do not predict protected categories or use data in ways that endanger individual well-being.

D.5 Deception & Harassment

Our technology is designed with safeguards to prevent its misuse in deceptive or harmful interactions. We implement verification mechanisms to detect and prevent impersonation and fraudulent activities. We actively work to ensure our system cannot be used to promote hate speech, abuse, or influence political processes maliciously.

D.6 Environment

We are mindful of the environmental impact of our research. While our work primarily involves computational resources, we strive to optimize our algorithms to minimize energy consumption. We do not engage in activities that promote fossil fuel extraction or increase societal consumption. Our focus is on developing sustainable and efficient technologies.

D.7 Human Rights

Our research adheres to ethical standards and legal requirements, ensuring that it does not facilitate illegal activities or deny individuals their rights to privacy, speech, health, liberty, security, legal personhood, or freedom of conscience or religion. We are committed to protecting and promoting human rights through our work.

D.8 Bias and Fairness

Our goal is to create fair and inclusive technology that benefits all users equally, regardless of their background.

E Broader Impacts

E.1 Positive Impacts

Improved Communication for Individuals with Disabilities Our MEG-to-text translation technology has the potential to significantly improve the quality of life for individuals with severe speech and motor impairments. By enabling these individuals to communicate effectively, we can help them achieve greater independence, participate more fully in society, and reduce their reliance on caregivers.

Advancements in Neurotechnology This research contributes to the broader field of neurotechnology, advancing our understanding of brain activity and its relationship to language. These advancements could lead to new therapies and interventions for a variety of neurological conditions, potentially benefiting a wide range of patients.

Innovation and Economic Growth The development and commercialization of advanced neurotechnologies can stimulate economic growth by creating new industries and job opportunities. This innovation can drive progress in related fields such as healthcare, robotics, and artificial intelligence, fostering a collaborative and dynamic technological ecosystem.

E.2 Negative Impacts

Privacy and Security Risks The collection and analysis of neural data pose significant privacy and security concerns. Unauthorized access to such sensitive information could lead to misuse, including identity theft or unauthorized surveillance. Ensuring robust data protection measures is crucial to mitigate these risks.

Potential for Misuse There is a risk that the technology could be misused for purposes other than its intended therapeutic applications. For instance, it could be exploited for invasive surveillance or to manipulate individuals by decoding their thoughts without consent. Strict ethical guidelines and regulations are necessary to prevent such misuse.

Bias and Discrimination If not carefully managed, the technology could inadvertently encode or exacerbate existing biases. For example, if the training data is not representative of diverse populations, the system might perform poorly for certain groups, leading to unequal access to its benefits. Ongoing efforts to ensure fairness and inclusivity are essential to address these concerns.

EIT-1M: One Million EEG-Image-Text Pairs for Human Visual-textual Recognition and More

Xu Zheng[1] * Ling Wang[1] * Kanghao Chen[1] † Yuanhuiyi Lyu[1] † Jiazhou Zhou[1] Lin Wang[1,2] ‡
[1]AI Thrust, HKUST(GZ) [2]Dept. of CSE, HKUST
{yuanhuiyilv, jiazhouzhou}@hkust-gz.edu.cn, zhengxu128@gmail.com, linwang@ust.hk

Abstract

Recently, electroencephalography (EEG) signals have been actively incorporated to decode brain activity to visual or textual stimuli and achieve object recognition in multi-modal AI. Accordingly, endeavors have been focused on building EEG-based datasets from visual or textual single-modal stimuli. However, these datasets offer limited EEG epochs per category, and the complex semantics of stimuli presented to participants compromise their quality and fidelity in capturing precise brain activity. The study in neuroscience unveils that the relationship between visual and textual stimulus in EEG recordings provides valuable insights into the brain's ability to process and integrate multi-modal information simultaneously. Inspired by this, we propose a novel large-scale multi-modal dataset, named ***EIT-1M****, with over 1 million EEG-image-text pairs. Our dataset is superior in its capacity of reflecting brain activities in simultaneously processing multi-modal information. To achieve this, we collected data pairs while participants viewed alternating sequences of visual-textual stimuli from 60K natural images and category-specific texts. Common semantic categories are also included to elicit better reactions from participants' brains. Meanwhile, response-based stimulus timing and repetition across blocks and sessions are included to ensure data diversity. To verify the effectiveness of EIT-1M, we provide an in-depth analysis of EEG data captured from multi-modal stimuli across different categories and participants, along with data quality scores for transparency. We demonstrate its validity on two tasks: 1) EEG recognition from visual or textual stimuli or both and 2) EEG-to-visual generation.*

1. Introduction

Electroencephalography (EEG) is a widely applied neuroimaging modality in cognitive neuroscience. It is known for its ability to decipher intricate brain activity patterns during various cognitive processes [29]. In the early days, research focused on constructing EEG datasets for medical purposes, such as detecting and predicting seizures [36]. Recently, EEG signals have been broadly incorporated to decode brain activity to visual or textual stimuli and achieve object recognition in multi-modal artificial intelligence (AI) [3, 7, 28, 31–33]. This enriches the data landscape, allowing for more nuanced and accurate models of brain activity and cognitive processes.

Accordingly, research endeavors have been focused on building EEG-based datasets [10, 11, 14, 18, 34], as summarized in Tab. 1. For instance, ZuCo 1.0 [13] is a pioneering EEG-Text dataset that records neural processes underlying reading and language comprehension during the reading tasks. On the other hand, Brain2Image [17] is a representative EEG-image dataset that includes evoked responses to visual stimuli from 40 classes. However, these datasets have two distinct shortcomings: ***1)*** They offer limited EEG epochs per category, and the complex semantics of stimuli presented to participants compromise their quality and fidelity in capturing precise brain activity. ***2)*** They only encompass EEG signals recorded from single-modal stimuli, either visual or textual. This makes them less possible to be used for training high-performance multi-modal AI models.

The study in neural science reveals that EEG recordings reveal a significant relationship between visual and textual stimuli, offering valuable insights into the brain's capacity to integrate multi-modal information simultaneously [29]. This integration is crucial for understanding how the brain processes complex, real-world scenarios where multiple types of sensory input are encountered simultaneously. Inspired by this, we introduce a novel large-scale multi-modal EEG dataset, **EIT-1M**, comprising paired EEG, visual, and textual data for the benefit of research communities. The key insight of our dataset is to record human brain activities while simultaneously processing multi-modal information. To achieve this, data was collected from five participants exposed to random sequences of 60K natural images and their corresponding category descriptions. To date, we have

*† equal contribution. ‡ corresponding author.

Dataset	Year	Equipment	Modality	Epochs	CEA	MEA	Purpose
Brain2Image [18]	2017	Brainvision BrainAmp DC	EEG-Image	11,466	×	×	Decoding
EVSR [21]	2018	Emotiv EPOC+	EEG-Image	13,800	✓	×	Recognition
ZuCo 1.0 [13]	2018	EGI Geodesic Hydrocel system	EEG-Text	259,788	×	×	NLP
ZuCo 2.0 [14]	2020	EGI Geodesic Hydrocel system	EEG-Text	272,484	×	×	NLP
THINGS EEG1 [11]	2022	Brainvision actiCHamp	EEG-Image	1,112,400	✓	×	Recognition
THINGS EEG2 [9]	2022	Brainvision actiCHamp	EEG-Image	821,600	×	×	Recognition
Alljoined1 [37]	2024	BioSemi ActiveTwo	EEG-Image	46,080	×	×	Decoding
EIT-1M (Ours)	2024	Brainvision actiCHamp Plus	EEG-Image-Text	1,200,000	✓	✓	Recognition & Decoding

Table 1. EEG datasets. (CEA: Category-level ERP Analysis, MEA: Multi-modal ERP Analysis.)

gathered over **1 million** epochs of brain responses using a 64-channel EEG headset (actiCHamp Plus[1]).

Specifically, we utilize the 10-category dataset CIFAR-10 [20] to construct the visual and textual stimulus. This dataset harnesses an image resolution of 32×32 pixels without excessive details, Empirically, as shown in Fig. 2, we find that low-resolution visual stimuli stimulate more stable neural responses, suggesting they are appropriate and manageable within a brief viewing period. We present visual and textual stimuli sequentially to maintain continuous engagement with the objects and concepts, as shown in Fig. 3. Moreover, our dataset features response-based stimulus timing, repetition across blocks and sessions, and diverse visual and textual classes. To verify the effectiveness, we provide an in-depth analysis of EEG data captured from multi-modal stimuli across different categories and participants. The data analysis includes EEG topographic maps, corresponding signals analysis and ERP analysis. These analysis highlight the distinct ERP characteristics from visual and textual stimuli, providing insights in the multi-modal information processing of brains. For transparency, we include data quality scores (See Tab. 3).

To benchmark our EIT-1M, we demonstrate its validity on two tasks: 1) EEG recognition from visual or textual stimuli or both (See Sec. 5.1) and 2) EEG-to-visual generation (See Sec. 5.2). We expect our dataset to be a benchmark contributor for advancing the research for multi-modal AI [5, 6, 25, 26, 38–41] and potentially for cognitive neuroscience.

2. Related Work

EEG Datasets with Visual Stimuli. They capture EEG waveforms while participants view visual stimuli, facilitating studies of brain activity, as shown in Tab. 1. A representative dataset is Brain2Image [17], which includes evoked responses to visual stimuli from 40 classes, each with 50 images, totaling 2K images. However, this dataset is impeded by its lack of train-test separation during recording, block-specific stimuli patterns, and inconsistency across frequency bands [4, 23]. In contrast, the THINGS-EEG1 [11] and THINGS-EEG2 [9] datasets address these issues by incorporating both main and validation sessions to ensure data quality and consistency. These two datasets contain human EEG responses from 50 subjects to 22,248 images in the THINGS stimulus set.

Regarding the diversity of stimuli, while studies like Brain2Image and [1] involve 40 classes, other studies focus on only 10 different image classes [34]. This limited representation allows for more controlled studies but fails to capture the continuous and diverse nature of naturalistic stimuli due to the limited samples from each category. Other datasets like MindBigData [35] and [1] capture a wide range of images but are derived from a single individual, limiting their potential for training image reconstruction models that generalize to other individuals. Recently, to address these limitations, Alljoined1 [37] includes 10K images per participant from object categories in MS-COCO [24], thereby accounting for the diversity and continuity of real-world images.

EEG Datasets with Textual Stimuli They are primarily developed for brain signal decoding. Notable examples include ZuCo 1.0 [13] and ZuCo 2.0 [14], captured with 128-channel EEG devices. These datasets provide insights into the neural processes underlying reading and language comprehension by recording EEG signals during reading tasks. EEG2Text [2] focuses on translating brain signals into textual descriptions, supporting the development of AI models for decoding and generating text from EEG signals. Despite these advancements, there remains a need for datasets that integrate both visual and textual stimuli to capture the com-

[1] https://brainvision.com/products/actichamp-plus/

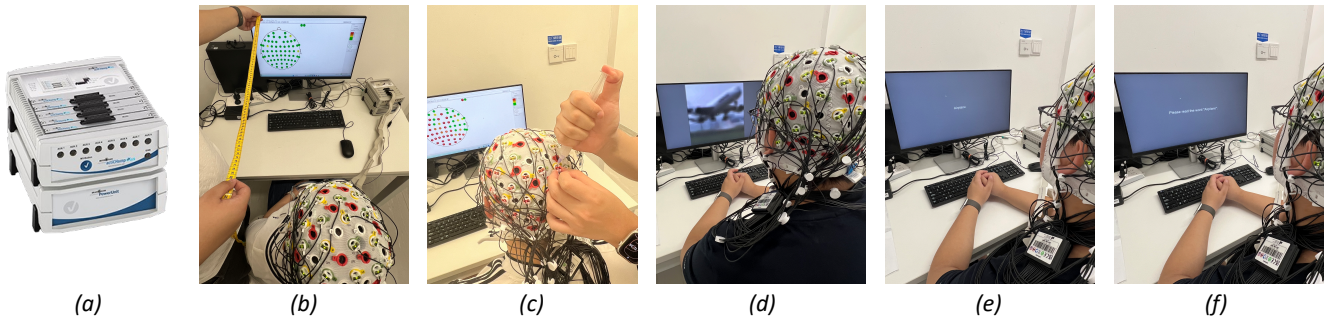


Figure 1. (a) actiCHamp Plus device. (b) Experimental setup with monitor 80 cm from participant. (c) Injecting conductive gel. (d) Visual stimuli. (e) Textual stimuli. (f) Speech stimuli.

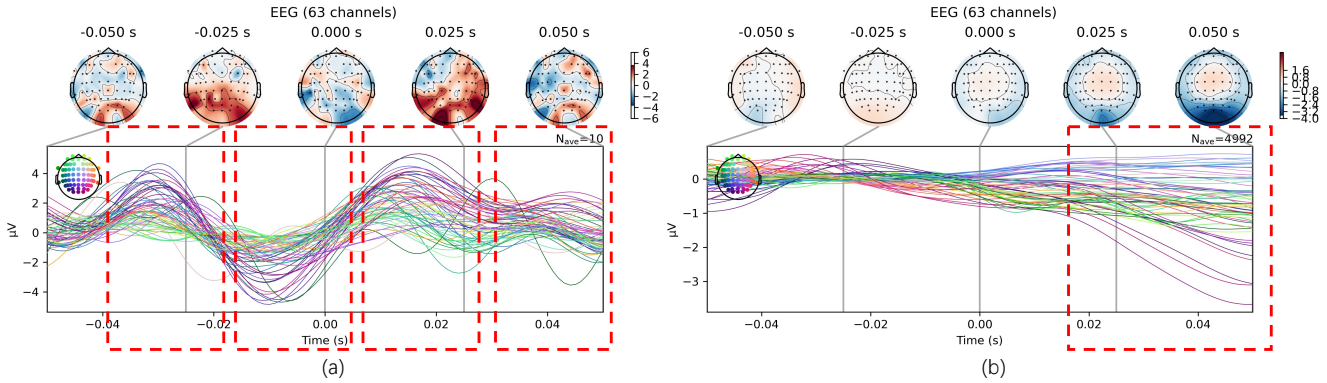


Figure 2. (a) EEG signals from high-resolution visual stimuli.(b) EEG signals from our visual stimuli.

plex interplay between different modalities in the brain. In a nutshell, all these datasets primarily focus on single-modal stimuli, limiting their fidelity for training multi-modal AI models. *Our EIT-1M dataset addresses this gap by providing paired EEG, visual, and textual data, enabling comprehensive multi-modal analysis. Thus our dataset is superior in its capacity of reflecting brain activities in simultaneously processing multi-modal information.*

3. Dataset Collection Methods

Tab. 2 provides an overview of one experiment involving five participants, aged 20-30 years, with a gender distribution of two females and four males. Each participant underwent two 300-minute sessions, during which 1,200,000 events were recorded, including 600K visual and 600K textual stimuli. The stimuli were drawn from ten CIFAR-10 categories for visuals and ten textual categories. EEG recordings were made using a 64-channel headset at a 1000 Hz sampling rate. The dataset ensures high quality with an average signal-to-noise ratio as in Tab. 3, maintaining impedance levels at or below 20 kΩ. Each session featured an average of 10K events, with each event lasting 50 ms and an inter-event interval of 1 second. Preprocessing involved 1-40 Hz band-pass filtering and epoching from -20 to 30 ms relative to stimulus onset, with baseline correction at -20 ms. This dataset aims to support research in EEG analysis and multi-modal recognition.

3.1. Experimental Settings

Participants Five adults (mean age 24.83 years; 1 female, 4 male) participated in this study, all with normal or corrected-to-normal vision, and none of them have suffered or are suffering neurological or psychiatric problems such as ADHD and epilepsy. Each participant provided informed written consent and received monetary reimbursement for their involvement. The study procedures were approved by the ethical committee. It is important to acknowledge the potential limitations of this study, such as the gender imbalance among participants and the low age disparity.

Stimuli. All images used in this study as visual stimuli are sourced from the CIFAR-10 dataset [20]. This dataset is a well-known benchmark in machine learning and computer vision, comprising 60K color images across 10 different classes, with 6K images per class. These classes represent a variety of everyday objects and animals, including airplanes, automobiles, birds, cats, deer, dogs, frogs, horses, ships, and trucks. This dataset is selected for our study due

Item	Description	Details
Participants	Number Age Range Gender Distribution	5 20-30 years 1 females, 4 males
Sessions	Number per Participant Duration	2 4 hours each
Total Events	Total Visual Stimuli Text Stimuli	1,200,000 600,000 600,000
Stimuli	Categories Description	10 from CIFAR-10 Visual: Images from CIFAR-10; Text: category names
Recording Details	Sampling Rate EEG Channels Equipment	1000 Hz 64 actiCHamp Plus[2]
Data Quality	Impedance Levels	≤ 20 kΩ
Event Details	Average Events/Session Event Duration Inter-event Interval	120,000 50 ms 50 ms
Preprocessing	Filtering Epoching	1-40 Hz band-pass -50 to 50 ms relative to stimulus onset, baseline correction at -50 ms

Table 2. Overview of our proposed EEG-Image-Text Dataset

Category	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
Average SNR / dB (raw data)	6.14	5.76	6.09	5.69	5.44	4.82	4.47	4.74	4.30	3.67

Table 3. Example raw data quality of participant 04 in our proposed dataset across different blocks (categories) within the first session.

to its diversity and balanced categories, which provide a robust set of stimuli for examining neural responses across different visual contexts. Utilizing this dataset allows for an in-depth exploration of how the brain processes various types of visual information and supports the development of multi-modal models that can generalize across different categories of visual stimuli. Each image in this dataset has a resolution of 32x32 pixels, making it ideal for stimulating brain activities for visual stimuli from participants. The textual stimuli are derived from the category names within the CIFAR-10 dataset: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck.

Hardware Setup We recorded data using a 64-electrode actiCHamp Plus system, digitized at a rate of 1024 Hz with 24-bit A/D conversion. The montage was arranged according to the international 10-20 System, and the electrode offset was kept below 40 mV. A 22-inch Dell monitor with a resolution of 1080p at 60 Hz was used to display the visual and textual stimuli. As shown in Fig. 1, the monitor was centrally positioned at a distance of 80 cm from the participant, maintaining a 3.5-degree angle of stimuli. We ensured that the angle remained small to minimize the occurrence of gaze drift.

3.2. Data Collection Procedure

Before viewing the stimuli, conductive gel was injected into each electrode to ensure the resistance was less than 20 ohms, facilitating better signal capture. Participants were then shown images and text over the course of four sessions, each four hours long. Each session comprised multiple blocks, with each block containing images from the same class and the corresponding category name text. The visual and textual stimuli were randomly arranged in a visual-textual-visual-textual order within each block. Different blocks contained stimuli from different classes. Within each block, 1,000 visual stimuli images and 1,000 text stimuli category names from CIFAR-10 were presented.

Within each trial, an image was presented for 50 ms, followed by 50 ms of a black screen. The corresponding category name of the image was also presented for 50 ms, followed by 50 ms of a black screen. A white fixation cross was visible on the screen throughout the entire trial. To ensure focus, participants were prompted to press the space bar after completing two consecutive blocks. Additionally, five to ten-minute breaks were provided between blocks based on participants' needs for better data recording. Fig. 3

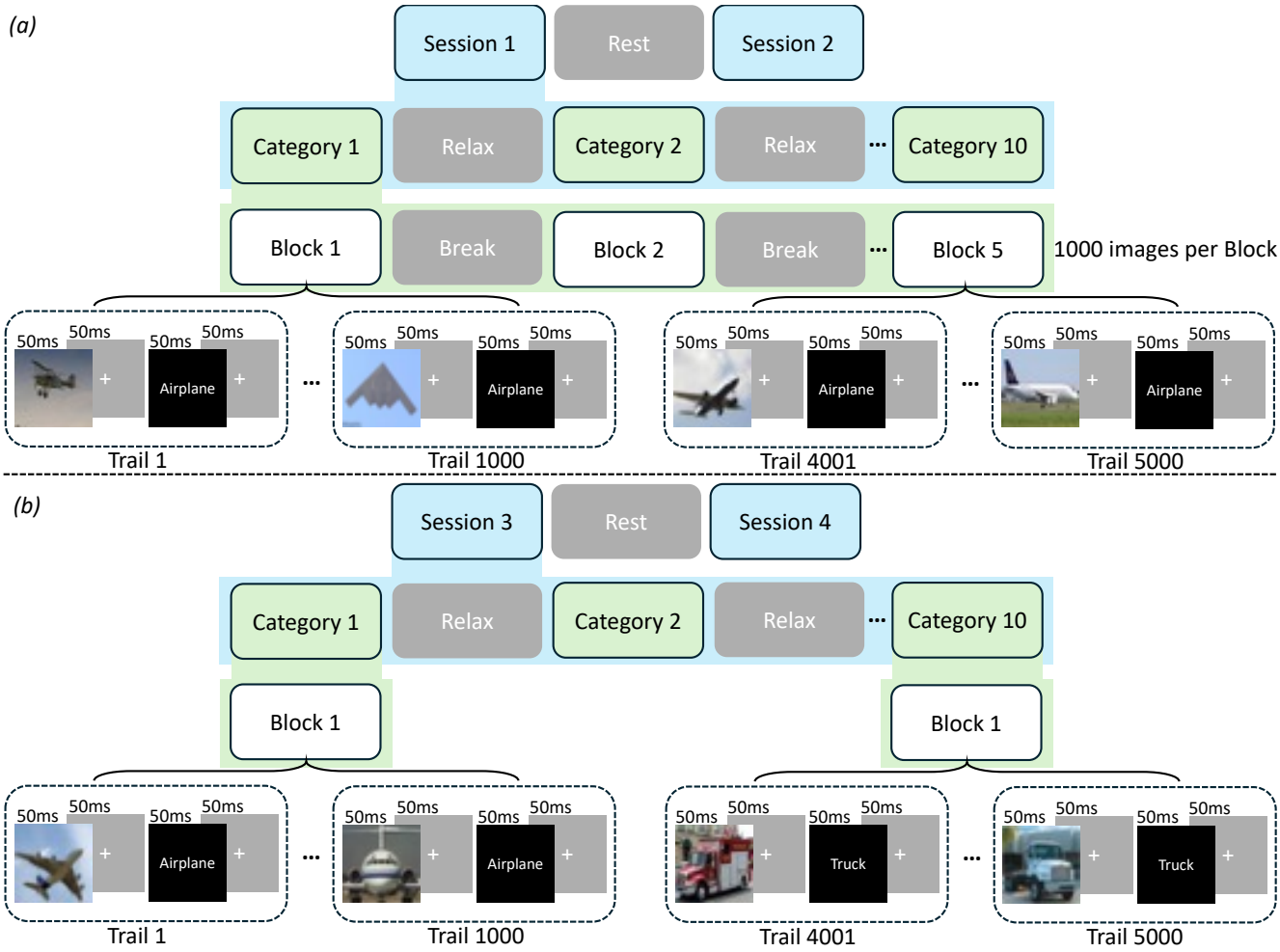


Figure 3. Schematic overview of the structure of trials, blocks, categories and sessions with RSVP paradigm. (a) Training set of CIFAR-10 dataset; (b) Testing set of CIFAR-10 dataset.

(a) shows a schematic overview of the structure of trials, blocks, categories, and sessions, which follows rapid serial visual presentation (RSVP) paradigm [11, 16, 19]. Each of the 5 block-specific CIFAR-10 training images and label text is presented once within each block, and each of the 10 category-specific blocks is presented once within each session. Each participant performed two sessions on different days. Each of the 2 sessions thus consists of 50,000 images and texts within and across blocks. Fig. 3 (b) illustrates that Sessions 3 and 4 consist of 10,000 images and labels from the CIFAR-10 testing set.

4. Data Analysis

4.1. EEG Topographic Maps and Corresponding Signals Analysis

Fig. 4 presents the comparison of EEG signals by showcasing topographic maps and corresponding signals averaged across 63 electrodes (channel FCz as reference) for different stimuli conditions, *i.e.*, visual and textual stimuli with airplane and frog categories. Each column of the figure represents a different stimulus type: visual stimuli (left column) and textual stimuli (right column). The visual stimuli include images from the CIFAR-10 dataset, and the textual stimuli comprise category names from the same dataset. Each row represents different categories, specifically airplane and frog.

Visual Stimuli (Left Column of Fig. 4) The topographic maps show the distribution of brain activity across the scalp at various time points (-0.050s, -0.025s, 0.000s, 0.025s, and 0.050s) after the stimulus onset. The maps reveal distinct patterns of neural activation, indicating how the brain processes visual stimuli over time. For instance, the airplane category (1st row) shows significant activation in the occipital and parietal regions, which are known to be involved in visual processing [27]. The corresponding ERP signals

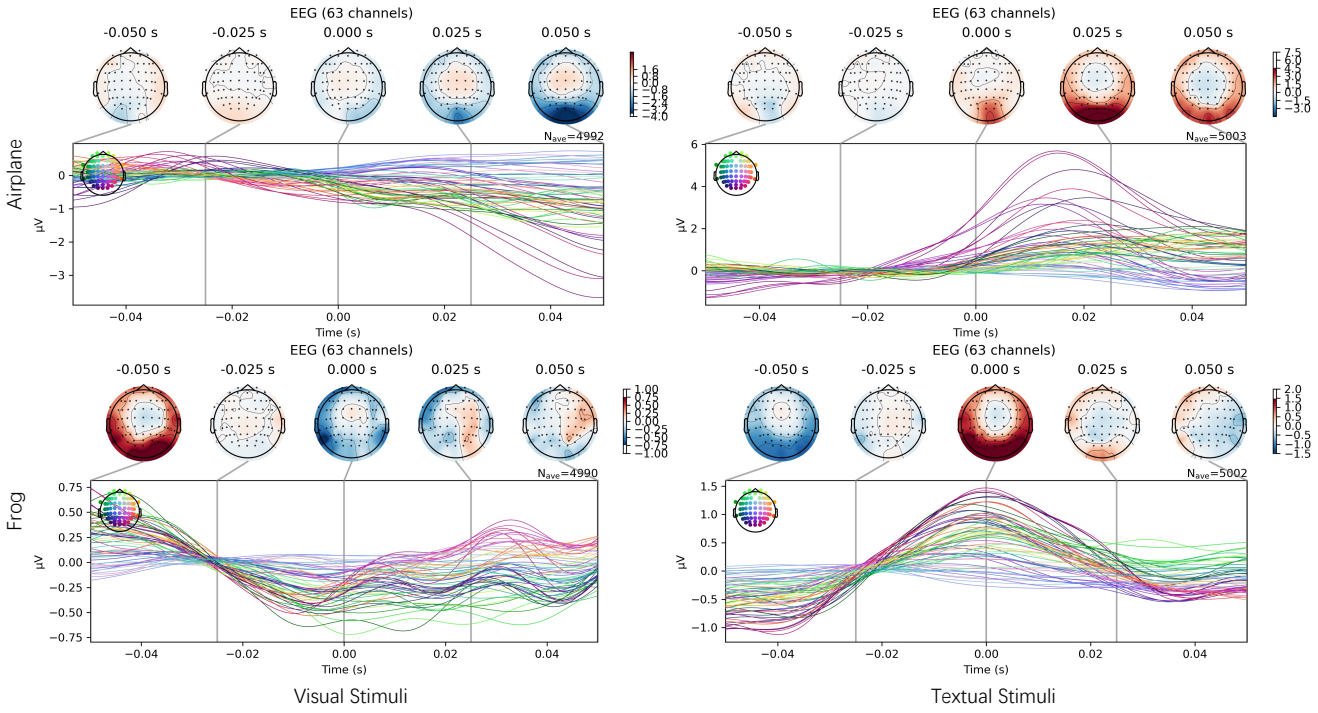


Figure 4. EEG topographic maps and corresponding signals averaged over events for the participant viewing visual stimuli (**left column**) viewing the airplane (1st row) and frog (2nd row) images from the CIFAR-10 dataset, and events for the participant viewing textual stimuli (**right column**) viewing the airplane (1st row) and frog (2nd row) text.

show the average response over time for all electrodes. The signals depict the dynamic changes in brain activity, with notable peaks and troughs corresponding to different cognitive processes. For the visual stimuli, there are clear ERP components around 20ms and 40ms, which might correspond to early visual processing and higher-level cognitive processing, respectively.

Textual Stimuli (Right Column of Fig. 4) Similar to the visual stimuli, the topographic maps for textual stimuli show brain activity at the time points in 50 ms later as visual stimuli. Note that the visual and textual stimuli are presented with a gap of 50 ms. There are noticeable differences in the activation patterns compared to visual stimuli, highlighting the distinct neural processes involved in reading and understanding texts of the participants. For the airplane text (1st row), there is significant activation in the temporal and frontal regions, areas associated with language processing [15]. The ERP signals for textual stimuli also display characteristic peaks, though the patterns differ from those elicited by visual stimuli. The airplane text category shows a strong response between 20 ms to 40 ms, likely reflecting early semantic processing [8].

According to these visualizations, we have the following findings: **(I) Individual and Common Patterns**: Fig. 4 highlights both individual and common brain activity patterns associated with both image and text presentation. This indicates that while there are distinct neural processes for visual and textual stimuli, there are also commonalities in how the brain responds to different types of information. **(II) Temporal Dynamics:** The temporal dynamics of the ERP signals provide insights into the timing of cognitive processes. Early components (within the last 20ms) are typically associated with sensory processing, while earlier components (before 20ms) are linked to cognitive and semantic processing. **(III) Gap Influence**: The 50 ms gap between visual and textual stimuli presentations allowed us to observe the sequential processing of different modalities, showing how the brain transitions between visual and textual information processing. **(IV) ERP Characteristics**: The ERP characteristics, such as the peaks around 20 ms and 40 ms for visual stimuli and between 0 ms to 20 ms for textual stimuli, provide valuable hints for understanding the stages of information processing in the brain.

Unlike previous EEG datasets, such as the THINGS-EEG dataset [11], which use high-resolution images as visual stimuli and introduce a vast number of object concepts (1854), our datasets can address the following limitations of previous ones. Fig. 2 illustrates the differences in EEG signal responses between high-resolution and lower-resolution visual stimuli. The more stable and less variable neural re-

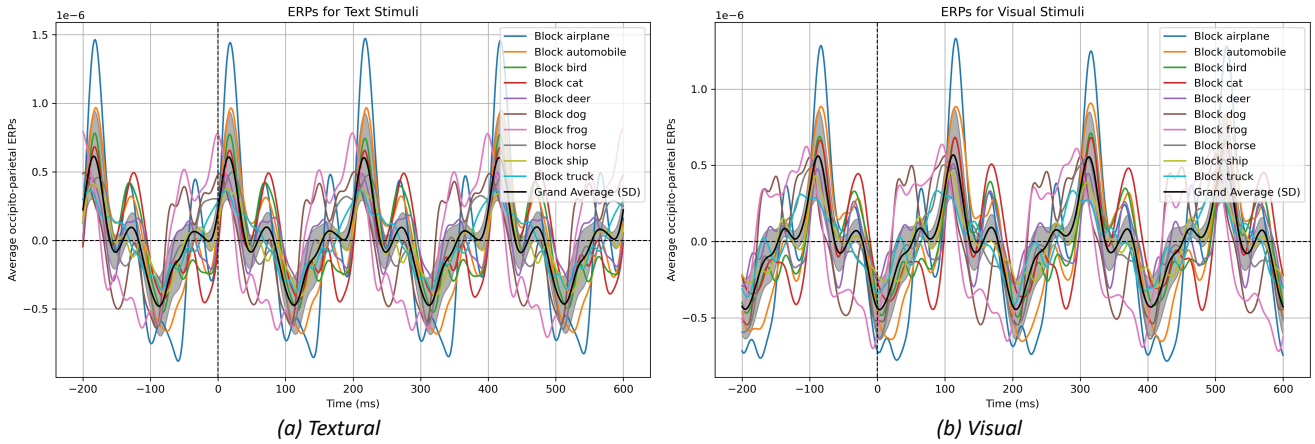


Figure 5. ERPs averaged over occipital and parietal electrodes for the participant viewing stimuli from (a) visual images and (b) the category text. Shaded areas around the grand average ERP represent standard deviations at each time point.

sponses to the lower-resolution images suggest their suitability for creating robust EEG datasets. High-resolution images, on the other hand, require more time for participants to process content and details, making them less suitable for effectively capturing quick neural responses at the millisecond level.

4.2. ERP Analysis

Fig. 5 presents the event-related potentials (ERPs) averaged over occipital and parietal electrodes for a participant viewing visual images (right panel) and category text (left panel). Both plots display ERP data from -200 ms to 600 ms relative to stimulus onset (0 ms), with the average occipito-parietal ERPs fluctuating between approximately -0.5 and 1.5 microvolts for both visual and text stimuli. Each trace, with a distinct color, represents a specific category, including airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck.

Regarding the text stimuli (left panel), a significant initial deflection is noticed around 0 ms, showing the brain's quick response to text stimuli. Early components, like peaks and troughs, are seen around 100 ms and 200 ms post-stimulus, typical of early ERP components such as the P1 and N1, which are linked to sensory processing. Additional peaks around 300 ms and beyond likely indicate higher-order cognitive processing. The shaded area around the grand average ERP line signifies the standard deviation, reflecting variability across different trials and categories. This variability is higher at certain peaks, suggesting differences in how the brain processes various text categories.

Concerning the visual stimuli (right panel), a comparable initial deflection is observed around 0 ms. Distinct peaks are evident at approximately 100 ms and 200 ms, corresponding to the P1 and N1 components, which are more pronounced and consistent across different visual categories compared to textual stimuli. Significant peaks around 300 ms and later may denote the P3 component, indicating cognitive processing associated with visual categorization. The standard deviation shading around the grand average indicates less variability compared to text stimuli, suggesting more consistent brain responses to visual stimuli across various categories.

In comparison, visual stimuli evoke more consistent ERPs across categories than text stimuli, as indicated by the smaller standard deviation areas. Both types of stimuli elicit similar amplitude ranges in the ERP responses, reflecting comparable levels of neural activity. The timing of early and late ERP components is similar for both text and visual stimuli, suggesting that initial sensory processing and subsequent cognitive processing occur within similar time frames for both types of stimuli. In conclusion, the ERPs for both textual and visual stimuli exhibit characteristic early and late components, indicative of sensory and cognitive processing stages. Visual stimuli elicit more consistent responses across categories, whereas text stimuli exhibit greater variability. This analysis provides insights into the sensory and cognitive functions associated with different types of stimuli.

5. Experiments with EIT-1M Dataset

Implementation Details. For preprocessing, a band-pass filter is applied to retain frequencies between 1 and 40 Hz within the raw EEG data. Subsequently, the continuous data is segmented into epochs, each commencing 50 ms prior to the stimulus onset and concluding 50 ms following each event. To train and evaluate the recognition models, the EEG data from one participant (Tab. 4) and two participants (Tab. 5) are divided using an 80/20% split to create training

Models	Image			Text			Image & Text		
	Acc	Recall	F1	Acc	Recall	F1	Acc	Recall	F1
EEGNet [22]	25.42	25.63	24.75	25.62	26.30	24.77	20.72	20.96	19.69
MobileNet_v2 [30]	40.84	41.61	40.79	40.17	39.64	39.52	49.76	49.57	49.32
ResNet18 [12]	**56.57**	**56.41**	**56.46**	56.38	56.17	56.17	**63.53**	**63.65**	**63.55**
ResNet34 [12]	56.41	56.15	56.24	**56.47**	**56.34**	**56.39**	58.77	59.22	58.89
ResNet50 [12]	49.45	48.49	48.80	49.93	49.46	49.61	49.34	50.26	49.38

Table 4. Benchmark experiments within one session of one participant.

Models	Image			Text			Image & Text		
	Acc	Recall	F1	Acc	Recall	F1	Acc	Recall	F1
EEGNet [22]	22.08	22.24	21.80	24.15	24.11	23.56	20.73	20.68	19.83
MobileNet_v2 [30]	41.90	42.18	41.66	42.67	43.32	41.59	48.97	49.28	48.86
ResNet18 [12]	53.49	**53.73**	**53.45**	**54.11**	**54.10**	**54.10**	58.60	58.65	58.54
ResNet34 [12]	**54.06**	53.27	52.42	53.57	53.30	52.65	**60.69**	**60.91**	**60.76**
ResNet50 [12]	49.80	49.60	48.53	49.82	48.81	48.17	56.07	54.98	54.98

Table 5. Benchmark experiments across different sessions of two participants.

Figure 6. Generation results of our dataset using the ThoughtVis Model.

and evaluation sets, respectively. The models are trained using the Adam optimizer, coupled with a step learning rate schedule, across 500 epochs. The default settings for the learning rate, weight decay, and batch size are 1×10^{-3}, 1×10^{-5}, and 2048, respectively. We apply three widely-used metric to evaluate the recognition performance on EIT-1M, including Accuracy, recall, and F1 score.

5.1. Recognition

The results of experiments conducted within one session of a single participant are shown in Tab. 4, illustrating the effectiveness of our dataset in the individual collection procedure. The results in Tab. 4 include the performance across various models with EEG signals captured from visual and textual stimuli. Note that $Image\&Text$ refers to the combined EEG signals from both visual and textual stimuli for recognition. The evaluated models include EEGNet [22], MobileNet-v2 [30], ResNet18 [12], ResNet34 [12], and ResNet50 [12].

Combining EEG signals from image and text stimuli generally enhances performance metrics across all models, suggesting that multi-modal data provides richer information, leading to better classification accuracy and robustness. The consistent performance improvements observed from MobileNet-v2 to ResNet architectures indicate that our EIT-1M dataset is well-suited for various deep-learning models. ResNet models, in particular, show significant improvements, highlighting the dataset's capacity to support complex neural networks. Similar performance metrics for image and text stimuli alone indicate that the dataset offers a balanced representation of both modalities. This balance is crucial for training models to generalize well across different types of stimuli. Additionally, the high F1 scores, especially for the ResNet models, reflect good data quality, ensuring that the recorded EEG signals are reliable and effective for training AI models. Tab. 5 summarizes benchmark experiments across different sessions of two participants. The results consistently show that combining EEG signals from both visual and textual stimuli improves performance across all models compared to using either visual or textual stimuli alone. For both visual and textual stimuli, ResNet models maintain consistently high performance, indicating the robustness of ResNet architectures in processing and learning from EEG data.

The analysis of Tab. 4 and Tab. 5 supports the rationality of our EIT-1M dataset. By providing high-quality, balanced, and scalable data, our dataset proves to be an excellent resource for advancing research in multi-modal AI and cognitive neuroscience. The observed improvements in combined image and text stimuli further highlight the importance of multi-modal datasets in capturing the intricate interplay between different types of information.

5.2. Generation

We follow the classic EEG-to-Image generation task proposed by ThoughtVis [34], which obtains images from EEG signals. As shown in the generation results in Fig. 6, our proposed EIT-1M dataset shows the capability to support the EEG-to-Image generation task.

6. Conclusion, Limitations, and Future Work

In this paper, we presented EIT-1M, a large-scale multi-modal dataset comprising 1 million EEG-image-text pairs. We collected the data pairs while participants viewed alternating sequences of visual-textual stimuli from 60K natural images and corresponding label texts. Our EIT-1M is superior in its capacity of recording brain activities in simultaneously processing multi-modal information, *i.e.*, images and text. It features response-based stimulus timing and repetition across blocks and sessions. To verify the effectiveness of EIT-1M, we provided an in-depth analysis of the EEG signals in EIT-1M across different categories and sessions and conducted experiments on two tasks.

Limitations. Despite the robustness of our dataset, there are areas for enhancement. Our current dataset includes data from multiple participants and sessions, but increasing the number of participants and sessions could yield a more comprehensive understanding of neural responses and improve the generalizability of the models trained on this data. Additionally, while we used a well-defined set of visual and textual stimuli, expanding the variety of stimuli, especially for the textual stimuli, could further enhance the dataset's fidelity for studying more diverse and complex neural processes.

Future work. It could be a good direction to integrate additional modalities, such as audio or tactile feedback, to create an even richer multi-modal dataset. This integration could provide deeper insights into the interplay between different sensory inputs and brain activity, advancing research in multi-modal AI and neuroscience. By addressing these limitations and expanding the dataset's scope, we can significantly contribute to the understanding and development of multi-modal AI models.

Broader Impact. EIT-1M advances neuroscience and AI by enabling deeper insights into cognitive processes and sensory integration. It improves brain-computer interfaces and personalized learning. Ethical considerations regarding neural data privacy are crucial for responsible applications.

References

[1] Hamad Ahmed, Ronnie B. Wilbur, Hari M. Bharadwaj, and Jeffrey Mark Siskind. Object classification from randomized EEG trials. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3845–3854. Computer Vision Foundation / IEEE, 2021. 2

[2] A. Author and B. Author. Eeg2text: Decoding text from brain signals. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 12345–12353, 2019. 2

[3] Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: toward real-time reconstruction of visual perception. *arXiv preprint arXiv:2310.19812*, 2023. 1

[4] Hari M. Bharadwaj, Ronnie B. Wilbur, and Jeffrey Mark Siskind. Still an ineffective method with supertrials/erps - comments on "decoding brain representations by multimodal learning of neural activity and visual features". *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(11):14052–14054, 2023. 2

[5] Jiahang Cao, Xu Zheng, Yuanhuiyi Lyu, Jiaxu Wang, Renjing Xu, and Lin Wang. Chasing day and night: Towards robust and efficient all-day object detection guided by an event camera. *arXiv preprint arXiv:2309.09297*, 2023. 2

[6] Jialei Chen, Daisuke Deguchi, Chenkai Zhang, Xu Zheng, and Hiroshi Murase. Clip is also a good teacher: A new learning framework for inductive zero-shot semantic segmentation. *arXiv preprint arXiv:2310.02296*, 2023. 2

[7] Michael X Cohen. Where does eeg come from and what does it mean? *Trends in neurosciences*, 40(4):208–218, 2017. 1

[8] Michelle E Costanzo, Joseph J McArdle, Bruce Swett, Vladimir Nechaev, Stefan Kemeny, Jiang Xu, and Allen R Braun. Spatial and temporal features of superordinate semantic processing studied with fmri and eeg. *Frontiers in human neuroscience*, 7:293, 2013. 6

[9] Alessandro T Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M Cichy. A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022. 2

[10] Alessandro T Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M Cichy. A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022. 1

[11] Tijl Grootswagers, Ivy Zhou, Amanda K Robinson, Martin N Hebart, and Thomas A Carlson. Human eeg recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Scientific Data*, 9(1):3, 2022. 1, 2, 5, 6

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 8

[13] Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):1–13, 2018. 1, 2

[14] Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. Zuco 2.0: A dataset of physiological recordings during natural reading and annotation. *arXiv preprint arXiv:1912.00903*, 2019. 1, 2

[15] Nora Hollenstein, Cedric Renggli, Benjamin Glaus, Maria Barrett, Marius Troendle, Nicolas Langer, and Ce Zhang. Decoding eeg brain activity for multi-modal natural language processing. *Frontiers in Human Neuroscience*, 15:659410, 2021. 6

[16] Helene Intraub. Rapid conceptual identification of sequentially presented pictures. *Journal of Experimental Psychology: Human Perception and Performance*, 7(3):604, 1981. 5

[17] Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, and Mubarak Shah. *Brain2Image*: Converting brain signals into images. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 1809–1817. ACM, 2017. 1, 2

[18] Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, and Mubarak Shah. Brain2image: Converting brain signals into images. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1809–1817, 2017. 1, 2

[19] Christian Keysers, D-K Xiao, Peter Földiák, and David I Perrett. The speed of sight. *Journal of cognitive neuroscience*, 13(1):90–101, 2001. 5

[20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2, 3

[21] Pradeep Kumar, Rajkumar Saini, Partha Pratim Roy, Pawan Kumar Sahu, and Debi Prosad Dogra. Envisioned speech recognition using eeg sensors. *Personal and Ubiquitous Computing*, 22:185–199, 2018. 2

[22] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018. 8

[23] Ren Li, Jared S. Johansen, Hamad Ahmed, Thomas V. Ilyevsky, Ronnie B. Wilbur, Hari M. Bharadwaj, and Jeffrey Mark Siskind. The perils and pitfalls of block design for EEG classification experiments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(1):316–333, 2021. 2

[24] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755. Springer, 2014. 2

[25] Yuanhuiyi Lyu, Xu Zheng, Dahun Kim, and Lin Wang. Omnibind: Teach to build unequal-scale modality interaction for omni-bind of all. *arXiv preprint arXiv:2405.16108*, 2024. 2

[26] Yuanhuiyi Lyu, Xu Zheng, and Lin Wang. Image anything: Towards reasoning-coherent and training-free multi-modal image generation. *arXiv preprint arXiv:2401.17664*, 2024. 2

[27] Amanda K Robinson, Praveen Venkatesh, Matthew J Boring, Michael J Tarr, Pulkit Grover, and Marlene Behrmann. Very high density eeg elucidates spatiotemporal aspects of early visual processing. *Scientific reports*, 7(1):16248, 2017. 5

[28] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 16(5): 051001, 2019. 1

[29] Maham Saeidi, Waldemar Karwowski, Farzad V Farahani, Krzysztof Fiok, Redha Taiar, Peter A Hancock, and Awad Al-Juaid. Neural decoding of eeg signals with machine learning: A systematic review. *Brain Sciences*, 11(11):1525, 2021. 1

[30] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 8

[31] Prajwal Singh, Pankaj Pandey, Krishna Miyapuram, and Shanmuganathan Raman. Eeg2image: image reconstruction from eeg brain signals. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1

[32] Prajwal Singh, Dwip Dalal, Gautam Vashishtha, Krishna Miyapuram, and Shanmuganathan Raman. Learning robust deep visual representations from eeg brain recordings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7553–7562, 2024.

[33] Michal Teplan et al. Fundamentals of eeg measurement. *Measurement science review*, 2(2):1–11, 2002. 1

[34] Praveen Tirupattur, Yogesh Singh Rawat, Concetto Spampinato, and Mubarak Shah. Thoughtviz: Visualizing human thoughts using generative adversarial network. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 950–958, 2018. 1, 2, 9

[35] David Vivancos and Felix Cuesta. Mindbigdata 2022 A large dataset of brain signals. *CoRR*, abs/2212.14746, 2022. 2

[36] Sheng Wong, Anj Simmons, Jessica Rivera-Villicana, Scott Barnett, Shobi Sivathamboo, Piero Perucca, Zongyuan Ge, Patrick Kwan, Levin Kuhlmann, Rajesh Vasa, et al. Eeg datasets for seizure detection and prediction—a review. *Epilepsia Open*, 8(2):252–267, 2023. 1

[37] Jonathan Xu, Bruno Aristimunha, Max Emanuel Feucht, Emma Qian, Charles Liu, Tazik Shahjahan, Martyna Spyra, Steven Zifan Zhang, Nicholas Short, Jioh Kim, Paula Perdomo, Ricky Renfeng Mao, Yashvir Sabharwal, Michael Ahedor Moaz Shoura, and Adrian Nestor. Alljoined - A dataset for eeg-to-image decoding. *CoRR*, abs/2404.05553, 2024. 2

[38] Xu Zheng and Lin Wang. Eventdance: Unsupervised source-free cross-modal adaptation for event-based object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17448–17458, 2024. 2

[39] Xu Zheng, Yexin Liu, Yunfan Lu, Tongyan Hua, Tianbo Pan, Weiming Zhang, Dacheng Tao, and Lin Wang. Deep learning

for event-based vision: A comprehensive survey and benchmarks. *arXiv preprint arXiv:2302.08890*, 2023.

[40] Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. E-clip: Towards label-efficient event-based open-world understanding by clip. *arXiv preprint arXiv:2308.03135*, 2023.

[41] Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. Exact: Language-guided conceptual reasoning and uncertainty estimation for event-based action recognition and more. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18633–18643, 2024. 2

EDPNet: An Efficient Dual Prototype Network for Motor Imagery EEG Decoding

Can Han[a], Chen Liu[a], Crystal Cai[a], Jun Wang[b,*], Dahong Qian[a,*]

[a]School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China
[b]School of Computer and Computing Science, Hang Zhou City University, Hangzhou, 310015, China

Abstract

Motor imagery electroencephalograph (MI-EEG) decoding plays a crucial role in developing motor imagery brain-computer interfaces (MI-BCIs). However, decoding intentions from MI remains challenging due to the inherent complexity of EEG signals relative to the small-sample size. In this paper, we propose an Efficient Dual Prototype Network (EDPNet) to enable accurate and fast MI decoding. EDPNet employs a lightweight adaptive spatial-spectral fusion module, which promotes more efficient information fusion between multiple EEG electrodes. Subsequently, a parameter-free multi-scale variance pooling module extracts more comprehensive temporal features. Furthermore, we introduce dual prototypical learning to optimize the feature space distribution and training process, thereby improving the model's generalization ability on small-sample MI datasets. Our experimental results show that the EDPNet outperforms state-of-the-art models with superior classification accuracy and kappa values (84.11% and 0.7881 for dataset BCI competition IV 2a, 86.65% and 0.7330 for dataset BCI competition IV 2b). Additionally, we use the BCI competition III IVa dataset with fewer training data to further validate the generalization ability of the proposed EDPNet. We also achieve superior performance with 82.03% classification accuracy. Benefiting from the lightweight parameters and superior decoding accuracy, our EDPNet shows great potential for MI-BCI applications. The code is publicly available at https://github.com/hancan16/EDPNet.

Keywords: prototype learning, attention mechanism, lightweight, brain-computer interface, motor imagery

1. Introduction

Brain-computer interface (BCI) systems enable non-muscular communication between users and machines by interpreting users' neural activity patterns [1]. In BCI applications, Electroencephalography (EEG) has become increasingly popular due to its non-invasive nature and cost-effectiveness. Motor Imagery (MI) [2] is the mental rehearsal of movement execution without any physical movement. When participants visualize moving parts of their body, specific areas of the brain experience energy changes known as event-related desynchronization/synchronization (ERD/ERS). These changes can be recorded via EEG and used to discriminate motor intent [3, 4]. The MI-based BCI has garnered significant attention as it enables the decoding of user motor intentions from EEG signals. It has been successfully applied in various fields, such as stroke rehabilitation [5], wheelchair control [6], and exoskeleton robot arm control [7].

Advancements in deep learning (DL) have significantly increased the accuracy of decoding EEG signals for MI-based BCI applications [8, 9, 10, 11], yet several issues still hinder EEG-based models from reaching practical use [12]. When developing a practical and accurate EEG-MI decoding algorithm, several key challenges need to be taken into consideration.

- **Complex characteristics of EEG.** EEG signals are contaminated with noises and artifacts, leading to a low signal-to-noise ratio (SNR). Moreover, complex spatial-spectral coupling characteristics and high temporal vari-

*Corresponding author.
Email addresses: hancan@sjtu.edu.cn (Can Han), wangjun@hzcu.edu.cn (Jun Wang), dahong.qian@sjtu.edu.cn (Dahong Qian)

ability further complicate the decoding of MI-EEG signals [13]. Therefore, extracting discriminative features from EEG signals is challenging yet crucial.

- **Limited data.** EEG signals frequently encounter constraints due to a scarcity of training samples, caused by several issues such as cumbersome calibration procedures, uncertainty in annotations due to variability in participants' responses to MI tasks, and data privacy issues [14, 15]. Without a massive amount of training data, the model may overfit. This poses challenges to the model's generalization ability on new test data.
- **Computational cost.** In practical BCI applications, computational resources are often limited. Therefore, lightweight and fast models are more suitable for practical scenarios[16].

Existing research primarily focuses on addressing one of the aforementioned challenges. Some research efforts leverage advanced DL techniques, such as multi-branch designs [17, 18, 19], transformers [20, 21, 22, 23, 24], and attention mechanisms [16, 25, 26, 27], to extract highly discriminative features from EEG data, thereby improving EEG-MI decoding accuracy. However, these methods overlook the overfitting issue caused by limited training data and tend to have high computational complexity. Other research efforts [28, 29, 30, 31] employ transfer learning (TL) to mitigate the small-sample issue. Nevertheless, these TL methods still require a relatively large amount of additional data from other subjects to achieve good performance, which may not be practical in real-world scenarios. The Sinc-ShallowNet proposed by Borra et al. [32] offers advantages in lightweight design and interpretability, but its decoding accuracy is unsatisfactory due to the lack of effective mechanisms for extracting discriminative features. Because of incomplete consideration and resolution of these challenges, there remains a gap between existing research and practical applications.

Considering all the above challenges, this paper aims to design a lightweight neural network architecture that effectively extracts highly discriminative and robust features from complex EEG signals for accurate MI classification, even with limited training data. To achieve this goal, we propose an Efficient Dual Prototype Network (EDPNet), inspired by the recognition mechanism of the human brain. Human brains can establish cognitive understanding from a small amount of learnable data and effectively generalize it to new data based on memories and template/prototype matching. The EDPNet is composed of two main components, i.e., a feature extractor and the prototypes for all MI classes. The feature extractor simulates the sensory system of humans for transforming original data into abstract representations. Moreover, the prototypes for each class act as abstract memories of the corresponding class in our brains. As in human recognition, the decision in EDPNet is made by matching the feature (abstract representation) with prototypes (memories) of each class.

For the feature extractor component of EDPNet, based on ERD/ERS prior knowledge, we design two novel modules: Adaptive Spatial-Spectral Fusion (ASSF) module and Multi-scale Variance Pooling (MVP). The ASSF module focuses on modeling the relationship between EEG electrode channels that reflect levels of brain activation [33]. Equipped with a lightweight attention mechanism, the ASSF module can adaptively adjust the weights of each EEG channel according to its importance, thereby effectively extracting spatial-spectral features relevant to specific MI tasks. Then, the MVP module captures multi-scale long-term temporal features based on signal variance which represents the EEG spectral power [34]. The MVP module has no trainable parameters and is computationally efficient, serving as a superior method for extracting powerful temporal features in MI-EEG decoding tasks. The combination of the ASSF module and MVP module enables the feature extractor of EDPNet to extract discriminative spatial-spectral-temporal features from the complex EEG signals.

Furthermore, to address the small-sample dilemma in EEG-MI decoding, we design a new prototype learning (PL) approach to optimize the distribution of prototypes and features, aiming to obtain a robust feature space. To the best of our knowledge, this is the first study to apply the PL to MI-EEG decoding. The classic PL method [35] employs a prototype loss to push feature vectors towards corresponding prototypes, making the features within the same class more compact, which is beneficial for classification and model generalization. Based on the classic PL method, we propose a Dual Prototype Learning (DPL) approach for our EDPNet to decouple inter-class separation and intra-class compactness in training processes. The DPL not only enhances intra-class compactness, but also explicitly increases inter-class margins. Compared to the classic PL, our DPL further improves the model's generalization capability.

The major contributions of this paper can be summarized as follows:

1. Inspired by clinical prior knowledge of EEG-MI and human brain recognition mechanisms, we propose a high-performance, lightweight, and interpretable MI-EEG decoding model EDPNet. The EDPNet simultaneously

considers and overcomes three major challenges in MI-BCIs.

2. To extract highly discriminative features from EEG signals, we design two novel modules, ASSF and MVP, for the feature extractor of EDPNet. The ASSF module extracts effective spatial-spectral features, and the MVP module extracts powerful multi-scale temporal features.

3. To overcome the small-sample issue of MI tasks, we propose a novel DPL approach to optimize the distribution of features and prototypes, aiming to obtain a robust feature space. This enhances the generalization capability and classification performance of our EDPNet.

4. We conduct experiments on three benchmark public datasets to evaluate the superiority of the proposed EDPNet against state-of-the-art (SOTA) MI decoding methods. Additionally, comprehensive ablation experiments and visual analysis demonstrate the effectiveness and interpretability of each module in the proposed EDPNet.

2. Related Works

2.1. Deep Learning based EEG-MI Decoding

With recent advancements in deep learning, researchers are increasingly using various deep learning architectures to decode EEG signals. DeepConvNet [8] employed multiple convolutional layers with temporal and spatial feature extraction kernels. Sakhavi et al. [9] utilized FBCSP for feature extraction, followed by CNN-based classification. Lawhern et al. [10] introduced a compact network, EEGNet, employing depthwise and separable convolution for spatial-temporal feature extraction. However, due to the lack of effective mechanisms for extracting highly discriminative features, the improvements of these methods are limited.

Attention mechanisms, which have recently gained significant recognition in various fields, have been successfully applied to MI-EEG decoding. TS-SEFFNet [25] combines a channel attention module based on the wavelet packet sub-band energy ratio with a temporal attention mechanism, followed by a feature fusion architecture. LMDA-Net [16] combines a custom channel recalibration module with a feature channel attention module from ECA-Net [36]. Wimpff et al. [26] applied various attention mechanisms to the proposed BaseNet and made a very comprehensive comparison of the different variations. M-FANet [27] uses a convolution with a small kernel size to extract local spatial information and a SE [37] module to extract information from multiple perspectives. These methods mainly apply attention mechanisms to deep features extracted by the neural network and improve the MI decoding accuracy to some extent. Nonetheless, few studies have employed attention mechanisms to model the relationships between EEG electrode channels that reflect levels of brain activation [33].

Besides, research efforts have also been devoted to extracting more effective temporal features from high temporal resolution EEG signals. Lately, transformer models have made waves in natural language and computer vision due to the inherent perception of global dependencies [38]. Transformers also emerged in MI decoding and achieved good performance [20, 21, 22, 23, 24], by leveraging long-term temporal relationships. ATCNet [22] uses self-attention to highlight the most important information in EEG signals. Conformer [23] stacks transformer blocks to extract long-term dependency features based on local temporal features extracted by CNN. However, transformer models have high parameters and computational costs, making them hard to be used for real-time MI decoding.

2.2. Prototype Learning

PL simulates the way humans learn by memorizing typical examples to understand and generalize to new situations. In PL methods, a set of representative samples (prototypes) is learned during training, and during testing test samples are assigned to the closest prototype to determine their categories. In [39], Snell et al. proposed to apply the prototype concept for few-shot learning. However, this method learns the prototypes and the feature extractor separately using discriminative loss. Yang et al. [35] introduced the prototype model into the DL paradigm and designed different discriminative loss as well as generative loss. This significantly improves the performance and robustness of DL models in classification tasks. Following the study [35], a substantial amount of research [40, 41, 42] has been devoted to using PL to learn a compact feature space for addressing open-world recognition. Another line of research [43, 44] continues to explore the potential of PL in few-shot learning.

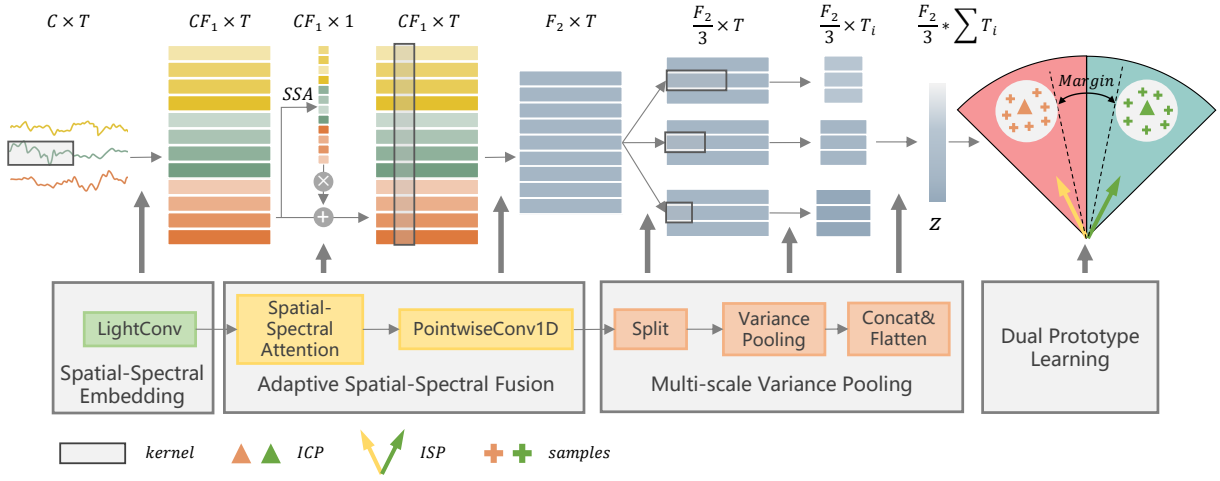


Figure 1: The overall framework of the proposed EDPNet. C and T denote the number of EEG channels and the number of time points, respectively. F_1 and F_2 denote the numbers of temporal filters and spatial-spectral filters, respectively. T_i represents the output length of the variance layer with different kernel sizes. The SSE, ASSF, and MVP make up the feature extractor, while DPL is used for training optimization and classification.

3. Method

As shown in Figure 1, our proposed EDPNet consists of four modules. The Spatial-Spectral Embedding (SSE) module, the Adaptive Spatial-Spectral Fusion module, and the Multi-scale Variance Pooling module constitute the feature extractor for extracting highly discriminative features. The Dual Prototype Learning module is used to optimize the feature space and make classification decisions.

3.1. EEG Representation

In this paper, we feed raw MI-EEG signals into the proposed model without any additional time-consuming preprocessing. Given a set of m labeled MI trials $S = \{X_i, y_i\}_{i=1}^{m}$, where $X_i \in \mathbb{R}^{C \times T}$ consists of C channels (EEG electrodes) and T time points, $y_i \in \{1, ..., n\}$ is the corresponding class label, and n is the total number of predefined classes for set S, our EDPNet model first maps a motor imagery trail X_i to the feature space Z and obtains $z_i = f(X_i) \in \mathbb{R}^d$, where f is the feature extractor, as shown in Figure 1. Then, the DPL module maps the feature z_i to its corresponding class y_i.

3.2. Feature Extractor of EDPNet

3.2.1. Spatial-Spectral Embedding

Since different MI classes may differ in their corresponding spectral-spatial sensorimotor rhythm (SMR) patterns [10], most existing studies first extract multi-view spectral information from each EEG electrode channel to form a spatial-spectral representation. Some works follow the practice of EEGNet [10] and use a 2D convolution to extract spectral features, while others follow the practice of FBCNet [9] and use multiple narrow-band digital filters to manually extract different spectral features. Our SSE module is similar to the former, but uses a 1D convolution for end-to-end extraction of spectral features.

Unlike these popular methods [16, 22, 23, 25, 26, 27] following EEGNet, we do not incorporate an additional feature dimension to create a 2D representation $X_i \in \mathbb{R}^{1 \times C \times T}$ for the raw EEG signal $X_i \in \mathbb{R}^{C \times T}$. We directly treat the EEG electrode dimension C in $X_i \in \mathbb{R}^{C \times T}$ as the feature dimension and use 1D convolution to extract spectral features. Specifically, we introduce the LightConv [45], a depthwise convolution that shares certain output channels, to act as the 1D temporal convolution. LightConv first divides the input signal $X_i \in \mathbb{R}^{C \times T}$ into h groups along the channel dimension. Therefore, each group has C/h channels, and each channel within the same group shares convolutional

weights. The implementation steps are as follows:

$$X_h = \text{Reshape}(X_i) \in \mathbb{R}^{(C/h)\times h\times T} \tag{1}$$

$$X_{dw} = \text{DWConv1D}(X_h, W) \in \mathbb{R}^{(C/h)\times(h*F_1)\times T} \tag{2}$$

$$X_{sse} = \text{Reshape}(X_{dw}) \in \mathbb{R}^{CF_1\times T}, \tag{3}$$

where DWConv1D denotes 1D depthwise convolution. Additionally, $W \in \mathbb{R}^{(h*F_1)\times 1\times k}$ is the learnable convolution parameter, and $X_{sse} \in \mathbb{R}^{CF_1\times T}$ corresponds to the resultant spatial-spectral embedding. It is important to note that W contains $h * F_1$ filters, as each channel uses F_1 filters with kernel size k to generate different spectral characteristics. Compared to depthwise convolution, LightConv reduces the number of parameters by a factor of C/h.

Moreover, by setting different h values and arranging electrode channels, different brain regions can be easily decoded by different temporal filters, which helps extract more comprehensive information. By setting $h = C$ in Eq. (1), the LightConv is equivalent to a depthwise convolution, where each electrode channel uses different filters to extract spectral features. To reduce the number of parameters and accelerate training, we set h as 1 in this paper. All electrode channels share F_1 temporal filters to produce spatial-spectral embedding X_{sse}, as shown in Figure 1.

3.2.2. Adaptive Spatial-Spectral Fusion

An EEG equipment uses multiple electrodes distributed in different regions of the cerebral cortex to capture neuronal activity in these brain regions. When performing different MI tasks, the EEG signal amplitudes recorded by different electrodes may increase or decrease in specific spectral bands [46]. This phenomenon is known as ERD and ERS. Therefore, it is critical to emphasize the relationship between different spectral features among multiple EEG electrodes. When extracting features in the spatial-spectral dimension, each channel should not be treated as equal, but rather, focus should be placed on areas and spectral bands relevant to the specific MI task. Consequently, leveraging the attention mechanism, we design a Spatial-Spectral Attention (SSA) to extract effective spatial-spectral information for all electrodes.

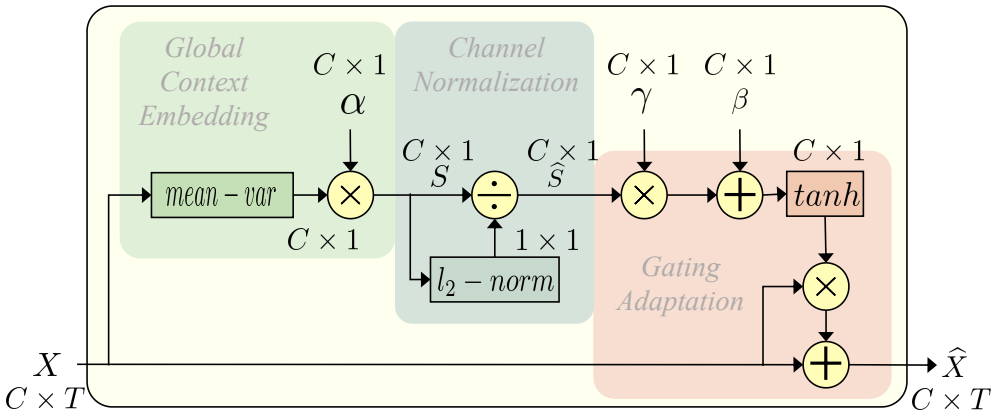


Figure 2: The structure of the proposed spatial-spectral attention.

Inspired by the gated channel transformation [47], our spatial-spectral attention consists of three parts: global context embedding, channel normalization, and gating adaptation, as shown in Figure 2. For an input $X \in \mathbb{R}^{C\times T}$, global context embedding employs a mean-var operation to aggregate temporal information from each channel. This involves calculating the variance within each 1-second window, followed by averaging them. Subsequently, α is responsible for controlling the weight of each channel:

$$s = \alpha \cdot \text{mean-var}(X). \tag{4}$$

Then, we use a channel normalization component to model channel relations:

$$\hat{s} = \frac{\sqrt{C}s}{\|\mathbf{s}\|_2} = \frac{\sqrt{C}s}{[(\sum_{c=1}^{C} s^2) + \epsilon]^{\frac{1}{2}}}, \tag{5}$$

where ϵ is a small constant to avoid the problem of derivation at the zero point. The gating weight and bias, γ and β are responsible for adjusting the scale of the input feature channel-wise:

$$\text{Attention} = 1 + \tanh(\gamma\hat{s} + \beta) \tag{6}$$

$$\hat{X} = X \cdot \text{Attention}. \tag{7}$$

The scale of each channel of $X_{sse} \in \mathbb{R}^{CF_1 \times T}$ output by the SSE module will be adjusted by the corresponding attention weight. Additionally, SSA leverages global temporal information to model channel relationships and modulate feature maps on the channel-wise level. Therefore, we can effectively fuse the weighted spatial-spectral features using a simple 1D pointwise convolution. As shown in Figure 1, the pointwise convolution uses F_2 filters to simultaneously fuse spectral features of all electrodes to get $X_{assf} \in \mathbb{R}^{F_2 \times T}$.

3.2.3. Multi-scale Variance Pooling

It is crucial to acquire long-term dependencies and global temporal information for EEG decoding. Transformer-based [22, 20, 21, 23] models can capture global information well by using self-attention mechanism, but they have a large number of parameters and high computational complexity. In fact, under the constrained EEG training data, it is difficult for the transformer-based model to achieve optimal performance as in the computer vision field. Therefore, it is necessary to design a new method for EEG decoding that can extract long-term dependencies information.

Metaformer [48] has proposed that using a simple pooling layer in place of self-attention in transformers can also perform well. Therefore, we consider using a pooling layer with a large kernel to extract the global temporal information from the EEG signals. Inspired by [34] and considering that various classes of MI differ in their spectral power (ERD/ERS), a variance operation which represents the spectral power in the given time series becomes a more suitable option for EEG temporal characterization. In order to achieve this, we first design a 1D variance pooling layer, VarPool, which is more compatible with the neural network architecture. For a time series signal $x \in \mathbb{R}^t$, the relationship between its variance Dx and mean Ex is as follows:

$$\begin{aligned} Dx &= E(x - Ex)^2 \\ &= E(x^2) + (Ex)^2 - 2 * (Ex)^2 \\ &= E(x^2) - (Ex)^2. \end{aligned} \tag{8}$$

Therefore, for the EEG representation $X \in \mathbb{R}^{C \times T}$, we can utilize average pooling to calculate variance pooling, as

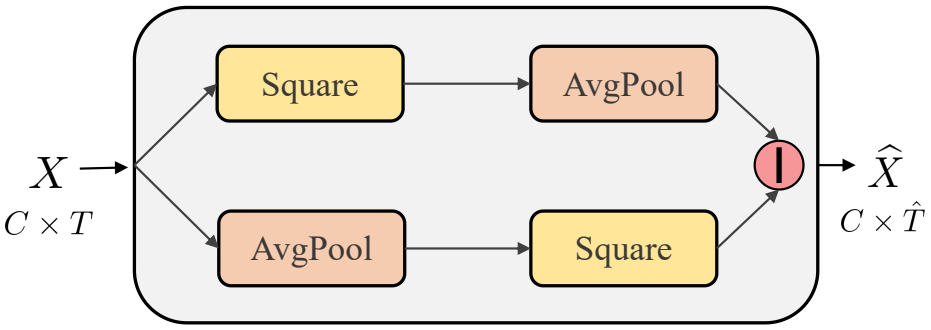


Figure 3: An illustration of the proposed VarPool layer.

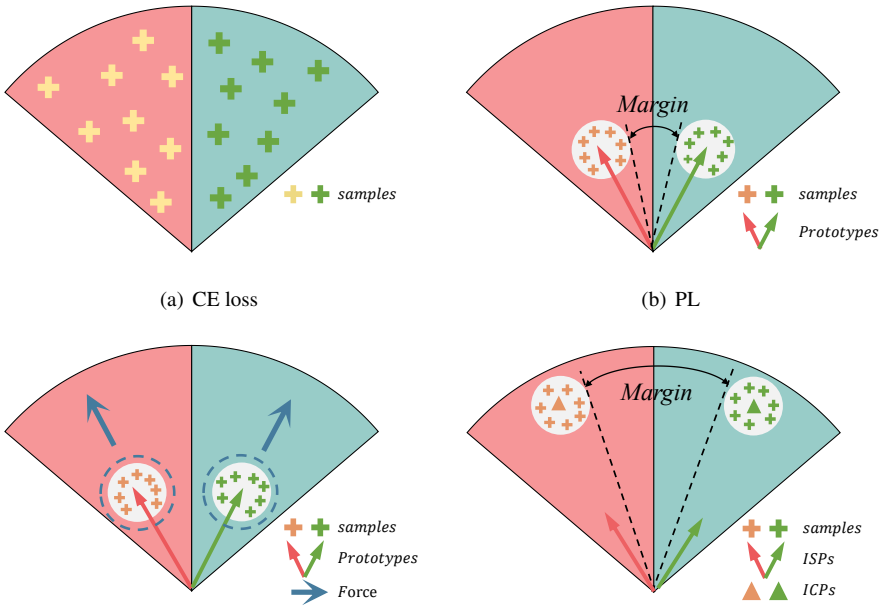
shown in Figure 3:

$$\mathrm{VarPool}(X)_{k,s} = \mathrm{AvgPool}(X^2)_{k,s} - \mathrm{AvgPool}(X)^2_{k,s}, \tag{9}$$

where k and s represent the specified window length and sliding step size. For an input $X \in \mathbb{R}^{C\times T}$, the VarPool layer slides along the time dimension to calculate the variance within each window to obtain the output $\hat{X} \in \mathbb{R}^{C\times \hat{T}}$:

$$\hat{T} = \left\lfloor \frac{T + 2 \times \mathrm{padding} - (k-1) - 1}{s} + 1 \right\rfloor. \tag{10}$$

Furthermore, to extract multi-scale long-term temporal features, we design the Multi-scale Variance Pooling layer. Specifically, the output X_{assf} of the ASSF module is split into three groups along the channel dimension. VarPool layers with different large kernel sizes (i.e., 50, 100, and 200) are used for each group to extract temporal features. Then, the outputs of the three groups are flattened and concatenated to obtain the final feature vector z_i. It is noteworthy that our MVP module contains no trainable parameters and is computationally efficient.



(a) CE loss (b) PL

(c) One way to improve PL (d) DPL

Figure 4: An illustration of the feature space distribution for CE loss, PL, and our proposed DPL.

3.3. Dual Prototype Learning

In classification tasks, existing methods input the feature vector z into a multi-layer perceptron (MLP) to obtain classification results, and optimize the parameters of the neural network using CE loss. Recent studies [21, 11] indicate that CE loss may be lacking in the effectiveness of reducing intra-class variation, especially when considering the non-stationarity of EEG signals. As shown in Figure 4 (a), CE loss only optimizes samples towards the decision boundary of the corresponding class, resulting in a loose distribution of sample features. When applying PL to classification, the first step is to assign a prototype to each class. A classification loss optimizes the sample features to be closest to its corresponding prototypes for classification. Additionally, a prototype loss is used to further push the sample features towards its corresponding prototypes, which can increase intra-class compactness as shown in Figure 4 (b), while also acting as a form of regularization to prevent model overfitting.

Although PL has been widely applied in the field of computer vision, its potential in EEG-MI decoding has been scarcely researched. PL methods focus on utilizing prototype loss to increase intra-class compactness, thereby implicitly enhancing inter-class distance to form a margin, as illustrated in Figure 4 (b). Benefitting from larger

margins, the PL methods outperform CE loss in both general classification tasks [35, 40] and few-shot learning [39]. Therefore, in this paper, we are dedicated to further increasing inter-class margins based on the PL method, in order to enhance the model's generalization capability in MI decoding tasks with small samples. A natural idea is to extend the clustered features along the direction of their corresponding prototypes, as shown in Figure 4 (c).

We achieve this goal using Dual Prototype Learning, ultimately obtaining a larger inter-class margin as shown in Figure 4 (d).

Specifically, we develop two prototypes for each class: the Inter-class Separation Prototype (ISP) and the Intra-class Compact Prototype (ICP), aiming to achieve inter-class separation and intra-class compactness, respectively. Based on the ISPs, we utilize softmax and CE loss to achieve inter-class separation:

$$\mathcal{L}_S(s,z) = -\frac{1}{m}\sum_{i=1}^{m} log\frac{e^{s_{y_i}\cdot z_i}}{\sum_{j=1}^{n} e^{s_j\cdot z_i}}, \tag{11}$$

where m is the number of training samples, n is the number of classes, z_i is the feature of the i-th sample, y_i is the corresponding label in range $[1, n]$, s represents the ISPs, and $s_j \in \mathbb{R}^d$ is the ISP of class j. Minimizing $\mathcal{L}_S$ can facilitate the separation of features from different classes, resulting in a feature space similar to Figure 4 (a).

Furthermore, we use intra-class compactness loss to compress the distance between samples belonging to the same class in the feature space, which is defined as:

$$\mathcal{L}_C(c,z) = \sum_{i=1}^{m} D(z_i, c_{y_i}), \tag{12}$$

where c represents the ICPs, $c_{y_i} \in \mathbb{R}^d$ is the ICP of class y_i, and D is the distance function. To prevent training oscillations and mitigate the influence of outlier samples, we use the Huber loss $\mathcal{L}_\delta(z,c)$ with $\delta = 1$ as the distance function $D(z,c)$, which is defined as:

$$\mathcal{L}_\delta(z,c) = \begin{cases} \frac{1}{2}(z-c)^2 & \text{if} \quad |z-c| \le \delta \\ \delta|z-c| - \frac{1}{2}\delta^2 & \text{if} \quad |z-c| > \delta \end{cases}. \tag{13}$$

$\mathcal{L}_C$ can represent the compactness of each class's feature vectors. By minimizing $\mathcal{L}_C$, we can increase the intra-class compactness so that features of the same class are clustered together like Figure 4 (b).

Previous PL methods use a single prototype for each class, as described in Figure 4 (b). In comparison, we decouple inter-class separation and intra-class compactness by using ISPs and ICPs. On one hand, this decouple enhances the robustness of the training process. On the other hand, it provides the conditions for further increasing the inter-class margins. Specifically, we apply an implicit force and an explicit force to the features to achieve the feature space optimization from Figure 4 (c) to Figure 4 (d).

- **Implicit force.** Due to the softmax's properties [49], the $\mathcal{L}_S$ tends to increase $s_{y_i}\cdot z_i$ until it converges to a constant value during training. This procedure simultaneously displaces the feature vector z_i and its corresponding ISP s_{y_i} away from the origin of the feature space until convergence. Furthermore, if we constrain the norm of ISPs to a smaller value, i.e., $\|s_i\|_2 \le S$ (weight-normalization), the feature vectors will be pushed further away from the origin. This constraint can act as an implicit force.

- **Explicit force.** To complement the implicit force, we design a simple loss function, $\mathcal{L}_{EF}$, to increase the norms of ICPs:

  $$\mathcal{L}_{EF}(c) = -\|c\|_2. \tag{14}$$

  By minimizing $\mathcal{L}_{EF}$, the norms of ICPs increase, thereby guiding the features to be pushed away from the origin.

During the training phase, the optimization objective of the proposed DPL is as follows:

$$\begin{aligned} &\text{minimize} \quad \mathcal{L}_S(s,z) + \lambda\mathcal{L}_C(c,z) + \alpha\mathcal{L}_{EF}(c) \\ &\text{subject to} \quad \|s_i\|_2 \le S, \quad \forall i = 1,2,\dots n \end{aligned}, \tag{15}$$

where λ and α are the trade-off scalar to balance the three losses, and S is set to 1. During the testing phase, the test sample X_i is classified by calculating the dot product between its feature vector z_i and the ISP for each class:

$$\hat{y}_i = \underset{j}{\operatorname{argmax}}(z_i \cdot s_j), \quad j = 1, 2, \ldots n, \tag{16}$$

where $\hat{y}_i$ denotes the predicted result.

4. Experiments and results

4.1. Evaluation Datasets

To demonstrate the effectiveness of our EDPNet, we evaluate it on two public MI-EEG datasets, namely, BCI Competition IV 2a (BCIC-IV-2a) [50] and BCI Competition IV 2b (BCIC-IV-2b) [51]. Additionally, we use the BCI competition III IVa (BCIC-III-IVa) [52] with fewer training data to further validate the generalization ability of the proposed method.

Dataset I. BCIC-IV-2a provided by Graz University of Technology consists of EEG data from 9 subjects. There were four motor imagery tasks, covering the imagination of moving the left hand, right hand, both feet, and the tongue. Two sessions on different days were collected with 22 Ag/AgCl electrodes at a sampling rate of 250 Hz. One session contained 288 EEG trials, i.e., 72 trials per task. We use [2, 6] seconds of each trial and all 22 electrodes in the experiments.

Dataset II. BCIC-IV-2b provided by Graz University of Technology consists of EEG data from 9 subjects. There were two motor imagery tasks, covering the imagination of moving left and right hand. Five sessions were collected with three bipolar electrodes (C3, Cz, and C4) at a sampling rate of 250 Hz and each session contained 120 trials. We use the [3, 7] seconds of each trial in the experiments.

Dataset III. BCIC-III-IVa, recorded at 100 Hz using 118 electrodes, contains 280 trials per subject and comprises two distinct classes: right hand, and foot. This dataset distinguishes itself from other datasets through its imbalanced division into training and testing trials. The quantity of training trials fluctuates between 28 and 224, varying with the subject (al: 224, aa: 168, av: 84, aw: 56, ay: 28), with the residual trials designated for testing. Each trial lasts 3.5 seconds. To preclude overfitting by reducing the number of data points per trial, we select the three channels shared (C3, Cz, and C4).

As the competition guidelines [51] for BCIC-IV-2a and BCIC-IV-2b datasets, we apply hold-out analysis to evaluate the performance of our EDPNet and all comparison methods. As such, the model is trained and tested completely in different sessions. This evaluation method is more in line with practical application scenarios and can better test the generalization ability of the model. For the BCIC-III-IVa dataset, we follow its official protocol to further validate the advantages of our method on small-sample training datasets.

4.2. Experimental Setups

4.2.1. Experimental Details

In this study, we implement our EDPNet using the PyTorch library, based on Python 3.10 with an Nvidia Geforce 2080Ti GPU. We use the AdamW optimizer with default settings (learning rate = 0.001, weight decay = 0.01) to train the feature extractor of our EDPNet. Additionally, we use another Adam optimizer to optimize ISPs and ICPs, with a learning rate of 0.001 on Datasets I and II, and a learning rate of 0.01 on the small-sample Dataset III. Moreover, for the hyperparameters of the model in Figure 1, on Dataset I and II, we empirically set the kernel size of LightConv as 75, F_1 and F_2 as 9 and 48, and the kernel sizes of different scales of the MVP layer as 50, 100, and 200. On Dataset III, due to the differences in sampling rate, we set the kernel size of LightConv as 50, and the kernel sizes of different scales of the MVP layer as 50, 100, and 150.

To prevent overfitting and reduce the number of epochs needed to train the model, a two-stage training strategy as in [8] is used in this work. Specifically, during the training phase, the training data is split into a training set and a validation set. In the first stage, only the training set is used, and the training is stopped if there is no decrease in the validation set loss for N_e consecutive epochs or reach the maximum training epoch N_1. During the second training stage, all training data are employed, then continue training for N_2 epochs. Due to the different sizes of the datasets used, we set N_1, N_e, and N_2 to be 1000, 200, and 300 respectively for Dataset I, and 300, 150, and 200, respectively, for Dataset II. For Dataset III, we set them to be 300, 150, and 150.

Table 1: Classification Accuracy(%) and Kappa Comparisons with SOTA Methods on Dataset I.

Methods	A01	A02	A03	A04	A05	A06	A07	A08	A09	Average	Std	Kappa	p-value
FBCSP [53]	76.00	56.50	81.25	61.00	55.00	45.52	82.75	81.25	70.75	67.75	12.89	0.5700	0.0020
EEGNet [10]	85.76	61.46	88.64	67.01	55.90	52.08	89.58	83.33	79.51	74.50	13.85	0.6600	0.0020
TS-SEFFNet [25]	82.29	49.79	87.57	71.74	70.83	63.75	82.92	81.53	81.94	75.17	11.32	0.6630	0.0020
LMDA-Net [16]	83.90	60.30	88.10	78.20	56.20	57.20	88.40	82.70	84.30	75.40	12.78	0.6700	0.0020
Basenet-SE [26]	81.60	52.08	90.28	73.96	76.39	62.85	86.81	80.56	79.51	76.00	11.22	0.6794	0.0020
M-FANet [27]	86.81	**75.00**	91.67	73.61	76.39	61.46	85.76	75.69	87.17	79.28	8.84	0.7259	0.0137
ATCNet [22]	86.81	68.40	92.01	73.61	78.82	62.15	86.46	87.15	83.33	79.86	9.37	0.7312	0.0020
Conformer[23]	87.85	54.86	86.46	76.04	58.33	59.72	89.58	83.33	81.25	75.27	13.06	0.6702	0.0020
FBMSNet [11]	87.85	66.32	92.36	76.74	72.57	62.15	80.21	86.46	87.85	79.17	9.91	0.7235	0.0020
EDPNet	**89.58**	71.88	**93.06**	**82.64**	**81.25**	**70.14**	**89.93**	**89.24**	**89.24**	**84.11**	**7.83**	**0.7881**	-

Best performances are highlighted in bold, while the second-best with underlined.

Table 2: Classification Accuracy(%) and Kappa Comparisons with SOTA Methods on Dataset II.

Methods	B01	B02	B03	B04	B05	B06	B07	B08	B09	Average	Std	Kappa	p-value
FBCSP [53]	70.00	60.36	60.94	**97.50**	93.12	80.63	78.13	92.50	86.88	80.01	13.06	0.6000	0.0059
EEGNet [10]	71.50	58.65	81.12	96.25	86.23	77.88	85.12	91.10	80.15	80.89	10.43	0.6321	0.0020
TS-SEFFNet [25]	72.81	65.71	75.75	96.25	91.25	85.00	88.63	91.87	82.18	83.27	9.51	0.6637	0.0020
LMDA-Net [16]	75.80	63.20	65.20	97.30	94.30	84.50	82.40	92.90	87.00	82.51	12.40	0.6500	0.0039
Basenet-SE [26]	72.50	67.86	81.13	96.86	93.44	84.69	88.75	**93.44**	84.69	84.82	9.20	0.6918	0.0057
ATCNet [22]	72.50	67.64	80.31	95.94	96.06	**88.12**	86.88	89.69	**90.94**	85.34	9.38	0.7068	0.0645
Conformer [23]	74.56	57.00	62.50	97.01	92.36	83.44	85.00	**93.44**	87.19	81.39	13.16	0.6265	0.0086
FBMSNet [11]	71.30	55.20	80.55	97.15	95.00	84.66	85.23	91.10	87.13	83.04	12.25	0.6692	0.0039
EDPNet	**77.50**	**68.93**	**82.81**	96.88	**96.25**	87.50	**89.06**	**93.44**	87.50	**86.65**	**8.58**	**0.7330**	-

Best performances are highlighted in bold, while the second-best with underlined.

4.2.2. Performance Metrics

In the experiments, the classification accuracy (ACC) and the Cohen's kappa coefficient (Kappa) are used as two metrics for performance evaluation. The mathematical formula of Cohen's kappa coefficient is defined as follows:

$$Kappa = \frac{P_0 - P_e}{1 - P_e}, \tag{17}$$

where p_0 represents the classification accuracy of the model and p_e represents the expected consistency level. Nevertheless, a one-sided Wilcoxon signed-rank test is used to verify the significance of improvement.

4.3. Overall Performance Comparison

We conduct extensive experiments and compare our method with numerous SOTA approaches across three public datasets. Table 1 displays the classification performance of all methods on Dataset I. Our EDPNet method achieves the highest average accuracy of 84.11% and the highest average kappa value of 0.7881. Moreover, Our method achieves a level of significance with $p < 0.05$ compared to all benchmark methods. The results demonstrate that the classification accuracy of our proposed EDPNet is not only 16.36% higher than the competition champion solution FBCSP ($p < 0.01$) but also significantly surpasses the classic EEGNet by 9.61% ($p < 0.01$). The four latest attention-based methods all apply the attention mechanism to deep feature dimensions. In contrast, our EDPNet utilizes lightweight attention in the spatial-spectral dimension. Consequently, our approach is more interpretable and has an accuracy of 4.83% higher than the best attention-based method M-FANet. Compared to transformer-based methods, we utilize a non-parametric and computationally efficient MVP module to extract long-term temporal features. The results demonstrate that our method achieves accuracy improvements of 4.25% and 8.84% compared to ATCNet and Conformer, respectively.

Table 3: Classification Accuracy(%) and Kappa Comparisons with SOTA Methods on Dataset III.

Methods	al	aa	av	aw	ay	Average	Kappa
	224/256	168/112	84/196	56/224	28/252		
EEG-Net [10]	100	68.75	58.16	79.46	51.59	71.59	0.4331
LMDA-Net [16]	100	70.13	61.33	78.94	52.11	72.50	0.4567
Basenet+SE [26]	100	79.46	<u>64.80</u>	75.89	<u>64.68</u>	<u>76.97</u>	<u>0.5374</u>
ATCnet [22]	100	75.00	61.22	79.02	53.57	73.76	0.4750
FBMSNet [11]	100	<u>82.14</u>	57.14	<u>82.04</u>	59.33	76.13	0.5270
EDPNet	100	**88.39**	**70.41**	**83.48**	**67.86**	**82.03**	**0.6426**

Best performances are highlighted in bold, while the second-best with underlined.

Compared to FBMSNet, our EDPNet not only increases intra-class compactness but also further enlarges inter-class margins, displaying an accuracy improvement of 4.94%.

The experimental results on Dataset II are presented in Table 2, which shows similar results to those on Dataset I. Our EDPNet also achieves the highest average accuracy of 86.65% with the smallest standard deviation and the highest average kappa value of 0.7330. And our method demonstrates significant advantages compared to most of the comparison methods. Notably, on Datasets I and II, our method achieves the best or second-best accuracy for nearly all subjects. Particularly, on Dataset I, for subjects A02 and A06 where other methods do not perform well, our method achieves an accuracy above 70%. As suggested in [11], a BCI system with > 70% binary classification accuracy is generally considered to be usable for healthy subjects and stroke patients. This demonstrates the potential of our EDPNet for MI-based BCI applications.

Moreover, on the smaller training Dataset III, we reproduce several SOTA methods suitable for comparison on this dataset. As shown in Table 3, the number of training samples for the 5 subjects in Dataset III decreases from 224 to 28. Our EDPNet achieves an average recognition accuracy of 82.03% and a kappa value of 0.6426, which are 5.06% and 0.1052 higher than the second-best method, respectively. Especially for the subject "ay" with only 28 training samples, our method still achieves a recognition accuracy of 67.86%. This experimental result demonstrates the superior generalization ability of our method in small-sample EEG decoding tasks.

4.4. Ablation Study

The significant improvement of our EDPNet can be attributed to three novel designs: the Adaptive-Spatial-Spectral fusion module, the Multi-scale Variance Pooling module, and the Dual Prototype Learning approach. To further analyze the impact of these three modules on model performance, we conduct ablation experiments on Datasets I and II. Four models, named Model1, Model2, Model3 and Model4, are utilized, which represent four scenarios as follows:

- **Model1** The model is realized by removing the ASSF module and adopting a depthwise convolution used in EEGNetto fuse the information between EEG electrodes.
- **Model2** This model is implemented by using a single kernel size (100) variance pooling layer, to verify the importance of multi-scale temporal information.
- **Model3** This model removes the DPL module and uses CE loss to optimize parameters.
- **Model4** This model removes the DPL module and uses the PL method [40] to optimize parameters.

Figure 5 shows the ablation data of the accuracy for each subject on Dataset I. It can be clearly seen that the ASSF module brings a 4.83% average accuracy improvement for Model1. This is because the ASSF module uses attention mechanisms to highlight specific spectral features of EEG electrodes related to the specific MI task and more effectively fuses spatial-spectral information. Similarly, the MVP can bring an average accuracy improvement of 2.98% for Model2, as it better adapts to changes in the length of MI-related activity segments between different

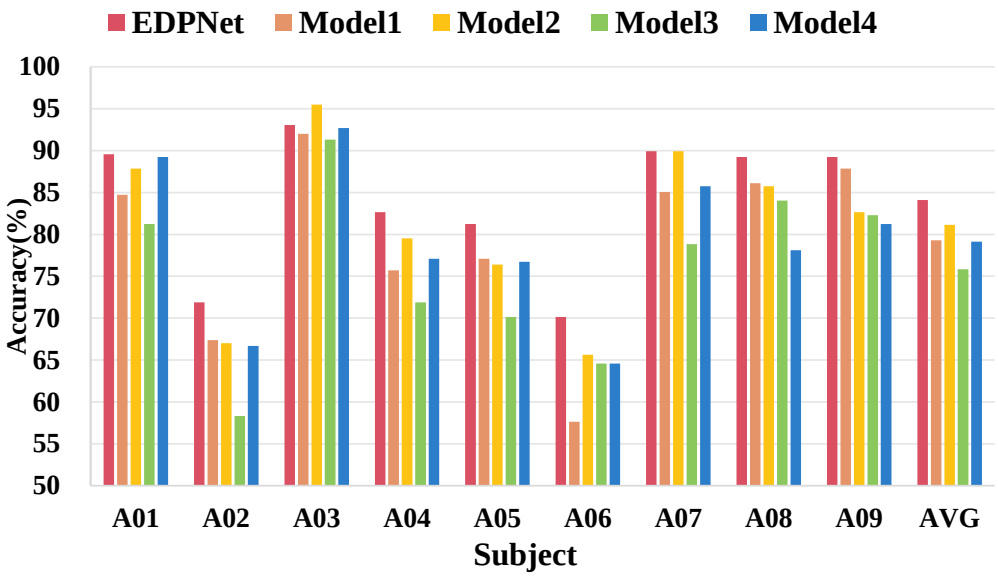


Figure 5: The accuracy comparison of each subject in Dataset I.

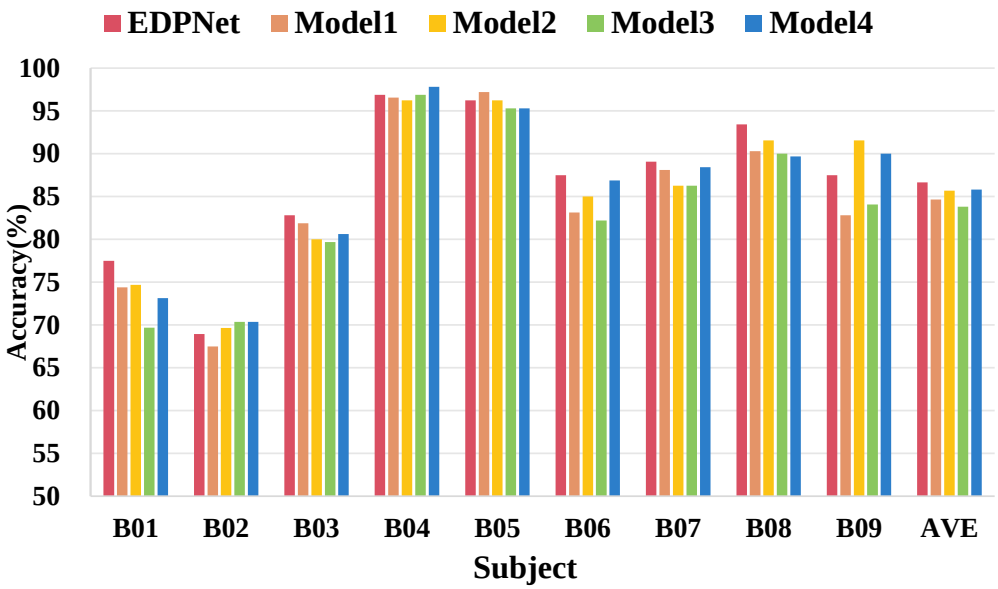
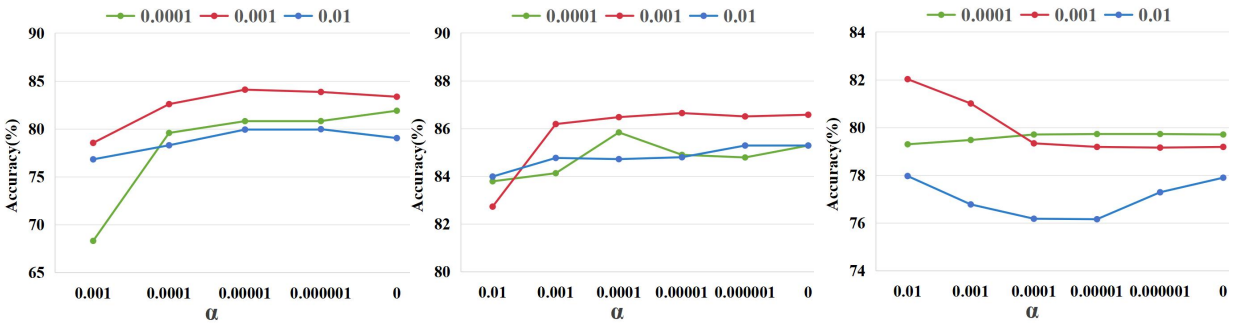


Figure 6: The accuracy comparison of each subject in Dataset II.

subjects. Employing dual prototype learning yields the most substantial enhancements. Compared with Model3, it not only results in an 8.26% average accuracy improvement, but also results in consistent improvement on each subject. Our DPL approach, when compared to CE loss, not only increases intra-class compactness but also further enlarges inter-class margins. This greatly enhances the model's generalization capability and recognition performance. Moreover, our EDPNet has improved the accuracy by 4.98% compared to Model4. This demonstrates that our DPL method has effectively improved upon the classical PL methods. A similar result is also observed on Dataset II. As shown in Figure 6, on Dataset II, our EDPNet achieves accuracy improvements of 2.00%, 0.96%, 2.82%, and 0.85% compared to Model1, Model2, Model3, and Model4, respectively.



(a) Dataset I (b) Dataset II (c) Dataset III

Figure 7: The accuracy of EDPNet across various settings of λ and α on three datasets.

4.5. Parameter Sensitivity

Our EDPNet employs a combination of $\mathcal{L}_s$, $\mathcal{L}_c$ and $\mathcal{L}_{EF}$ as the final loss function, as shown in Eq. (15). While the $\mathcal{L}_s$ loss aims to minimize misclassification of subject movement intent, the $\mathcal{L}_c$ loss minimizes the sum of the embedded space distance of samples in a class to its center, making the samples belonging to the same class compact in the feature space. And $\mathcal{L}_{EF}$ provides an explicit force, pushing the features away from the origin of the feature space to achieve larger inter-class margins. The λ and α are used to balance the impact of these three losses. To evaluate the influence of the λ and α, an empirical investigation compares the performance of EDPNet across various settings on all three datasets.

As shown in Figure 7, $\lambda = 0.001$ is the most suitable value across all three datasets. When λ is increased to 0.01 or decreased to 0.0001, there is a noticeable decrease in accuracy on Datasets I and II. When λ is fixed at 0.001 and α varies between 0 and 0.001, the accuracies are consistently high and reach the best performance at $\alpha = 0.00001$ on Dataset I and II. It is worth noting that even when $\alpha = 0$, the accuracies on Datasets I and II remain high. This indicates that our DPL can automatically learn larger inter-class margins relying solely on the implicit force. In contrast, on Dataset III, the best performance is achieved when α is increased to 0.01. This implies that increasing the $\mathcal{L}_{EF}$ can further improve classification accuracy on datasets with fewer training samples.

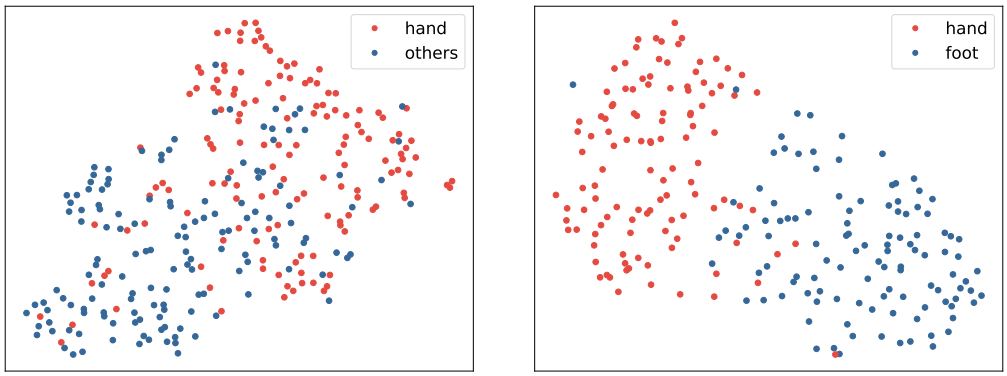
In summary, our EDPNet is relatively robust to the values of the hyperparameters λ and α. When $\lambda = 0.001$ and α is set to a small value (i.e., $\alpha < 0.001$), EDPNet can achieve good performance. If the amount of training data is very limited, further increasing α can be considered.

5. Further Analysis

5.1. Effect of Adaptive Spatial-Spectral Fusion

The key to our ASSF module lies in using the SSA to model the relationships between EEG electrodes. The signal amplitude of specific spectral bands of different EEG electrodes may increase or decrease when performing different MI tasks (such as imagining the hand movement and the foot movement). Our SSA exploits this phenomenon to help the model focus on it for classification. SSA generates adaptive attention weights in the spatial-spectral dimension based on the input EEG representations to re-weight the EEG representations. This enables the model to focus more on spatial-spectral features relevant to the current task.

To further verify the role of the SSA mechanism in imagining movements of different body parts, we use the t-SNE [54] method to visualize the attention vectors defined in Eq. (6) when performing different tasks. As shown in Figure 8(a), there are distinct differences in the distribution of attention vectors when A03 performs MI of the hand and MI of others on Dataset I. Similarly, when the subject "al" in Dataset III performing MI tasks of the hand and foot, the distribution of attention vectors also shows clear boundaries and clusters. This fully demonstrates the effectiveness and interpretability of our SSA mechanism.



(a) A03 in Dataset I (b) "al" in Dataset III

Figure 8: The distribution of attention vectors when performing different MI tasks on two datasets. All attention vectors are mapped to the 2D space using the t-SNE method.

5.2. *Effect of Multi-scale Variance Pooling*

Our MVP module innovatively uses a pooling layer with a large kernel size to extract long-term temporal information. Moreover, utilizing the crucial prior knowledge of spectral power in EEG signals, we design a variance pooling layer. This integrates the EEG prior into the architecture design of the neural network. Most importantly, compared to transformer-based models, our MVP module has no learnable parameters and is computationally efficient. The ablation experiments in Figure 5 and Figure 6 demonstrate that using only a single kernel size (100) for variance pooling achieves an accuracy of 81.13% and 85.69% on Datasets I and II, respectively. This result is superior to the comparison methods in Table 1 and Table 2.

Moreover, within one trial, the start point and duration of the actual MI period showing the appropriate ERS and ERD pattern can be different from trial to trial [20]. This phenomenon is more significant among trials between different subjects. In order to adapt to these differences between trials and extract more discriminative temporal information, we group the EEG representations along the channel dimension and use variance pooling with different kernel sizes to extract multi-scale temporal information. As shown in Figure 9, a smaller kernel size of 50 performs better on subjects A02, A04, and A09, while a larger kernel size of 200 performs better on subjects A05, A07, and A08. Altogether, the MVP integrates information from different scales and achieves the best overall performance across all subjects.

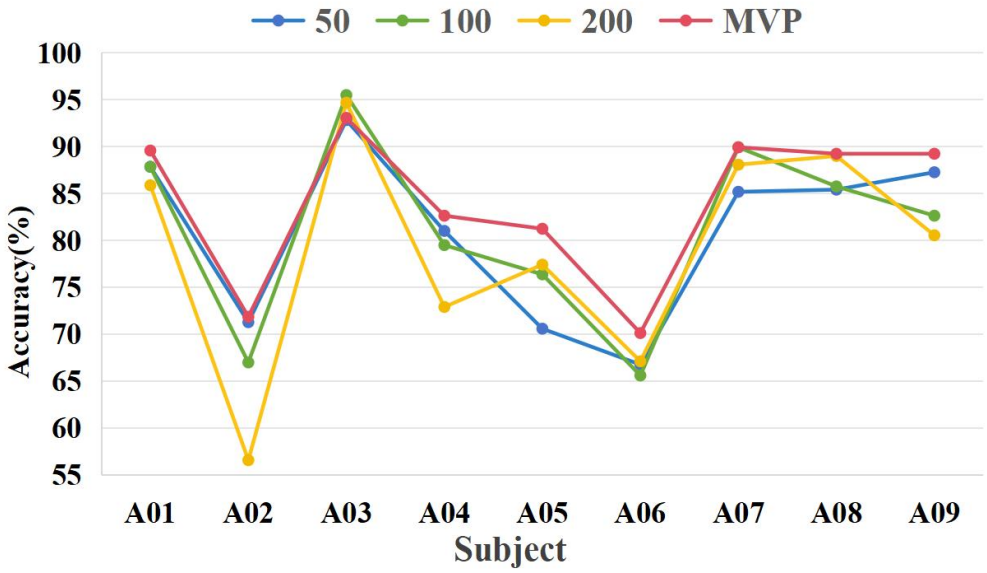


Figure 9: Comparison of the accuracy of single-scale variance pooling and MVP.

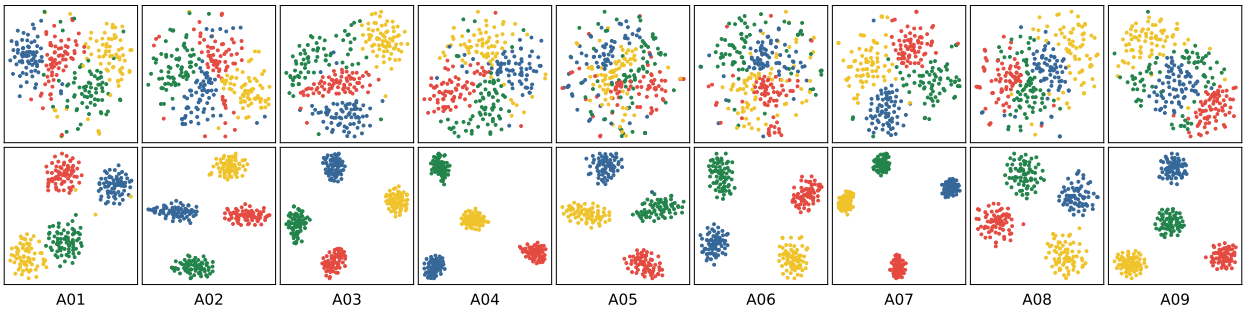


Figure 10: t-SNE visualization of the feature distribution for each subject trained on Dataset I using CE loss (first row) and DPL (second row). The points with different colors denote features from different classes.

5.3. Effect of Dual Prototype Learning

Our EDPNet is the first to apply prototype learning methods to MI-EEG decoding. Furthermore, we decouple inter-class separation and intra-class compactness by using two prototypes, ISP and ICP, for each class. Figure 5 and Figure 6 demonstrate that our DPL method significantly improves the recognition accuracy of MI tasks compared to CE loss. To further explain the effectiveness of DPL, we visualize the distribution of feature vectors z using the t-SNE method. Figure 10 displays the feature distribution of all subjects on Dataset I under both CE loss and DPL optimization approaches. It is apparent that our DPL method achieves greater intra-class compactness and larger inter-class margins compared to CE loss. Therefore, our DPL significantly enhances the model's generalization ability. This also intuitively explains why our DPL achieves better classification accuracy.

To verify that our DPL pushes the features further away compared to the PL method, thereby achieving larger inter-class margins, we visualize the feature norm distribution of our DPL method and the PL method. Figure 11 shows the L2 norm distribution of deep features trained with DPL and PL on Dataset I. It can be clearly seen that across all subjects, the feature norms trained using DPL are statistically larger than those trained using PL. This realizes the feature space optimization process from Figure 4 (b) to Figure4 (d), demonstrating the superiority of DPL.

5.4. Computational Expenses

As BCI systems typically operate in online or closed-loop mode on devices with limited computational resources [26], it is crucial to examine the computational expense of new algorithms. Table 4 displays the preprocessing methods, the classification accuracy, model parameters, and Floating Point Operations (FLOPs) for Dataset I. LMDA-Net needs to use Euclidean alignment (EA) to achieve high recognition performance, but EA is not suitable for real-time testing. FMBSNet and M-FANet require time-consuming multi-narrow-band band-pass filtering. In comparison, our EDPNet does not require any preprocessing steps and achieves the highest recognition accuracy with the lowest FLOPs. This suggests that our model optimally balances the accuracy and speed of MI-EEG decoding.

Table 4: Comparisons with SOTA Methods in the Computational Expenses and Recognition Performance on Dataset I.

Methods	Preprocessing	Acc(%)	Kappa	Parameters(k)	FLOPs(M)
LDMA-Net [16]	EA & BP	75.40	0.6700	**3.71**	50.38
FBMSNet [11]	MBP	79.17	0.7235	16.23	99.95
M-FANet [27]	MBP	79.28	0.7259	4.08	23.39
Conformer [23]	BP	75.27	0.6702	789.57	63.86
ADCNet [22]	no preprocessing	79.86	0.7312	113.73	29.81
EDPNet	no preprocessing	**84.11**	**0.7881**	15.21	**9.65**

BP: band-pass filtering, MBP: multi-narrow-band band-pass filtering.

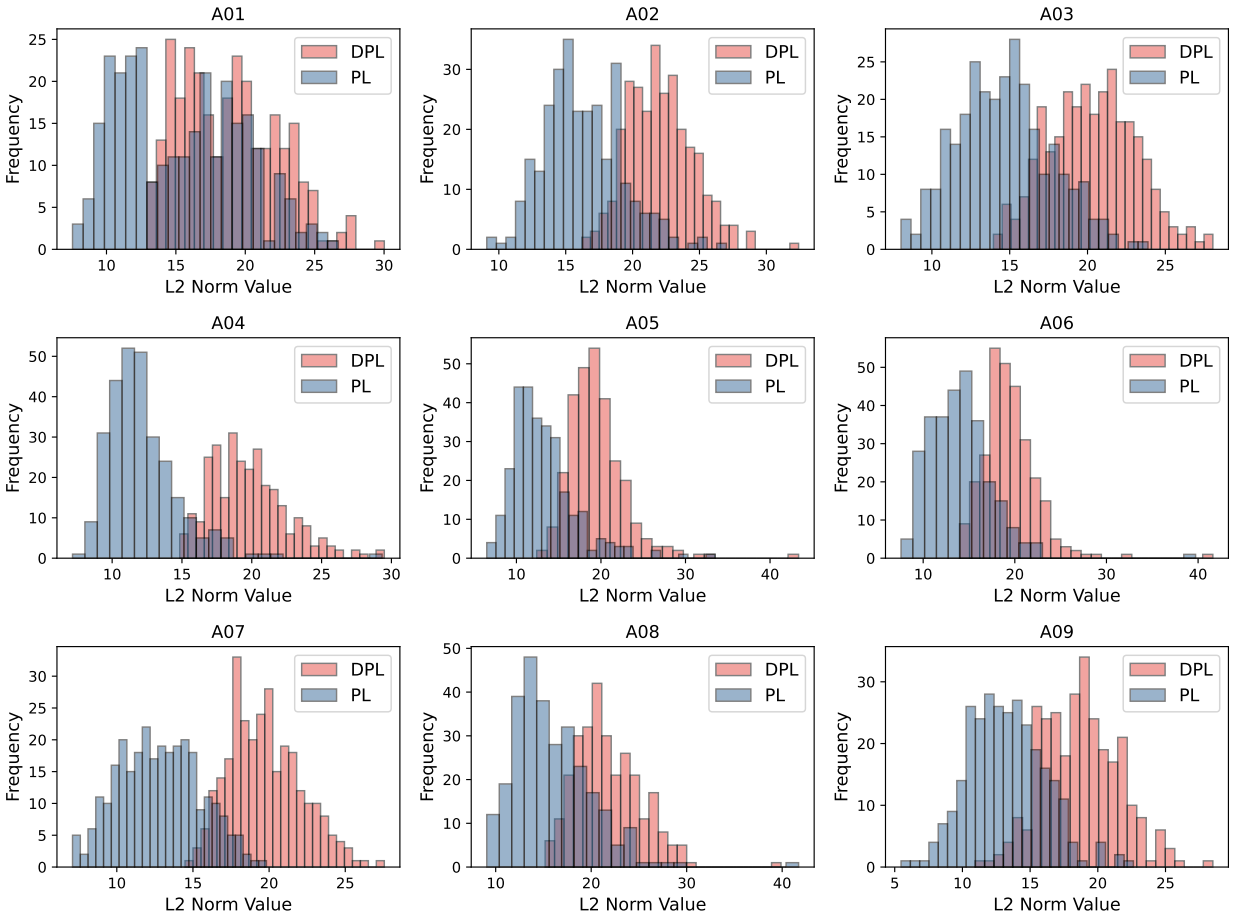


Figure 11: Histogram of the L2 norm distribution of features for each subject trained on Dataset I using PL and DPL, respectively.

Moreover, the number of parameters for our model is much less than transformer-based methods, and is on par with lightweight CNN methods.

5.5. Limitation and Future Work

Although the proposed EDPNet addresses major challenges in EEG-MI decoding and achieves excellent performance, there are some limitations in our current work. First, incorporating more prior knowledge into neural network design is worth exploring, such as considering the mirror distribution of EEG electrodes and the functional partitioning of the brain [24, 55]. Although the LightConv proposed in the SSE module could potentially leverage this prior knowledge, we did not further explore this aspect as it is not the focus of this work. Second, the potential of the brain-inspired DPL framework remains to be fully explored. By decoupling inter-class separation and intra-class compactness, we simply constrain the prototype and feature distribution to achieve superior performance. Nonetheless, more effective and discriminative loss functions deserve further investigation. Finally, EDPNet has only undergone offline testing on public datasets and has not yet been validated in an online BCI environment. In the future, we will continue to enhance EDPNet based on these avenues to achieve good performance in online BCI applications.

6. Conclusion

In this paper, we propose a lightweight and efficient dual prototype network for MI-EEG decoding. Based on neurophysiological priors and EEG data characteristics, we design the ASSF module and MVP module to extract

high discriminative features from EEG signals. The ASSF module utilizes a lightweight SSA mechanism to model the relationship between EEG electrodes for the extraction of powerful spatial-spectral features. Then, the MVP module is used to capture multi-scale long-term temporal features. Moreover, inspired by the recognition mechanism of the human brain, we propose a novel DPL approach to explicitly increase intra-class compactness and inter-class margins in the feature space. The DPL enhances the model's generalization capability, thereby helping to alleviate the limited sample issue. We conduct extensive experiments on three public datasets, and the results confirm that our method surpasses other SOTA methods. The proposed EDPNet holds promising potential for MI-based BCI applications due to its remarkable performance combined with low computational expenses.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by OYMotion Technologies.

References

[1] X. Gao, D. Xu, M. Cheng, S. Gao, A bci-based environmental controller for the motion-disabled, IEEE Transactions on neural systems and rehabilitation engineering 11 (2) (2003) 137–140.
[2] R. Abiri, S. Borhani, E. W. Sellers, Y. Jiang, X. Zhao, A comprehensive review of eeg-based brain–computer interface paradigms, Journal of neural engineering 16 (1) (2019) 011001.
[3] K. Sakai, K. Goto, J. Tanabe, K. Amimoto, K. Kumai, H. Kamio, Y. Ikeda, Effects of visual-motor illusion on functional connectivity during motor imagery, Experimental Brain Research 239 (7) (2021) 2261–2271.
[4] P. D. E. Baniqued, E. C. Stanyer, M. Awais, A. Alazmani, A. E. Jackson, M. A. Mon-Williams, F. Mushtaq, R. J. Holt, Brain–computer interface robotics for hand rehabilitation after stroke: A systematic review, Journal of neuroengineering and rehabilitation 18 (2021) 1–25.
[5] K. K. Ang, C. Guan, Eeg-based strategies to detect motor imagery for control and rehabilitation, IEEE Transactions on Neural Systems and Rehabilitation Engineering 25 (4) (2016) 392–401.
[6] J. Long, Y. Li, H. Wang, T. Yu, J. Pan, F. Li, A hybrid brain computer interface to control the direction and speed of a simulated or real wheelchair, IEEE Transactions on Neural Systems and Rehabilitation Engineering 20 (5) (2012) 720–729.
[7] B. J. Edelman, J. Meng, D. Suma, C. Zurn, E. Nagarajan, B. S. Baxter, C. C. Cline, B. He, Noninvasive neuroimaging enhances continuous neural tracking for robotic device control, Science robotics 4 (31) (2019) eaaw6844.
[8] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, T. Ball, Deep learning with convolutional neural networks for eeg decoding and visualization, Human brain mapping 38 (11) (2017) 5391–5420.
[9] S. Sakhavi, C. Guan, S. Yan, Learning temporal information for brain-computer interface using convolutional neural networks, IEEE transactions on neural networks and learning systems 29 (11) (2018) 5619–5629.
[10] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, B. J. Lance, Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces, Journal of neural engineering 15 (5) (2018) 056013.
[11] K. Liu, M. Yang, Z. Yu, G. Wang, W. Wu, Fbmsnet: A filter-bank multi-scale convolutional neural network for eeg-based motor imagery decoding, IEEE Transactions on Biomedical Engineering 70 (2) (2022) 436–445.
[12] X. Gu, F. Deligianni, J. Han, X. Liu, W. Chen, G.-Z. Yang, B. Lo, Beyond supervised learning for pervasive healthcare, IEEE Reviews in Biomedical Engineering (2023).
[13] X. Zhang, L. Yao, X. Wang, J. Monaghan, D. Mcalpine, Y. Zhang, A survey on deep learning-based non-invasive brain signals: recent advances and new frontiers, Journal of neural engineering 18 (3) (2021) 031002.
[14] S. Li, H. Wu, L. Ding, D. Wu, Meta-learning for fast and privacy-preserving source knowledge transfer of eeg-based bcis, IEEE Computational Intelligence Magazine 17 (4) (2022) 16–26.
[15] J. Han, X. Gu, G.-Z. Yang, B. Lo, Noise-factorized disentangled representation learning for generalizable motor imagery eeg classification, IEEE Journal of Biomedical and Health Informatics (2023).
[16] Z. Miao, M. Zhao, X. Zhang, D. Ming, Lmda-net: A lightweight multi-dimensional attention network for general eeg-based brain-computer interfaces and interpretability, NeuroImage 276 (2023) 120209.
[17] W. Tao, Z. Wang, C. M. Wong, Z. Jia, C. Li, X. Chen, C. P. Chen, F. Wan, Adfcnn: Attention-based dual-scale fusion convolutional neural network for motor imagery brain-computer interface, IEEE Transactions on Neural Systems and Rehabilitation Engineering (2023).
[18] X. Tang, C. Yang, X. Sun, M. Zou, H. Wang, Motor imagery eeg decoding based on multi-scale hybrid networks and feature enhancement, IEEE Transactions on Neural Systems and Rehabilitation Engineering 31 (2023) 1208–1218.
[19] X. Liu, S. Xiong, X. Wang, T. Liang, H. Wang, X. Liu, A compact multi-branch 1d convolutional neural network for eeg-based motor imagery classification, Biomedical Signal Processing and Control 81 (2023) 104456.
[20] P. Deny, S. Cheon, H. Son, K. W. Choi, Hierarchical transformer for motor imagery-based brain computer interface, IEEE Journal of Biomedical and Health Informatics (2023).

[21] H.-J. Ahn, D.-H. Lee, J.-H. Jeong, S.-W. Lee, Multiscale convolutional transformer for eeg classification of mental imagery in different modalities, IEEE Transactions on Neural Systems and Rehabilitation Engineering 31 (2022) 646–656.
[22] H. Altaheri, G. Muhammad, M. Alsulaiman, Physics-informed attention temporal convolutional network for eeg-based motor imagery classification, IEEE transactions on industrial informatics 19 (2) (2022) 2249–2258.
[23] Y. Song, Q. Zheng, B. Liu, X. Gao, Eeg conformer: Convolutional transformer for eeg decoding and visualization, IEEE Transactions on Neural Systems and Rehabilitation Engineering 31 (2022) 710–719.
[24] J. Zhang, K. Li, B. Yang, X. Han, Local and global convolutional transformer-based motor imagery eeg classification, Frontiers in Neuroscience 17 (2023) 1219988.
[25] Y. Li, L. Guo, Y. Liu, J. Liu, F. Meng, A temporal-spectral-based squeeze-and-excitation feature fusion network for motor imagery eeg decoding, IEEE Transactions on Neural Systems and Rehabilitation Engineering 29 (2021) 1534–1545.
[26] M. Wimpff, L. Gizzi, J. Zerfowski, B. Yang, Eeg motor imagery decoding: A framework for comparative analysis with channel attention mechanisms, Journal of Neural Engineering 21 (3) (2024) 036020.
[27] Y. Qin, B. Yang, S. Ke, P. Liu, F. Rong, X. Xia, M-fanet: Multi-feature attention convolutional neural network for motor imagery decoding, IEEE Transactions on Neural Systems and Rehabilitation Engineering (2024).
[28] K. Zhang, N. Robinson, S.-W. Lee, C. Guan, Adaptive transfer learning for eeg motor imagery classification with deep convolutional neural network, Neural Networks 136 (2021) 1–10.
[29] S. Pérez-Velasco, E. Santamaría-Vázquez, V. Martínez-Cagigal, D. Marcos-Martínez, R. Hornero, Eegsym: Overcoming inter-subject variability in motor imagery based bcis with deep learning, IEEE Transactions on Neural Systems and Rehabilitation Engineering 30 (2022) 1766–1775.
[30] H. W. Ng, C. Guan, Subject-independent meta-learning framework towards optimal training of eeg-based classifiers, Neural Networks 172 (2024) 106108.
[31] K. Yin, E. Y. Lim, S.-W. Lee, Gitgan: Generative inter-subject transfer for eeg motor imagery analysis, Pattern Recognition 146 (2024) 110015.
[32] D. Borra, S. Fantozzi, E. Magosso, Interpretable and lightweight convolutional neural network for eeg decoding: Application to movement execution and imagination, Neural Networks 129 (2020) 55–74.
[33] T. Hanakawa, I. Immisch, K. Toma, M. A. Dimyan, P. Van Gelderen, M. Hallett, Functional properties of brain areas associated with motor execution and imagery, Journal of neurophysiology 89 (2) (2003) 989–1002.
[34] R. Mane, N. Robinson, A. P. Vinod, S.-W. Lee, C. Guan, A multi-view cnn with novel variance layer for motor imagery brain computer interface, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2020, pp. 2950–2953.
[35] H.-M. Yang, X.-Y. Zhang, F. Yin, C.-L. Liu, Robust classification with convolutional prototype learning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3474–3482.
[36] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, Eca-net: Efficient channel attention for deep convolutional neural networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11534–11542.
[37] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[39] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, Advances in neural information processing systems 30 (2017).
[40] H.-M. Yang, X.-Y. Zhang, F. Yin, Q. Yang, C.-L. Liu, Convolutional prototype network for open set recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (5) (2020) 2358–2370.
[41] F. C. Borlino, S. Bucci, T. Tommasi, Contrastive learning for cross-domain open world recognition, in: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2022, pp. 10133–10140.
[42] Z. Xia, P. Wang, G. Dong, H. Liu, Adversarial kinetic prototype framework for open set recognition, IEEE Transactions on Neural Networks and Learning Systems (2023).
[43] B. Zhang, X. Li, Y. Ye, Z. Huang, L. Zhang, Prototype completion with primitive knowledge for few-shot learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 3754–3762.
[44] F. Zhou, P. Wang, L. Zhang, W. Wei, Y. Zhang, Revisiting prototypical network for cross domain few-shot learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 20061–20070.
[45] F. Wu, A. Fan, A. Baevski, Y. N. Dauphin, M. Auli, Pay less attention with lightweight and dynamic convolutions, arXiv preprint arXiv:1901.10430 (2019).
[46] H. Altaheri, G. Muhammad, M. Alsulaiman, S. U. Amin, G. A. Altuwaijri, W. Abdul, M. A. Bencherif, M. Faisal, Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: A review, Neural Computing and Applications 35 (20) (2023) 14681–14722.
[47] Z. Yang, L. Zhu, Y. Wu, Y. Yang, Gated channel transformation for visual recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11794–11803.
[48] W. Yu, C. Si, P. Zhou, M. Luo, Y. Zhou, J. Feng, S. Yan, X. Wang, Metaformer baselines for vision, IEEE Transactions on Pattern Analysis and Machine Intelligence (2023).
[49] F. Wang, X. Xiang, J. Cheng, A. L. Yuille, Normface: L2 hypersphere embedding for face verification, in: Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 1041–1049.
[50] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. R. Müller-Putz, et al., Review of the bci competition iv, Frontiers in neuroscience 6 (2012) 55.
[51] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, H. Zhang, Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b, Frontiers in neuroscience 6 (2012) 21002.
[52] B. Blankertz, K.-R. Muller, D. J. Krusienski, G. Schalk, J. R. Wolpaw, A. Schlogl, G. Pfurtscheller, J. R. Millan, M. Schroder, N. Birbaumer,

The bci competition iii: Validating alternative approaches to actual bci problems, IEEE transactions on neural systems and rehabilitation engineering 14 (2) (2006) 153–159.
[53] K. K. Ang, Z. Y. Chin, H. Zhang, C. Guan, Filter bank common spatial pattern (fbcsp) in brain-computer interface, in: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), IEEE, 2008, pp. 2390–2397.
[54] L. Van der Maaten, G. Hinton, Visualizing data using t-sne., Journal of machine learning research 9 (11) (2008).
[55] J. Luo, Y. Wang, S. Xia, N. Lu, X. Ren, Z. Shi, X. Hei, A shallow mirror transformer for subject-independent motor imagery bci, Computers in Biology and Medicine 164 (2023) 107254.

Scaling Law in Neural Data: Non-Invasive Speech Decoding with 175 Hours of EEG Data

Motoshige Sato[1], Kenichi Tomeoka[1], Ilya Horiguchi[1], Kai Arulkumaran[1], Ryota Kanai[1], and Shuntaro Sasai[1,*]

[1]Araya Inc. Tokyo, Japan,
[*]Correspondence: Shuntaro Sasai, Ph.D. (sasai_shuntaro@araya.org)

Abstract

Brain-computer interfaces (BCIs) hold great potential for aiding individuals with speech impairments. Utilizing electroencephalography (EEG) to decode speech is particularly promising due to its non-invasive nature. However, recordings are typically short, and the high variability in EEG data has led researchers to focus on classification tasks with a few dozen classes. To assess its practical applicability for speech neuroprostheses, we investigate the relationship between the size of EEG data and decoding accuracy in the open vocabulary setting. We collected extensive EEG data from a single participant (175 hours) and conducted zero-shot speech segment classification using self-supervised representation learning. The model trained on the entire dataset achieved a top-1 accuracy of 48% and a top-10 accuracy of 76%, while mitigating the effects of myopotential artifacts. Conversely, when the data was limited to the typical amount used in practice ($\sim$10 hours), the top-1 accuracy dropped to 2.5%, revealing a significant scaling effect. Additionally, as the amount of training data increased, the EEG latent representation progressively exhibited clearer temporal structures of spoken phrases. This indicates that the decoder can recognize speech segments in a data-driven manner without explicit measurements of word recognition. This research marks a significant step towards the practical realization of EEG-based speech BCIs.

1 Introduction

Motor speech disorder is a severe medical condition that leaves people barely or completely unable to speak, and occurs in 90% of Parkinson's disease patients [1], 45.2% of stroke patients [2], and 95% of amyotrophic lateral sclerosis (ALS) patients [3]. Typical communication aids for speech impairments, such as those using eye trackers, are significantly slower than spontaneous speech, especially in the late stages of ALS, where vision loss and eye movement deficits can cause fatigue problems and render use impossible [4–8]. Conversely, among recent advances in speech brain-computer interfaces (BCIs), invasive neural recordings such as intracortical microelectrode arrays and cortical electrograms (ECoG) have shown remarkable promise in achieving word production rates close to those of natural speech [9–14], while also providing a less burdensome alternative for users. However, these methods, which require surgery to implant electrodes in the brain, are highly invasive and present significant psychological and physical barriers. For this reason, there is a growing demand for a speech BCI using non-invasive neural recording technology, which has far lower barriers for use.

Among non-invasive recordings of neural activity, EEG enables portable and real-time BCI. fMRI and MEG require large magnetic field equipment and hence are impractical for everyday use. NIRS, on the other hand, is a neural recording method that is suitable for everyday use. However, because it measures blood flow changes as a secondary effect of neural activity, its temporal resolution is

Preprint. Under review.

insufficient for real-time decoding of continuous speech, where syllables occur at intervals of a few hundred milliseconds. EEG is a non-invasive and routinely available recording technique that can capture features of speech with a high enough temporal resolution at a relatively low cost.

On the other hand, EEGs measure signals that pass through the skull and skin, which weaken the signal amplitudes and are prone to noise and myopotential artifacts. Furthermore, since neural activity exhibits significant day-by-day variability [15], EEGs of the same subject are rarely recorded for long periods of time across days except for epilepsy monitoring or clinical purposes [16]. In fact, speech BCI systems utilizing EEG have collected data for each individual for only a few hours at most. Moreover, since the analysis has previously been limited to a maximum lexicon size of 13 words [17–22], it has not yet achieved a practical level that allows for highly flexible communication. To go beyond this and decode under open vocabulary conditions without limiting the number of words would therefore require a large amount of speech data containing a variety of words and phrases. Even then, it is not clear whether collecting such data by recording across many days would improve decoding performance because of the large day-by-day variability of EEG. In this work we investigated this problem, assessing decoding performance by recording the EEG of one subject during speech, over several months, for a total of 175 hours.

We employed the contrastive language-image pre-training (CLIP) [23] method, which uses self-supervised learning and hence does not require labels, on a large amount of paired EEG and speech data. Using CLIP we integrate the different modalities of EEG and speech and acquire representations that can be easily used for various downstream tasks, such as zero-shot classification of phrases, and speech reconstruction.

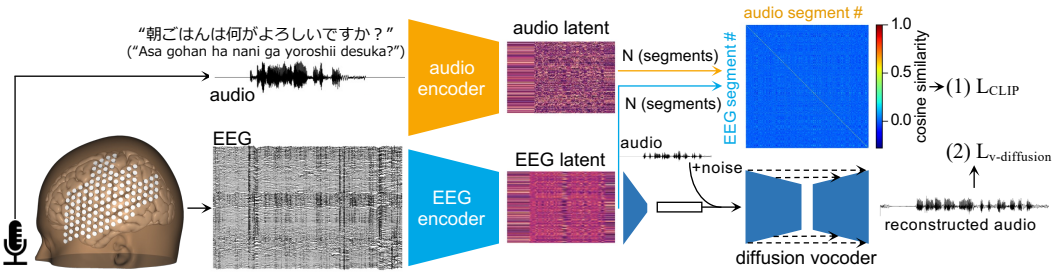


Figure 1: Decoding framework. EEG and speech were recorded simultaneously and converted to latent representations by a (fixed) audio encoder and an EEG encoder, respectively, for each 5-second segment. In step (1), the CLIP loss is applied on the N-segment pairwise cosine similarity matrix. In step (2), a diffusion vocoder was trained to reconstruct the speech waveform from the EEG latent representation.

2 Related Work

Here we review studies that have attempted to decode words from EEG in an open vocabulary setting.

Défossez et.al. [24] evaluated the zero-shot classification accuracy of EEG segments while participants listened to speech by optimizing the CLIP loss between the latent representations of EEG and audio. They used two datasets consisting of 19.2 h EEG from 19 participants, and 6.7 h EEG from 33 participants [25, 26]. They succeeded in achieving 25.7% in top-10 accuracy out of 190 segments for the former, and 17.7% out of 1,842 segments for the latter. The decoding performance was 5-30 times higher than chance levels. However, it is important to note that their experimental paradigm focused on decoding speech perception rather than directly targeting the decoding of the participant's own verbal output.

[27, 28] conducted decoding EEG during silent text comprehension, leveraging an EEG dataset known as ZuCo [29, 30], in which data was collected from 18 participants in total for 100-180 minutes per subject. Their approach involved the integration of eye-tracking data concurrent with EEG recordings to decode the process of reading. The EEG-to-text translation was conducted by aligning the latent representation of individual words with corresponding EEG signals captured during eye fixation periods, utilizing a machine-translation-based method [27] and a CLIP-based loss function with vector quantization [28]. However, subsequent to the publication of their findings,

the researchers acknowledged an overestimate in their accuracy due to the usage of teacher-forcing during evaluation. When teacher-forcing was omitted, they were unable to achieve performance significantly above random [31]. Liu et al. [32] successfully improved BLEU and ROUGE scores by up to 5% compare to DeWave[28] by implementing a multi-view transformer that combines a brain area independent transformer with an integrated transformer after pre-training with a masked auto-encoder and on the same ZuCo dataset.

3 Methods

3.1 Dataset

Table 1: Dataset properties. We summarized the properties of the two EEG datasets [25, 26] used by Défossez et.al. [24], and our dataset. Word overlap indicates the proportion of unique words in the test dataset that also appear in the training dataset. Japanese word segmentation on our dataset was performed using the janome tokenizer [33].

						vocabulary size		
dataset	language	task	channels	subjects	duration	train	test	word overlap
[26]	EN	listening	60	33	6.7 [h]	513	148	60%
[25]	EN	listening	128	19	19.2 [h]	1418	764	67%
ours	JP	speech	128	1	175 [h]	28159	10103	85%

3.1.1 Participant

A healthy male adult participated in this study. He had no history of neurological or psychiatric illnesses. He was initially given a briefing on the purpose of the study and experimental protocol, and he provided informed consent before we proceeded to the actual experiment. Our study obtained ethical approval by the Shiba Palace Clinic Ethics Review Committee. The experiments were undertaken in compliance with national legislation and the Code of Ethical Principles for Medical Research Involving Human participants of the World Medical Association (the Declaration of Helsinki).

3.1.2 EEG and EMG recording

EEG, electrooculogram (EOG), upper and lower orbicularis oris electromyogram (EMG), and speech voice signals were recorded simultaneously while the participant read aloud dialogues from text corpora [34], novels, and text-based games displayed on a computer screen. During the recording, pace instruction cues were not presented, and the subject continuously read the text aloud at a natural speed. The data recording experiment was conducted over a period of 48 days, with a total recording duration of 175 hours. EEG electrode placement was targeted around the language areas of the left hemisphere (Fig 1) and eight g.pangolin [35, 36] electrode sheets (16 electrodes per sheet, g.tec, Austria) were placed over Broca's area, auditory area, Wernicke's area, and premotor area. For EOG recording, the electrodes were placed above and below the left eye. For measurement of mouth muscle activity, each of two electrodes were placed on each of the left upper and lower orbicularis oris.

3.1.3 Preprocessing

EEG was acquired in real-time at 1200 Hz using the g.NEEDaccess Python API (g.tec, Austria). We denoised the raw EEG using MNE-Python [37]; details are shown in Figure 2a. To reduce the EMG contamination in the EEG signal, we used an adaptive filter [38, 39], through which the EEG signal linearly predicted from the EMG signals (EOG, and upper and lower orbicularis oris EMG) was removed from the EEG. The adaptive filter has been used in several studies [40–42] as a technique to remove EMG artifacts from EEG as it has a fast computation time and hence is suitable for real-time BCI. Details of the adaptive filter are described in Appendix B.1.

The participant's spoken voice was recorded with a microphone at 48 kHz and downsampled to 16 kHz. Furthermore, we applied Silero VAD [43] to remove noise from the audio signal. After

synchronizing the timestamps of the EEG and audio data, the signals were epoched into 5-second time windows without overlap, ensuring temporal correspondence between the EEG and audio modalities. For each window, the EEG data were Z-scored along the time axis and clipped to a range of ±5. Windows containing less than 20% of speech segments, as detected by Silero VAD, were excluded.

3.1.4 Data splits

EEG data, like other time series data, is known to exhibit temporal trends and variations. Therefore, in order to test the ability for our decoder to generalise over time, the dataset was split with chronological order, with the initial 80% assigned to train set, the following 10% to validation set, and the final 10% to test set.

3.2 Model

3.2.1 EEG preprocessing and encoder

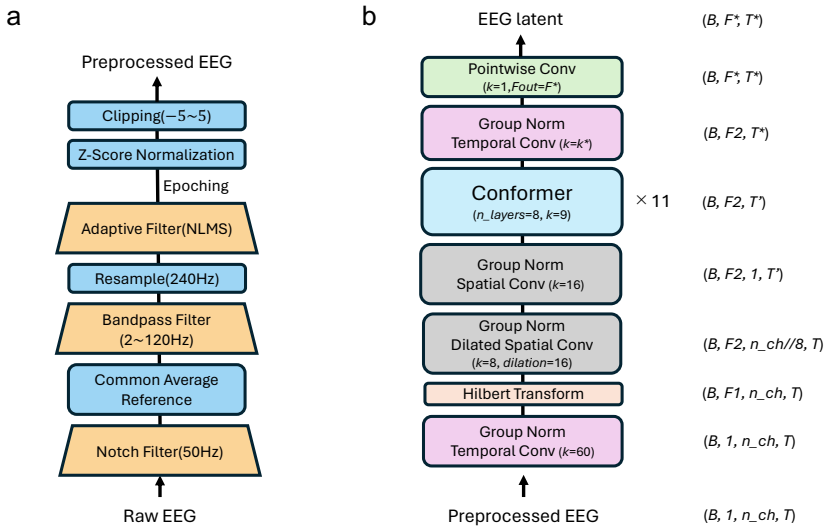


Figure 2: A diagram of EEG preprocessing and EEG encoder. (a) The pre-processing procedure. (b) EEG encoder architecture. The number of feature dimensions F^* and the number of time steps T^* in the latent representation differ depending on the audio encoder. The output shape of each layer is shown in the *right* column.

For the EEG encoder (Figure 2), we developed a model combining HTNet [44, 45] and Conformer [46]. For the audio encoder, we used the following three pre-trained models: wav2vec2.0 [47], Whisper [48] (only the encoder module was used), and Encodec [49].

For speech reconstruction, a 1D U-Net was trained to convert EEG latents to speech waveforms using diffusion vocoder architecture [50].

3.2.2 Decoder training

Encoders were applied to the simultaneously recorded speech and EEG signals, respectively, and the CLIP loss between EEG and audio latent representations was optimized. The loss function is expressed in Equation 1:

$$
\begin{aligned}
L_{\text{CLIP}} &= \text{CrossEntropy}(I, \hat{X}\hat{Y}) \\
&= -\frac{1}{N}\sum_{i}^{N} \log\left(\frac{\exp(\hat{X}_i \cdot \hat{Y}_i)}{\sum_j^N \exp(\hat{X}_i \cdot \hat{Y}_j)}\right)
\end{aligned} \tag{1}
$$

where, I denotes the identity matrix, N denotes the batch size, and x and y denote the audio and EEG data, respectively. $X = E_{\text{EEG}}(x)$ and, $Y = E_{\text{audio}}(y)$, where E_{EEG} and E_{audio} denote the

EEG encoder and audio encoder, respectively. The normalized representations are obtained as $\hat{X}_i = X_i/\|X_i\|$ and $\hat{Y}_i = Y_i/\|Y_i\|$.

For the audio encoder, a pre-trained wav2vec2.0, Whisper or Encodec model was applied, with weights kept fixed. When performing CLIP training of the EEG encoder with wav2vec2.0 embeddings, the randomly initialized EEG encoder was trained for 300 epochs with a batch size of 512 samples and the Lamb optimizer [51] (initial learning rate: 0.001, weight decay: 0.01). We used a cosine annealing learning rate schedule, with 1000 iterations of warmup (corresponding to 7.8 epochs). For CLIP training with either the Encodec or Whisper audio encoders, we instead used AdamW [52] (learning rate: 0.0001, weight decay: 0.01),since it achieved better performance than Lamb in a pilot study. The learning rate scheduler was the same as with wav2vec2.0.

For speech reconstruction, the diffusion vocoder was trained for 300 epochs with a batch size of 512 samples and the AdamW optimizer (learning rate: 0.001, weight decay: 0.01). The reconstructed speech waveform was generated efficiently by distillation model [53] of the vocoder with DDIM sampling [54]. 1000 time steps were used for the denoising process.

3.3 Evaluation

Generalization performance was evaluated by performing inference on the test dataset using the weights from the epoch with the lowest loss on the validation dataset.

To assess the effectiveness of pretraining using CLIP, zero-shot segment classification was conducted, following the same procedure as Défossez et.al. [24]. The cosine similarity matrix was computed between the EEG latents and audio latents of the 512 samples in the test set. For each EEG latent, the indices of the top-k most similar audio latents based on cosine similarity were extracted. The top-k accuracy was defined as the percentage of EEG latents for which the corresponding audio latent was among the top-k most similar latents. In other words, this score is higher when the latent predicted from the EEG is more similar to the corresponding audio latent than to the latents of other samples. The average performance ± 1 standard deviation (SD) was obtained from 16 batches consisting of 512 test samples per batch for Table 2, 3, 4 and Figure 3, 4.

For the evaluation of speech reconstruction, we computed the mel-cepstral distortion (MCD) metric, as used in previous studies [9, 14]. MCD is a widely accepted objective measure of the dissimilarity between two speech signals in the mel-cepstral domain [55], and is formulated by the following equation 2:

$$\mathrm{MCD}(\hat{y}, y) = \frac{10}{\log(10)} \sqrt{(\sum_{d=1}^{24} (\mathrm{mc}_d^y - \mathrm{mc}_d^{\hat{y}})^2)}, \tag{2}$$

where, $\hat{y}$ and y denotes the reconstructed speech and the recorded speech respectively, and mc_d denotes the dth dimension of the mel-cepstrum. A lower MCD value indicates a higher degree of similarity between the reconstructed and original speech signals. The average performance ± 1 SD was obtained from 8448 test samples for Figure 4.3.

4 Results

4.1 Zero-shot speech segment classification

Table 2: Zero-shot segment classification accuracy across different audio encoders.

Audio Encoder	top1	top10 (%)
wav2vec2.0	**48.5**	**76.0**
Whisper	10.6	52.9
Encodec	28.8	60.4

Table 2 presents the average zero-shot speech segment classification accuracy for 512 samples of decoders trained with all training data. The EEG encoder was trained to optimize the CLIP loss with

the respective audio latent of the pre-trained audio encoders: wav2vec2.0, Encodec and Whisper. Among the three types of speech encoders, wav2vec2.0 achieved the highest accuracy, with 48.5% for top-1 accuracy and 76.0% for top-10 accuracy. This performance is significantly higher than the chance level (top-1: 0.19%, top-10: 1.95%).

4.2 Data scaling in classification accuracy

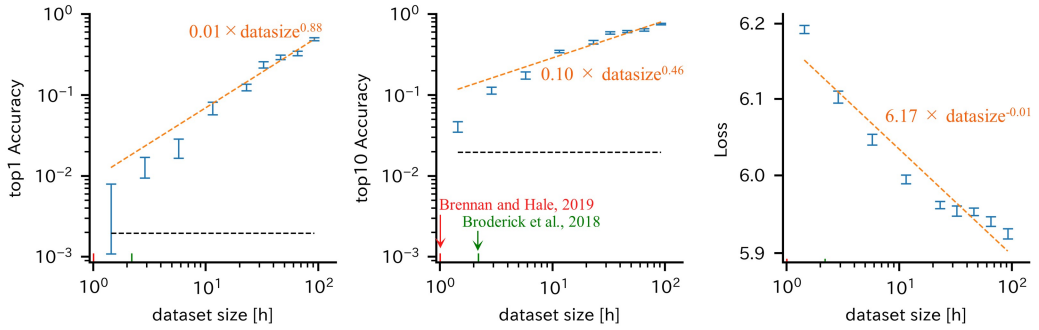


Figure 3: Data scaling. The relationship between the training dataset size (total segment length) and the top-1 accuracy (*left*), top-10 accuracy(*center*), and loss (*right*) on the test dataset. The *black* dashed line indicates chance level, and the *orange* dashed line indicates the best linear fit to the data. The dataset sizes of the *green* and *red* arrows are the dataset sizes of the datasets [25, 26] used in [24], respectively.

Figure 3 illustrates the relationship between zero-shot segment classification accuracy and the EEG encoder training loss versus the amount of training data. A common test set was applied to all models of each training data amount to evaluate performance. Figure 3(*left*, *center*) show that the classification accuracy improves as the amount of training data increases. Furthermore, the accuracy has not saturated with respect to the training data and continues to rise, suggesting that increasing the amount of data further would lead to further improvements in accuracy.

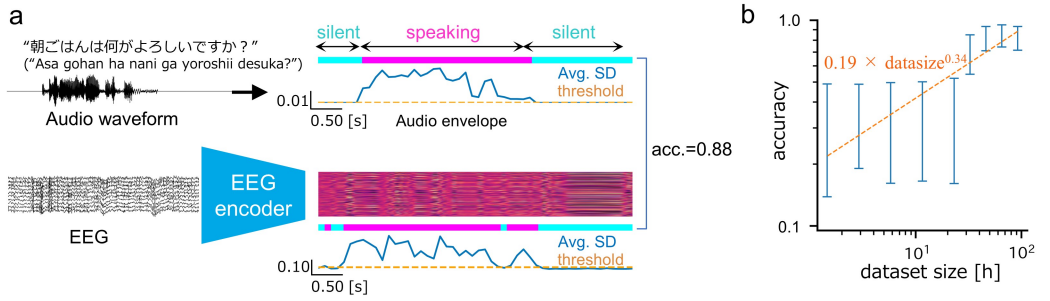


Figure 4: Voice activity detected from EEG latent representations without explicit training. (a) Process of speech interval detection. The speech waveform (*upper left*) and EEG (*lower left*) were each converted into a latent representation (*upper right color map*) through the encoder, and the variance for each feature dimension was taken in a sliding window of 100 ms and then averaged across feature domain (*lower blue line*). Intervals above the threshold (*orange line*) for this value were detected as speaking periods, and intervals below the threshold were detected as silent periods. The speech segment for the ground truth was determined by applying a threshold value to the waveform envelop. In this example, the overlap between the speech segments and the segments detected by the EEG latent (accuracy) was 0.88. (b) The relationship between the speech detection accuracy and the training dataset size.

Upon observing the EEG latent and audio latent representations, both exhibit similar temporal variations (Figure 1). The intervals without speech (periods between syllables) show nearly constant temporal variations across each feature dimension. The EEG latent and wav2vec2.0 latent representation were divided into 50 bins (=100ms) respectively, and the SD values were calculated along the

Table 3: Comparison of voice activity detection performance between different latents and raw EEG.

modality	predictor	accuracy	balanced accuracy (%)
audio	wav2vec2.0 latent	**99.9 ± 0.45**	**99.6 ± 3.3**
EEG	EEG latent	88.6 ± 10.3	78.7 ± 16.0
EEG	raw EEG	76.8 ± 22.4	60.6 ± 20.4

time axis within each bin. These SD values were averaged across the features and binarized using a threshold to detect the onset of each syllable. The ground truth for speech intervals was obtained by dividing the envelope of the speech waveform within each 5-second time window into 50 bins (=100 ms). The maximum value within each bin was taken as the representative value for that bin where the representative value exceeded a predetermined threshold were considered as speech, while those below the threshold were regarded as non-speech. Figure 4(a) illustrates the coincidence rate between the speech segments predicted from the EEG latent representation and those determined from the wav2vec2.0 latent representation. Figure 4(b) also confirms the data scaling of the speech segment agreement rate with respect to the learning data. The thresholds for determining speech intervals were set to 0.001 for the speech waveform, 0.047 for the speech latent representation, 0.01 for the raw EEG waveform, and 0.06 for the EEG latent representation. These values were obtained through a grid search to maximize the interval overlaps between the speech waveform and each respective data type.

4.3 Voice reconstruction from EEG latent representations

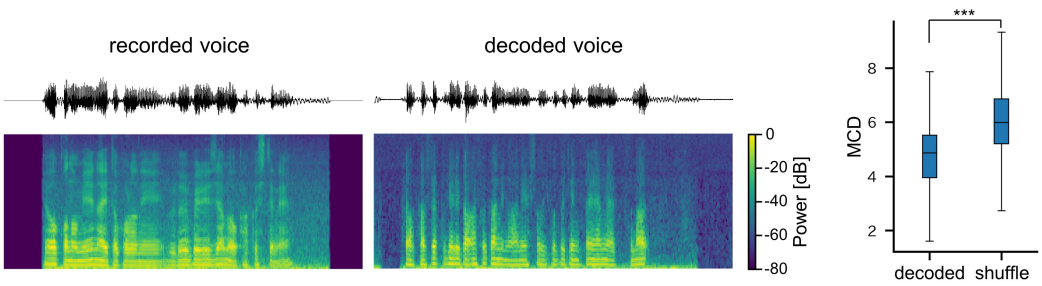


Figure 5: Representative recorded voice (*left*) and reconstructed voice (*middle*). The *top* panels show the voice waveforms and the *bottom* panels are mel-spectrograms. MCD scores (*right*), where smaller values indicate better performance. These were compared to a random model trained on a dataset with shuffled EEG and speech correspondences. The box plot illustrates the distribution of the scores obtained from 8,448 test samples. Decoding performance significantly outperformed chance ($p^{***} < 10^{-3}, t_{8447} = -73.5$, paired t-test).

The process of reconstruction of the participant's spoken voice from EEG is shown in Figure 5. After pre-training of EEG and audio latent representations with CLIP, a Vocoder diffusion model [53] was trained to convert EEG latent representations into audio waveforms, with fixed EEG encoder weights. The performance was evaluated by MCD, as with previous studies [9, 14]. The performance was significantly better than the MCD of models trained on the data where EEG and audio pairs were shuffled. We achieved an MCD of 4.68 dB on average, which is comparable to the results of SOTA (3.8-5.0 dB) of invasive speech decoding [9, 14]. It should be noted, however, that the MCD index is affected by differences in the sampling rate of the audio and silent segments, therefore a direct comparison is not feasible. While the reconstructed speech bears a resemblance to the participant's voice, it remains challenging to clearly discern the content of the speech. For the practical implementation of speech BCIs, it is still desirable to achieve a high level of clarity in the reconstructed speech. Consequently, improving the quality of the reconstructed speech remains a challenge for future research.

Table 4: Low susceptibility to EMG artifacts. The susceptibility to EMG artifacts of EEG encoder was evaluated by the top-10 decoding accuracy from EMG. This score was evaluated for EOG, upper orbicularis oris EMG (uEMG), lower orbicularis oris EMG (lEMG), and the mean of these three EMGs (mEMG) in each column. In order to train the model for inference independent of EMG, we applied data augmentation with synthetic data, mixing EMG scaled by a factor of α from different trials to EEG, which was multiplied by a factor of $1 - \alpha$. Here, α is a random value within range [0, 0.95]. The model was trained to minimize the CLIP loss with the audio latent for the EEG trial while ignoring the audio latent of the trial on the EMG side. Fig D provides a schematic overview of this operation. Each row indicates the EMG type mixed in this data synthesis.

	Predictor & top-10 accuracy(%)				
augment w/	EOG↓	uEMG↓	lEMG↓	mEMG↓	EEG↑
–	28.8	14.6	13.8	24.6	76.0
EOG	5.02	6.47	7.04	6.38	72.3
uEMG	2.88	2.95	3.05	2.69	70.1
lEMG	4.47	4.97	5.49	6.89	69.0
mEMG	3.19	4.43	4.53	4.09	70.9

4.4 Influence of EMG contamination on decoding performance

EEG is easily contaminated by myopotential artifacts[56, 57]. If these signals other than neural activities contribute to speech decoding, this technology will not work when used by patients with speech impairments, since similar levels of muscle activity will not be obtained from those patients. Therefore, to establish that speech can be reliably decoded from EEG independently of myopotential artifacts, we examined whether the decoder's inference performance was reduced when EMG was used as the input instead of EEG. To this end, we propose a method to train the decoder to infer audio latent representations while ignoring EMG by data augmentation where EMG signals from different trials were artificially added to EEG signals.

Table 4 shows that without data augmentation, the inference accuracy is significantly lower than the EEG decoding accuracy, but the speech is still decoded from EMG significantly above 2% of chance level. In other words, the accuracy of the EEG Encoder cannot be explained solely by the effect of EMG, but the influence of EMG artifacts cannot be completely denied.

The top-10 decoding performance from EMG signals dropped to around 3% for the models trained to disregard EMG signals using the mixed EMG and EEG signals This decline of performance to the near chance level of 2% suggests that the inference from EEG was not heavily influenced by EMG artifacts. Additionally, the decoding accuracy for EEG input remained at approximately 70% as measured by the top-10 accuracy, demonstrating the ability to decode speech without depending on muscle-related artifacts.

5 Limitations

Our current study achieved a top-1 accuracy of 48% on the 512-phrase classification task under open vocabulary conditions, which is unprecedented for EEG-based speech decoding. However, it is important to note that this does not imply that the system can be immediately used as a speech neuroprosthesis. Several challenges must be addressed in future work to fully develop a practical speech BCI.

Although we collected an exceptionally large sample of 175 hours of data, the data was obtained only from a single participant. As such, it is unclear whether this system can be transferred to other participants. However, we anticipate that it is possible to achieve a generally usable speech BCI without collecting such large-scale data from all participants. Typically, models pre-trained on large datasets can be fine-tuned with smaller amounts of new data. We hypothesize that this empirical observation will also hold for speech decoding. Future research should investigate the data amount required to enable transferability across participants.

Many highly performant speech neuroprostheses methods have been developed for invasive measurements from individuals with speech disabilities attempting speech. In contrast, our study demonstrates

that high-accuracy speech decoding is possible with non-invasive measurements by collecting data from healthy individuals during overt speech. This opens up a new avenue for highly accurate, non-invasive speech neuroprostheses for individuals with speech disabilities by using non-invasive measurements. However, since our study did not include data from individuals with speech disabilities, this hypothesis remains to be tested. Furthermore, it is essential to validate the potential of non-invasive speech decoding using the data collection methods for attempted speech that have been employed in invasive measurements. This validation is crucial for broader applications in patients with speech disabilities.

6 Discussion

Traditionally, the development of speech neuroprosthesis has relied on invasive measurements, such as ECoG [9, 10, 12, 14] and multi-unit recordings [11, 13], to capture neural activity during attempted speech. However, the physical and psychological burdens associated with invasive measurements have made these technologies impractical for many patients. On the other hand, non-invasive measurements like EEG have been considered infeasible due to their low signal quality, limiting their application to classification tasks with only a few dozen classes, far from the vocabulary sizes needed for everyday conversation. This has led to the belief that developing a BCI capable of supporting the vocabulary required for everyday conversation using EEG is unfeasible. Challenging this assumption, our study collected 175 hours of EEG data during speech from the same participant, achieving 48.5% top-1 and 76.0% top-10 accuracy in a 512-phrase classification task. Crucially, our results suggest that decoding accuracy can continue to increase significantly with further training data, based on a scaling law between performance and data size. Moreover, we demonstrated that identifying the onset of speech—a significant challenge for traditional language-decoding BCI—can be achieved without special training or reliance on non-neural signals such as eye tracking data [28–30], given sufficient data. These findings suggest that realizing an open vocabulary non-invasive speech neuroprosthesis using EEG is feasible by increasing data length.

Decoding speech from EEG data obtained during attempted or overt speech has seen little progress over years due to concerns about the contamination of muscle activities. This is because EEG data during speech contain substantial electromyographic (EMG) signals, which can overshadow the neural signals related to speech. Consequently, there has been an inevitable concern that models trained on such data might decode speech relying on EMG signals rather than neural activity. Our study addressed this concern by training the model to actively ignore EMG-related signals by artificially mixing EMG signals from other trials to the training dataset. In support of this approach, the models trained on the EMG-mixed data failed to predict the speech only from the EMG signals as shown by the near-chance decoding accuracy from EMG signals alone (Table 4). These findings indicate that our model primarily relied on EEG signals. Our study was conducted with participants in a static seated position, minimizing the influence of non-speech-related EMG signals. However, for speech neuroprosthesis technology to be practical, it needs to function while the user is moving. Future research should collect data from participants speaking while moving and train models to handle such dynamic conditions, aiming to develop more versatile and practical models.

The state-of-the-art in EEG-based language decoding prior to our study, achieved by Défossez et al. [24], involved decoding heard phrases from over 100 words with a top-10 accuracy of 25.7% out of 190 segments for dataset [26] and 17.7% out of 1842 segments for dataset [25]. In contrast, we achieved a top-10 accuracy of 76.0% for 512 segments when decoding phrases during speech. When the length of training data was reduced to 2.9 hours, comparable to what Défossez et al. used, our accuracy dropped to 10.5% out of 190 segments and 1.40% out of 1842 segments, lower than their reported accuracy. Additionally, the overlap rate of words between training and test data also dropped to a similar level as reported by Défossez et al[24]. (Appendix F). These results suggest that the task difficulty of speech decoding may be harder than that of listening. We further discovered that speech decoding accuracy increases logarithmically with the word overlap rate (Appendix E). Given the limited number of commonly used vocabulary words, the required data length to learn language representation in the brain might be on the order of 10^2 hours for both speech and listening. Future research should conduct similar analyses on reading data to demonstrate that the scaling law between data length and decoding accuracy universally applies to language decoding from the brain, regardless of whether it involves reading or speech.

Acknowledgments and Disclosure of Funding

We thank Masakazu Inoue and Sensho Nobe, for helpful discussions. This work was supported by the JST, Moonshot R&D Grant Number JPMJMS2012. The authors declare no competing interests.

References

[1] Moya-Galé, G., E. S. Levy. Parkinson's disease-associated dysarthria: prevalence, impact and management strategies. *Research and Reviews in Parkinsonism*, pages 9–16, 2019.

[2] Bahia, M. M., L. F. Mourao, R. Y. S. Chun. Dysarthria as a predictor of dysphagia following stroke. *NeuroRehabilitation*, 38(2):155–162, 2016.

[3] Ball, L. J., D. R. Beukelman, G. L. Pattee. Communication effectiveness of individuals with amyotrophic lateral sclerosis. *Journal of Communication Disorders*, 37(3):197–215, 2004.

[4] Hayashi, H., E. A. Oppenheimer. Als patients on tppv: totally locked-in state, neurologic findings and ethical implications. *Neurology*, 61(1):135–137, 2003.

[5] San Agustin, J., H. Skovsgaard, E. Mollenbach, et al. Evaluation of a low-cost open-source gaze tracker. In *Proceedings of the 2010 symposium on eye-tracking research & applications*, pages 77–80. 2010.

[6] Ball, L. J., A. S. Nordness, S. K. Fager, et al. Eye gaze access of aac technology for people with amyotrophic lateral sclerosis. *Journal of medical speech-language pathology*, 18(3):11, 2010.

[7] Spataro, R., M. Ciriacono, C. Manno, et al. The eye-tracking computer device for communication in amyotrophic lateral sclerosis. *Acta Neurologica Scandinavica*, 130(1):40–45, 2014.

[8] Chen, S.-H. K., M. O'Leary. Eye gaze 101: what speech-language pathologists should know about selecting eye gaze augmentative and alternative communication systems. *Perspectives of the ASHA Special Interest Groups*, 3(12):24–32, 2018.

[9] Anumanchipalli, G. K., J. Chartier, E. F. Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019.

[10] Makin, J. G., D. A. Moses, E. F. Chang. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature neuroscience*, 23(4):575–582, 2020.

[11] Willett, F. R., D. T. Avansino, L. R. Hochberg, et al. High-performance brain-to-text communication via handwriting. *Nature*, 593(7858):249–254, 2021.

[12] Moses, D. A., S. L. Metzger, J. R. Liu, et al. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3):217–227, 2021.

[13] Willett, F. R., E. M. Kunz, C. Fan, et al. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, 2023.

[14] Metzger, S. L., K. T. Littlejohn, A. B. Silva, et al. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620(7976):1037–1046, 2023.

[15] Huang, G., Z. Zhao, S. Zhang, et al. Discrepancy between inter-and intra-subject variability in eeg-based motor imagery brain-computer interface: Evidence from multiple perspectives. *Frontiers in neuroscience*, 17:1122661, 2023.

[16] Renzel, R., L. Tschaler, I. Mothersill, et al. Sensitivity of long-term eeg monitoring as a second diagnostic step in the initial diagnosis of epilepsy. *Epileptic Disorders*, 23(4):572–578, 2021.

[17] Lee, S.-H., M. Lee, S.-W. Lee. Eeg representations of spatial and temporal features in imagined speech and overt speech. In *Pattern Recognition: 5th Asian Conference, ACPR 2019, Auckland, New Zealand, November 26–29, 2019, Revised Selected Papers, Part II 5*, pages 387–400. Springer, 2020.

[18] Lee, Y.-E., S.-H. Lee, S.-H. Kim, et al. Towards voice reconstruction from eeg during imagined speech. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pages 6030–6038. 2023.

[19] Koizumi, K., K. Ueda, M. Nakao. Development of a cognitive brain-machine interface based on a visual imagery method. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1062–1065. IEEE, 2018.

[20] Watanabe, H., H. Tanaka, S. Sakti, et al. Synchronization between overt speech envelope and eeg oscillations during imagined speech. *Neuroscience research*, 153:48–55, 2020.

[21] Sereshkeh, A. R., R. Trott, A. Bricout, et al. Eeg classification of covert speech using regularized neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2292–2300, 2017.

[22] —. Online eeg classification of covert speech for brain–computer interfacing. *International journal of neural systems*, 27(08):1750033, 2017.

[23] Radford, A., J. W. Kim, C. Hallacy, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[24] Défossez, A., C. Caucheteux, J. Rapin, et al. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5(10):1097–1107, 2023.

[25] Broderick, M. P., A. J. Anderson, G. M. Di Liberto, et al. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Curr. Biol.*, 28(5):803–809.e3, 2018.

[26] Brennan, J. R., J. T. Hale. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLoS One*, 14(1):e0207741, 2019.

[27] Wang, Z., H. Ji. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pages 5350–5358. 2022.

[28] Duan, Y., C. Zhou, Z. Wang, et al. Dewave: Discrete encoding of eeg waves for eeg to text translation. In *Thirty-seventh Conference on Neural Information Processing Systems*. 2023.

[29] Hollenstein, N., J. Rotsztejn, M. Troendle, et al. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):1–13, 2018.

[30] Hollenstein, N., M. Troendle, C. Zhang, et al. Zuco 2.0: A dataset of physiological recordings during natural reading and annotation. *arXiv preprint arXiv:1912.00903*, 2019.

[31] Hyejeong Jo, J. H. Y. D. H. X. W. H. L., Yiqian Yang. Are eeg-to-text models working? In *Arxiv*. 2024.

[32] Liu, H., D. Hajialigol, B. Antony, et al. Eeg2text: Open vocabulary eeg-to-text decoding with eeg pre-training and multi-view transformer. *arXiv preprint arXiv:2405.02165*, 2024.

[33] Welcome to janome's documentation! (english) — janome v0.4 documentation (en). `https://mocobeta.github.io/janome/en/`. Accessed: 2024-5-15.

[34] Sonobe, R., S. Takamichi, H. Saruwatari. Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis. *arXiv preprint arXiv:1711.00354*, 2017.

[35] Lee, H. S., L. Schreiner, S.-H. Jo, et al. Individual finger movement decoding using a novel ultra-high-density electroencephalography-based brain-computer interface system. *Frontiers in Neuroscience*, 16:1009878, 2022.

[36] Schreiner, L., M. Jordan, S. Sieghartsleitner, et al. Mapping of the central sulcus using non-invasive ultra-high-density brain recordings. *Scientific reports*, 14(1):6527, 2024.

[37] Gramfort, A., M. Luessi, E. Larson, et al. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7(267):1–13, 2013.

[38] Diniz, P. S., et al. *Adaptive filtering*, vol. 4. Springer, 1997.

[39] Haykin, S. S. *Adaptive filter theory*. Pearson Education India, 2002.

[40] Correa, A. G., E. Laciar, H. Patiño, et al. Artifact removal from eeg signals using adaptive filters in cascade. In *Journal of Physics: Conference Series*, vol. 90, page 012081. IOP Publishing, 2007.

[41] Kher, R., R. Gandhi. Adaptive filtering based artifact removal from electroencephalogram (eeg) signals. In *2016 International Conference on Communication and Signal Processing (ICCSP)*, pages 0561–0564. IEEE, 2016.

[42] Molla, M. K. I., M. R. Islam, T. Tanaka, et al. Artifact suppression from eeg signals using data adaptive time domain filtering. *Neurocomputing*, 97:297–308, 2012.

[43] Team, S. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. `https://github.com/snakers4/silero-vad`, 2021.

[44] Lawhern, V. J., A. J. Solon, N. R. Waytowich, et al. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

[45] Peterson, S. M., Z. Steine-Hanson, N. Davis, et al. Generalized neural decoders for transfer learning across participants and recording modalities. *Journal of Neural Engineering*, 18(2):026014, 2021.

[46] Gulati, A., J. Qin, C.-C. Chiu, et al. Conformer: Convolution-augmented transformer for speech recognition. 2020.

[47] Baevski, A., H. Zhou, A. Mohamed, et al. wav2vec 2.0: A framework for Self-Supervised learning of speech representations. 2020.

[48] Radford, A., J. W. Kim, T. Xu, et al. Robust speech recognition via large-scale weak supervision. 2022.

[49] Défossez, A., J. Copet, G. Synnaeve, et al. High fidelity neural audio compression. 2022.

[50] Schneider, F., O. Kamal, Z. Jin, et al. Moûsai: Text-to-music generation with long-context latent diffusion, 2023.

[51] You, Y., J. Li, S. Reddi, et al. Large batch optimization for deep learning: Training BERT in 76 minutes. 2019.

[52] Loshchilov, I., F. Hutter. Decoupled weight decay regularization. 2017.

[53] Salimans, T., J. Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*. 2022.

[54] Song, J., C. Meng, S. Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*. 2021.

[55] Kubichek, R. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, vol. 1, pages 125–128. IEEE, 1993.

[56] Porcaro, C., M. T. Medaglia, A. Krott. Removing speech artifacts from electroencephalographic recordings during overt picture naming. *NeuroImage*, 105:171–180, 2015.

[57] Jiang, X., G.-B. Bian, Z. Tian. Removal of artifacts from eeg signals: a review. *Sensors*, 19(5):987, 2019.

[58] Cejnek, M., J. Vrba. Padasip: An open-source python toolbox for adaptive filtering. *Journal of Computational Science*, 65:101887, 2022.

A EEG and EMG recording details

Before electrode placement, the hair was shaved with a shaver and then cleaned with alcohol tissue. The size of the head was measured with a measuring tape to determine the location of the CZ, and eight electrode sheets coated with conductive gel (Elefix V, ZV-181E, NIHON KOHDEN, Japan) were attached at the marked positions. For the ground, the mastoid process behind the left ear was polished with Nuprep (Weaver and Company, US) and cleaned with alcohol tissue, and EMG/ECG/EKG electrode (Kendall^{TM}, CardinalHealth, US) was placed.

B Preprocessing details

B.1 Formulation of adaptive filter

In this study, we employed an adaptive filter implemented with Normalized Least-Mean-Square (NLMS) to remove EMG artifacts from the EEG data. The NLMS filter updates its weight matrix $w(t) \in \mathbb{R}^{\mathrm{ch\_emg} \times \mathrm{ch\_eeg}}$ according to equation 3, where the EEG data is represented by a vector $s(t) \in \mathbb{R}^{\mathrm{ch\_eeg}}$ and the EMG data is represented by a vector $n(t) \in \mathbb{R}^{\mathrm{ch\_emg}}$. In equation 3 the adaptation coefficient η was set to 0.1, and the parameter ϵ, which prevents divergence caused by division, was set to the default value of 0.001. The implementation was based on padasip [58] library.

$$
\begin{aligned}
w(t+1) &= w(t) + \mu \frac{n(t) \cdot (s(t) - \hat{s}(t))^{\top}}{\|n(t)\|^2 + \epsilon} \\
\hat{s}(t) &= w(t)^{\top} \cdot n(t)
\end{aligned}
\tag{3}
$$

C Computational resources

All models were trained in parallel with Distributed Data Parallel (DDP) using four NVIDIA A100 GPUs (4 $\times$ 80 GB). It took approximately 40 hours to train one EEG Encoder or a diffusion vocoder with the full training dataset.

D EMG analysis

Details of the sensitivity test for EMG, as described in Section 4.4. In training phase, the EEG encoder receives input X, defined as in Equation4.

$$
X = (1 - \alpha) \times EEG + \alpha \times EMG \tag{4}
$$

Here, the EMG is from a different segment than the EEG, and α is a value randomly sampled from range [0, 0.95]. In inference phase, either EEG or EMG is fed into the EEG encoder, and the classification accuracy is compared (Table 4) between the cases where EEG is used as input and where EMG is used as input. If the inference accuracy is higher when EMG is used compared to when EEG is used, it indicates that the learning of EEG encoder largely rely on EMG. However, if the inference accuracy is higher when EEG is used, it suggests that the learning of the EEG encoder prioritized EEG and tends to ignore EMG.

E Word overlap scaling

We investigated the impact of word overlap between the training data and the test data on decoding accuracy. Figure 7 shows that the accuracy for both top-1 and top-10 increases logarithmically as the overlap increases. Correspondingly, the loss also decreases logarithmically.

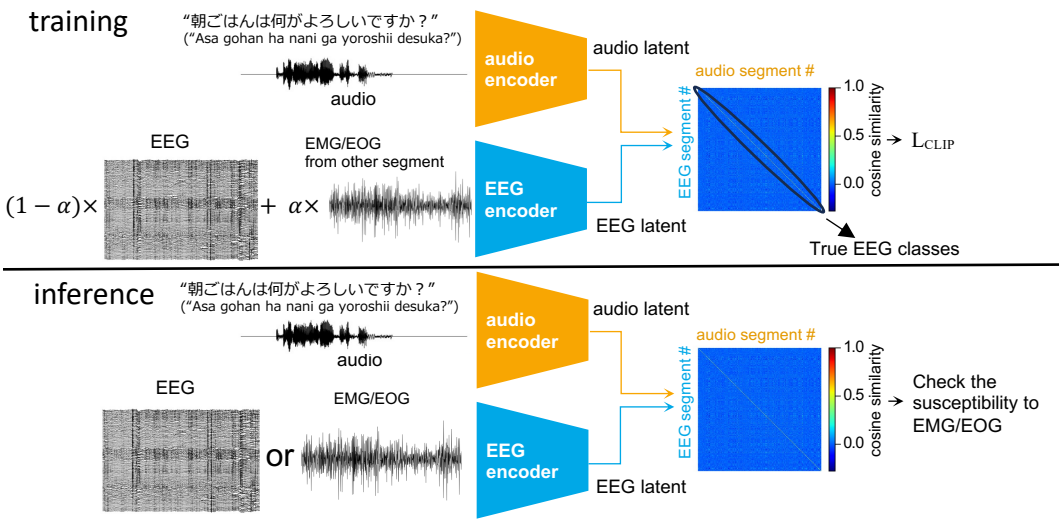


Figure 6: A schematic diagram illustrating the test procedure for evaluating the influence of EMG on the decoding process.

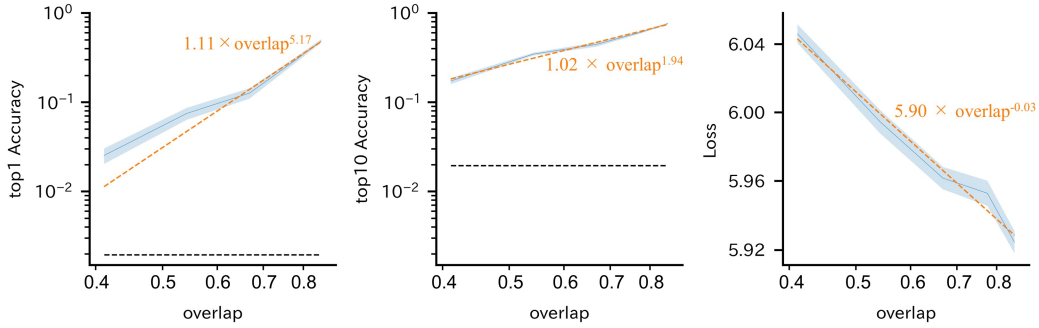


Figure 7: Word overlap increases logarithmically with recording duration.

F Performance comparison under the same test conditions as the previous study

Table 5: Zero-shot segment classification performance at the same amount of training data and number of segments as previous study [24].

dataset	dataset size [h]	test segments	top10 (%)
ours (1/32)	2.89	1842	3.64
[26]	1.01	1842	17.7
ours (1/32)	2.89	190	23.9
[25]	2.20	190	25.7

We compared the decoding accuracy using the same amount of data as the previous study [24]. Table 5 shows the comparison for two datasets [25, 26]. While their study focuses on decoding from hearing task, our task focuses on speech decoding.

G Data & code availability

The preprocessed EEG & EMG, audio latent, audio waveform, and transcription used in the analysis are split into train, validation and test sets, and have been uploaded to the following data repository. https://dataverse.harvard.edu/privateurl.xhtml?token=efad3304-6bcc-4cac-a542-27078dbd09e7

All codes used for the analysis, including EEG and audio pre-processing, training of EEG encoder and diffusion vocoder, and inference, evaluation, and visualization, will be made available in a public repository.

QEEGNet: Quantum Machine Learning for Enhanced Electroencephalography Encoding

Chi-Sheng Chen*†, Samuel Yen-Chi Chen‡, Aidan Hung-Wen Tsai†, Chun-Shu Wei*
*Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
m50816m50816@gmail.com, wei@nycu.edu.tw
‡Computational Science Initiative, Brookhaven National Laboratory, Upton NY, USA
ycchen1989@ieee.org
†Neuro Industry, Inc., CA, USA
{michael, aidan}@neuro-industry.com

***Abstract*—Electroencephalography (EEG) is a critical tool in neuroscience and clinical practice for monitoring and analyzing brain activity. Traditional neural network models, such as EEGNet, have achieved considerable success in decoding EEG signals but often struggle with the complexity and high dimensionality of the data. Recent advances in quantum computing present new opportunities to enhance machine learning models through quantum machine learning (QML) techniques. In this paper, we introduce Quantum-EEGNet (QEEGNet), a novel hybrid neural network that integrates quantum computing with the classical EEGNet architecture to improve EEG encoding and analysis, as a forward-looking approach, acknowledging that the results might not always surpass traditional methods but it shows its potential. QEEGNet incorporates quantum layers within the neural network, allowing it to capture more intricate patterns in EEG data and potentially offering computational advantages. We evaluate QEEGNet on a benchmark EEG dataset, BCI Competition IV 2a, demonstrating that it consistently outperforms traditional EEGNet on most of the subjects and other robustness to noise. Our results highlight the significant potential of quantum-enhanced neural networks in EEG analysis, suggesting new directions for both research and practical applications in the field.**

***Index Terms*—electroencephalography, EEG classification, quantum machine learning, quantum algorithm, deep learning, brain-computer interface**

I. Introduction

Electroencephalography (EEG) is a non-invasive technique widely used in neuroscience and clinical applications to measure electrical activity in the brain. The analysis of EEG data has been instrumental in understanding brain functions, diagnosing neurological disorders, and developing brain-computer interfaces. Traditional methods for EEG analysis often rely on conventional machine learning and deep learning techniques, such as the EEGNet model, which have demonstrated significant success in various EEG-based tasks. However, these models sometimes face limitations in capturing the complex and high-dimensional nature of EEG signals [1].

Recent advancements in quantum computing have opened a new era for enhancing machine learning algorithms. Quantum machine learning (QML) leverages the principles of quantum mechanics to process information in fundamentally different ways compared to classical computing, offering potential advantages in terms of computational efficiency and the ability to explore larger solution spaces [2]. The integration of QML with classical neural networks presents a promising hybrid approach that can potentially overcome some of the limitations of traditional deep learning models [3], [4].

In this paper, we propose QEEGNet, a novel hybrid neural network that combines quantum machine learning techniques with the EEGNet architecture [5] to enhance the encoding and analysis of EEG data. By incorporating quantum layers into the neural network, QEEGNet aims to leverage the power of quantum computing to improve the performance and robustness of EEG-based models. Our approach builds upon the strengths of EEGNet while introducing quantum elements that can capture more intricate patterns within EEG signals.

We evaluate QEEGNet on a famous benchmark EEG dataset, comparing its performance with the traditional EEGNet model. Our experimental results demonstrate that QEEGNet consistently outperforms these models in terms of accuracy and robustness to noise. These findings suggest that the integration of quantum machine learning can significantly enhance the capabilities of EEG analysis, paving the way for more effective and reliable applications in neuroscience and clinical settings.

The contributions of this paper are as follows:

- We introduce Quantum-EEGNet (QEEGNet), a hybrid neural network that integrates variational quantum circuits (VQC) with EEGNet for enhanced EEG encoding.
- We provide a comprehensive evaluation of QEEGNet on a popular EEG open dataset, demonstrating its partial superior performance compared to traditional models.
- We discuss the practical feature embedding ability of all the models, prove that QEEGNet has the more advantage of feature representation than EEGNet in the field of EEG analysis.

II. Related Work

A. Deep Learning on EEG Data

The analysis of EEG data using neural networks has been a focal point in recent research [6], aiming to improve the accuracy and efficiency of EEG-based applications. Traditional

models like artificial neural network, deep neural network models have shown significant promise but face limitations in handling the complex, high-dimensional nature of EEG signals. Recent advancements in quantum computing offer novel approaches to address these challenges.

EEGNet, tailored for classifying EEG Event-Related Potential (ERP) tasks, highlighting the potential of advanced neural networks in enhancing EEG signal processing. Their approach demonstrated improved classification accuracy, yet it also underscored the limitations of classical models in fully capturing the intricacies of EEG data.

There are several studies using deep learning dealing with different downstream tasks on EEG data. [7] using diffusion model to generate the EEG synthetic data, the proposed method could have a broader impact on neuroscience research by creating large, publicly available synthetic EEG datasets without privacy concerns. Artifact removal [8] is also a important field in EEG signal processing, [9] based on a pre-trained subject-independent model, was validated through multiple evaluations, showing it can maintain brain activity and outperform current artifact removal methods in decoding accuracy. In another study [10], explored the integration of both graph attention and self-attention mechanisms with convolutional neural network (CNN) as EEG encoder for EEG-based image recognition. Their work showed that incorporating attention mechanisms can significantly enhance the model's ability to focus on relevant features of multimodal data, thereby improving performance.

Despite these advancements, traditional deep learning models often fall short in fully addressing the high-dimensional and non-linear characteristics of EEG signals. As reviewed by [11], the incorporation of sophisticated deep learning components has shown potential benefits, yet there remains a need for more robust approaches to unlock the full potential of EEG data analysis. Recent advancements in quantum computing offer novel approaches to these challenges. Quantum machine learning (QML) models, with their ability to handle high-dimensional data and complex dependencies, present a promising direction for advancing EEG data analysis. Notably, [12] demonstrates the potential of QML approaches in time-series signal processing across several models, further underscoring the capabilities of QML in addressing the limitations of classical neural networks. By leveraging the principles of quantum mechanics, QML models can potentially overcome the limitations of classical neural networks, paving the way for more accurate and efficient EEG-based applications.

B. Quantum Machine Learning on EEG Data

Quantum machine learning (QML) represents a cutting-edge advancement that can further elevate EEG analysis. QML algorithms, especially the hybrid quantum-classical algorithms based on VQC, draw inspiration from several traditional deep learning algorithms, such as quantum convolutional neural networks (QCNN) [13], quantum generative adversarial networks (QGAN) [14], and quantum long short-term memory networks (QLSTM) [15], among others. A recent paper [16] discussed the application of hybrid quantum-classical neural networks for pattern recognition tasks. In this setting, a classically pre-trained model is used to preprocess the data, which is then sent to a variational quantum circuit (VQC) to further process. Their findings suggest that incorporating quantum layers can enhance classical models by providing greater computational power and the ability to explore larger solution spaces. Applying this hybrid quantum-classical approach to EEG analysis can potentially address the intricate patterns and dependencies in EEG data, potentially overcoming the limitations faced by traditional neural networks.

Moreover, QML has shown promise in improving model classification performance. [17] provided a QCNN to do the medical image classification. In the EEG QML application, [18] proposed a hybrid quantum-classical neural networks with a VQC in front of the traditional neural network, [19] and [20] exploring EEG classification task using quantum support vector machine (QSVM), [21] using different quantum circuit architectures as feature extractor with the classic multilayer perceptron (MLP) to do drowsiness detection. These researches emphasizing the potential to revolutionize various machine learning tasks, including those involving complex, high-dimensional data such as EEG. Their findings suggest that quantum algorithms can significantly enhance the representational power and efficiency of traditional neural networks, paving the way for more robust EEG analysis model. These studies collectively highlight the evolving landscape of EEG analysis, where integrating advanced neural network architectures and quantum computing techniques can lead to substantial performance gains. Building upon these foundations, our work differentiates from these aforementioned studies by specifically integrating VQC within the EEGNet architecture to create QEEGNet. While previous research has shown the promise of QML in various domains, including time series analysis and medical imaging, our approach focuses on leveraging quantum layers to enhance EEG encoding and analysis. By doing so, we aim to address the unique challenges posed by EEG data, such as its high dimensionality and complex temporal-spatial dependencies. The novelty of QEEGNet lies in being the first quantum-classical hybrid model that incorporates VQC quantum encoding layers at the end of the model. This unique architecture enables QEEGNet to capture more intricate patterns within EEG signals, improving performance and robustness in EEG-based applications. Our experimental results demonstrate that QEEGNet consistently outperforms traditional EEGNet models, highlighting the significant potential of quantum-enhanced neural networks in the field of EEG analysis.

III. Methodology

In this section, we detail the architecture and methodology of QEEGNet, our proposed hybrid quantum-classical neural network model that leverages quantum machine learning to enhance EEG encoding and analysis.

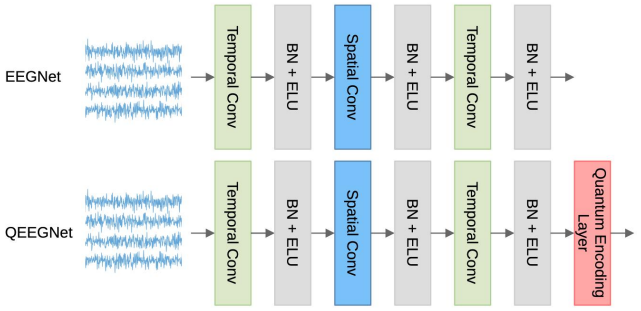


Fig. 1. Schematic illustration comparison of the EEGNet and the proposed Quantum-EEGNet (QEEGNet). In this diagram, Conv denotes convolution, BN denotes batch normalization, and ELU denotes the exponential linear unit activation function.

A. Architecture of EEGNet

In the field of brain-computer interface (BCI) and EEG signal processing, EEGNet [5] is a popular method, it a compact CNN model specifically designed for the analysis and classification of EEG data. The network consists of only a few layers, each designed to capture different aspects of the EEG signal, including temporal and spatial features. The EEGNet architecture illustrated in Fig 1.

B. Quantum Circuit

Quantum circuits can potentially provide computational advantages over classical methods for specific tasks. By exploiting the superposition and entanglement of qubits, quantum circuit can represent and process information in ways that classical bits cannot. This capability is mathematically represented as:

$$\text{Quantum State: } |\psi\rangle = \alpha|0\rangle + \beta|1\rangle, \tag{1}$$

where α and β are complex amplitudes, and $|\psi\rangle$ can exist in multiple states simultaneously. Quantum entanglement and superposition enable the exploration of complex data structures and correlations. For instance, a quantum state with n qubits can represent 2^n states simultaneously, providing a richer feature space for learning tasks:

$$|\psi\rangle = \sum_{i=0}^{2^n-1} \alpha_i |i\rangle, \tag{2}$$

where $|i\rangle$ represents the basis states of the quantum system, and α_i are the coefficients that determine the probability amplitudes of these states in the superposition. The index i helps to enumerate all possible states of the n-qubit system, which can be in any combination of the 2^n possible states.

This superposition allows the quantum encoding layer to encode and process more information than a classical layer of comparable size.

C. Quantum Encoding Layer Module

The quantum encoding layer module shows in in Fig 2. It is to integrate quantum computing capabilities into classical neural networks. By incorporating a quantum circuit within a classical deep learning model, the quantum encoding layer aims to leverage the unique properties of quantum mechanics, such as superposition and entanglement, to enhance the performance and capabilities of machine learning models. It leverages a parameterized quantum circuit defined by mapping classical EEG features as a set of qubits $\mathbf{x} \in \mathbb{R}^{n_{\text{qubits}}}$ and has trainable weights $\mathbf{w} \in \mathbb{R}^{n_{\text{layers}} \times n_{\text{qubits}}}$. The first step in the quantum encoding layer is encoding the input. Each qubit q_i is rotated based on the input data x_i using the rotation gate RY (rotation-Y gate):

$$RY(x_i) = \exp\left(-i\frac{x_i}{2}\sigma_y\right), \tag{3}$$

where σ_y is the Pauli-Y matrix:

$$\sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}. \tag{4}$$

Second, there are some parameterized layers in the quantum encoding layer module, in this QEEGNet model, we use a ring pattern of CNOT (controlled-NOT) gates. The CNOT gate is a two-qubit gate where the state of one qubit (the control qubit) determines whether to flip the state of another qubit (the target qubit). The CNOT gate is a two-qubit quantum gate represented by the following 4x4 unitary matrix:

$$CNOT = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{5}$$

A ring pattern involves applying CNOT gates such that each qubit is entangled with its neighbor, and the last qubit is entangled with the first qubit, forming a closed loop or "ring." This pattern ensures that entanglement is spread throughout the entire set of qubits. For each layer l in n_{layers}, a ring pattern of CNOT gates entangles the qubits:

$$CNOT(q_i, q_{(i+1) \mod n_{\text{qubits}}}) \quad \text{for} \quad i = 0, 1, \ldots, n_{\text{qubits}} - 1. \tag{6}$$

Each qubit q_i undergoes an additional rotation based on the trainable weight $w_{l,i}$:

$$RY(w_{l,i}) = \exp\left(-i\frac{w_{l,i}}{2}\sigma_y\right). \tag{7}$$

The final stage of the quantum encoding layer module is measurement the states of the qubits. The expectation value of the Pauli-Z operator σ_z is measured for each qubit, yielding the output:

$$\langle \sigma_z^i \rangle = \langle 0|U^\dagger \sigma_z^i U|0\rangle, \tag{8}$$

where U is the unitary operation representing the quantum circuit. The Pauli-Z operator, also known as the Pauli-Z matrix, is represented by the following 2x2 matrix:

$$\sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{9}$$

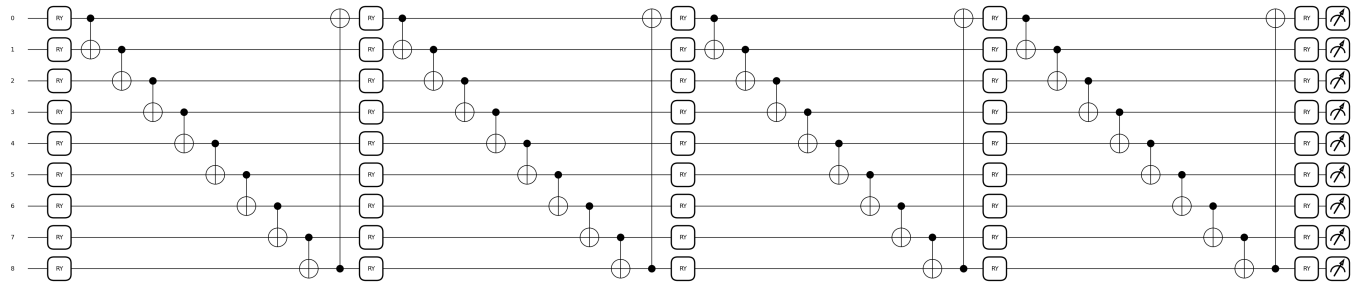


Fig. 2. The illustration of quantum encoding layer module of the proposed QEEGNET. The module contains nine qubits and four quantum circuit layers in this work.

This operator is used to measure the z-component of the spin of a qubit and has eigenvalues of +1 and -1, corresponding to the basis states $|0\rangle$ and $|1\rangle$, respectively.

D. Architecture of Quantum-EEGNet

The integration of quantum circuits into classical neural networks enables the creation of hybrid models that combine the strengths of both paradigms. We provided a novel classical-to-quantum Data encoding on EEG signal processing. The quantum encoding layer serves as a bridge, allowing the network to learn quantum-encoded features while being trained using classical optimization techniques. The overall architecture of QEEGNet is shown in TABLE I. The input EEG data will first encoding by the classical EEGNet, then the output features will encoding into the qubits through the quantum encoding layer module to exploit the high-dimensional information inherent in Hilbert space. The last stage of the model is output the quantum measurement result into fully connected layer to do the downstream classification task.

TABLE I
NETWORK STRUCTURE OF QUANTUMEEGNET

Layer	Input $(C \times T)$	Operation	Output
1	$1 \times 1 \times T$	Conv2D (1, 64)	$16 \times 1 \times (T-63)$
	$16 \times 1 \times (T-63)$	BatchNorm	$16 \times 1 \times (T-63)$
	$16 \times 1 \times (T-63)$	ELU	$16 \times 1 \times (T-63)$
2	$16 \times 1 \times (T-63)$	ZeroPad2D (16, 17, 0, 1)	$16 \times 1 \times (T-30)$
	$16 \times 1 \times (T-30)$	Conv2D (2, 32)	$32 \times 1 \times (T-61)$
	$32 \times 1 \times (T-61)$	BatchNorm	$32 \times 1 \times (T-61)$
	$32 \times 1 \times (T-61)$	ELU	$32 \times 1 \times (T-61)$
	$32 \times 1 \times (T-61)$	MaxPool2D (2, 4)	$32 \times 1 \times \frac{(T-61)}{4}$
	$32 \times 1 \times \frac{(T-61)}{4}$	Dropout (0.25)	$32 \times 1 \times \frac{(T-61)}{4}$
3	$32 \times 1 \times \frac{(T-61)}{4}$	ZeroPad2D (2, 1, 4, 3)	$32 \times 1 \times \frac{(T-52)}{4}$
	$32 \times 1 \times \frac{(T-52)}{4}$	Conv2D (8, 4)	$32 \times 1 \times \frac{(T-55)}{4}$
	$32 \times 1 \times \frac{(T-55)}{4}$	BatchNorm	$32 \times 1 \times \frac{(T-55)}{4}$
	$32 \times 1 \times \frac{(T-55)}{4}$	ELU	$32 \times 1 \times \frac{(T-55)}{4}$
	$32 \times 1 \times \frac{(T-55)}{4}$	MaxPool2D (2, 4)	$32 \times 1 \times \frac{(T-55)}{16}$
	$32 \times 1 \times \frac{(T-55)}{4}$	Dropout (0.25)	$32 \times 1 \times \frac{(T-55)}{16}$
4	$32 \times 1 \times \frac{(T-55)}{16}$	Quantum Encoding Layer	$9 \times N$
5	$9 \times N$	Fully Connected Layer	N

IV. RESULTS AND DISCUSSIONS

A. BCIC-IV-2a Dataset

We use the BCIC-IV-2a (Brain-Computer Interface Competition) dataset, which provides time-asynchronous EEG data. This dataset is one of the most popular public EEG datasets, released for the BCI Competition IV in 2008 [22]. It includes EEG recordings from nine subjects who performed a four-class motor-imagery task, repeated twice on different days. During the task, subjects imagined one of four movements (right hand, left hand, feet, and tongue) for four seconds after a cue. Each session had 288 trials, with 72 trials for each movement type. The EEG signals were recorded using 22 electrodes placed around the central region at a sampling rate of 250 Hz. We processed the EEG signals by down-sampling from 250 Hz to 128 Hz, applying a band-pass filter at 4-38 Hz, and segmenting the signals from 0.5 to 4 seconds after the cue, resulting in 438 time points per trial.

B. Experiment Details

For BCIC-IV-2a dataset, we used the first session of a subject for the training set, with one-fifth of it set aside for validation. Using one-fifth for validation instead of the more common one-eighth or one-ninth seen in some documents is intended to increase the difficulty of training and to increase the number of validation samples, with the hope of better highlighting the differences and generalization between quantum and classical models. The model that had the lowest validation loss within 100 epochs was then tested on the second session of the same subject. We trained QEEGNet, which has 9 qubits and 4 quantum layers. All the models are implemented by Pytorch and PennyLane frameworks with the simulator as backend. Input batch size for training is 32 with training 100 epochs, using AdamW as optimizer with 10^{-3} learning rate. We sellected the model with the best validation accuracy to do the prediction on the test dataset. The training time of QEEGNet per subject is almost 20 hours on CPUs in a Google Cloud Plateform a2-ultragpu-1g machine with 170 GB RAM.

C. Results

The experiment results are shown in TABLE II and TABLE III. The results highlight QEEGNet's consistent performance advantages over EEGNet in both validation and test datasets. This performance disparity can be attributed to the quantum layer integrated into QEEGNet, which likely enhances its ability to capture complex patterns and features within EEG data. The substantial improvements in validation and test accuracies for multiple subjects suggest that QEEGNet

TABLE II
COMPARISON OF THE HIGHEST VALIDATION ACCURACY RESULTS OF EEGNET AND QEEGNET ON THE BCIC-IV-2A DATASET.

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6	Subject 7	Subject 8	Subject 9	Average
EEGNet	42.7%	25.0%	46.4%	32.1%	25.0%	**42.9%**	**42.9%**	**32.1%**	69.6%	39.8%
QEEGNet	**50.0%**	**26.8%**	**50.0%**	**35.7%**	**32.1%**	41.1%	39.3%	30.4%	**73.2%**	**42.1%**

TABLE III
COMPARISON OF THE HIGHEST TEST ACCURACY RESULTS OF EEGNET AND QEEGNET ON THE BCIC-IV-2A DATASET.

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6	Subject 7	Subject 8	Subject 9	Average
EEGNet	47.9%	22.9%	**46.5%**	**32.6%**	26.4%	28.8%	33.7%	**32.3%**	62.2%	37.7%
QEEGNet	**49.3%**	**30.2%**	44.1%	30.6%	**26.7%**	**31.9%**	**36.8%**	28.1%	**65.3%**	**38.1%**

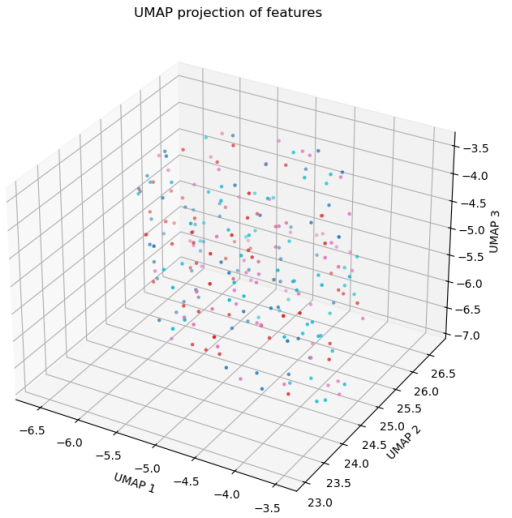


Fig. 3. 3D-UMAP projection of EEGNet features. Different classes are marked by different colors.

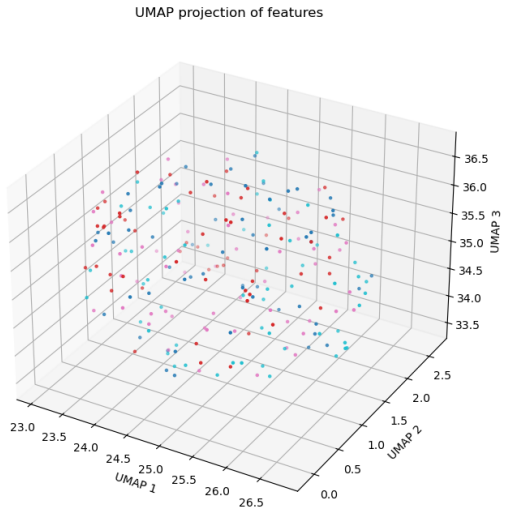


Fig. 4. 3D-UMAP projection of QEEGNet features. Different classes are marked by different colors.

can generalize better across different data variations. Moreover, the noticeable improvement in subjects with lower accuracies using EEGNet (e.g., Subject 2) emphasizes QEEGNet's robustness and potential for broader applicability.

D. Model Interpretation and Discussion

We use uniform manifold approximation and projection (UMAP) [23] method to compare the model embedding performance between two models. The provided UMAP projections visually represent the feature embeddings produced by EEGNet and QEEGNet models. These 3D scatter plots help in understanding the underlying structure and separability of the data as transformed by each model. The first UMAP projection in Fig 3. showcases the embeddings generated by the EEGNet model. The points in this plot are distributed with a noticeable spread, indicating that EEGNet captures a variety of features from the input data. However, the embeddings exhibit some degree of overlap, suggesting potential challenges in distinguishing between different classes or patterns. The spread of points suggests a moderate clustering tendency, but with some intermingling between clusters, indicating that while EEGNet can learn useful features, its ability to distinctly separate different classes might be limited. In contrast, the second UMAP projection in Fig 4. depicts the embeddings from the QEEGNet model. The points in this plot appear more tightly clustered compared to the EEGNet embeddings. This suggests that QEEGNet has a better capability to group similar features together, enhancing the separability between different classes or patterns. The clusters are more distinct and less overlapping, indicating that QEEGNet is more effective in capturing and distinguishing between complex features in the data. The comparison between the two UMAP projections highlights the improvements brought by integrating quantum layers into the QEEGNet model. QEEGNet's embeddings show a more pronounced clustering effect, which can be attributed to its enhanced feature extraction capabilities. The tighter and more distinct clusters suggest that QEEGNet is better at capturing the underlying structure of the EEG data, leading to improved performance in classification tasks, as evidenced by the higher validation and test accuracies discussed previously. Furthermore, the clearer separation in QEEGNet's embeddings implies that the model can more effectively learn and represent the unique characteristics of different classes. This can lead to more robust and reliable

predictions, particularly in complex datasets where traditional models like EEGNet might struggle.

V. Conclusion

In this paper, we introduced QEEGNet, a novel hybrid neural network that integrates quantum computing with the traditional EEGNet architecture to enhance the encoding and analysis of EEG data. Our experimental results on the BCIC-IV-2a dataset demonstrate that QEEGNet overall outperforms traditional EEGNet in terms of accuracy and robustness across most subjects. The integration of quantum layers within the neural network allows QEEGNet to capture more intricate patterns in EEG data, suggesting a significant potential for quantum-enhanced neural networks in the field of EEG analysis.

The findings highlight the practical feature embedding ability of QEEGNet, showcasing its advantage in feature representation over traditional models. The UMAP projections further validate these improvements by illustrating the superior clustering and separability of features learned by QEEGNet. These enhancements underline the potential of quantum machine learning to provide computational advantages and explore larger solution spaces, paving the way for more effective and reliable applications in neuroscience and clinical settings.

The performance improvements achieved by QEEGNet suggest new directions for research and practical applications, emphasizing the importance of further exploring and developing quantum-enhanced models for complex and high-dimensional data analysis tasks. The study opens up opportunities for integrating advanced quantum computing techniques with classical neural networks to achieve substantial gains in performance and robustness in EEG signal processing and beyond.

References

[1] M. Rashid, N. Sulaiman, A. Majeed, R. M. Musa, A. Fakhri, B. S. Bari, and S. Khatun, "Current status, challenges, and possible solutions of eeg-based brain-computer interface: A comprehensive review," *Frontiers in neurorobotics*, vol. 14, Jun 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7283463/

[2] J. D. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, "Quantum machine learning," *Nature*, vol. 549, no. 7671, p. 195–202, Sep 2017. [Online]. Available: https://www.nature.com/articles/nature23474

[3] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heimonen, J. S. Kottmann, T. Menke *et al.*, "Noisy intermediate-scale quantum algorithms," *Reviews of Modern Physics*, vol. 94, no. 1, p. 015004, 2022.

[4] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio *et al.*, "Variational quantum algorithms," *Nature Reviews Physics*, vol. 3, no. 9, pp. 625–644, 2021.

[5] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, p. 056013, Jul 2018. [Online]. Available: https://arxiv.org/abs/1611.08024

[6] G. Li, C. H. Lee, J. J. Jung, Y. C. Youn, and D. Camacho, "Deep learning for eeg data analytics: A survey," *Concurrency and computation*, vol. 32, no. 18, Feb 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.5199

[7] G. Tosato, C. M. Dalbagno, and F. Fumagalli, "Eeg synthetic data generation using probabilistic diffusion models," 2023. [Online]. Available: https://arxiv.org/abs/2303.06068

[8] X. Jiang, G.-B. Bian, and Z. Tian, "Removal of artifacts from eeg signals: A review," *Sensors*, vol. 19, no. 5, p. 987–987, Feb 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6427454/

[9] P.-H. Lai, B.-S. Wang, W.-C. Yang, H.-C. Tsou, and C.-S. Wei, "Cleegn: A convolutional neural network for plug-and-play automatic eeg reconstruction," 2022. [Online]. Available: https://arxiv.org/abs/2210.05988

[10] C.-S. Chen and C.-S. Wei, "Mind's eye: Image recognition by eeg via multimodal similarity-keeping contrastive learning," 2024. [Online]. Available: https://arxiv.org/abs/2406.16910

[11] K. M. Hossain, M. A. Islam, S. Hossain, A. Nijholt, and A. Rahman, "Status of deep learning for eeg-based brain–computer interface applications," *Frontiers in computational neuroscience*, vol. 16, Jan 2023. [Online]. Available: https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2022.1006763/full

[12] A. Padha and A. Sahoo, "Quantum deep neural networks for time series analysis," *Quantum Information Processing*, vol. 23, no. 6, May 2024. [Online]. Available: https://link.springer.com/article/10.1007/s11128-024-04404-y

[13] I. Cong, S. Choi, and M. D. Lukin, "Quantum convolutional neural networks," *Nature physics*, vol. 15, no. 12, p. 1273–1278, Aug 2019. [Online]. Available: https://arxiv.org/abs/1810.03787

[14] C. Zoufal, A. Lucchi, and S. Woerner, "Quantum generative adversarial networks for learning and loading random distributions," *npj quantum information*, vol. 5, no. 1, Nov 2019. [Online]. Available: https://www.nature.com/articles/s41534-019-0223-2

[15] S. Y.-C. Chen, S. Yoo, and Y.-L. L. Fang, "Quantum long short-term memory," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9747369

[16] A. Mari, T. R. Bromley, J. Izaac, M. Schuld, and N. Killoran, "Transfer learning in hybrid classical-quantum neural networks," *Quantum*, vol. 4, p. 340–340, Oct 2020. [Online]. Available: https://quantum-journal.org/papers/q-2020-10-09-340/

[17] M. Yousif, B. Al-Khateeb, and B. Garcia-Zapirain, "A new quantum circuits of quantum convolutional neural network for x-ray images classification," *IEEE access*, p. 1–1, Jan 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10517587

[18] T. Koike-Akino and Y. Wang, "queegnet: Quantum ai for biosignal processing," *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, Sep 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9926814

[19] R. Ho and K. Hung, "Exploring quantum machine learning for electroencephalogram classification," *2023 IEEE 13th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, May 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10165407

[20] G. Aksoy, G. Cattan, S. Chakraborty, and M. Karabatak, "Quantum machine-based decision support system for the detection of schizophrenia from eeg records," *Journal of medical systems*, vol. 48, no. 1, Mar 2024. [Online]. Available: https://link.springer.com/article/10.1007/s10916-024-02048-0

[21] I. D. Lins, M. Mendes, C. Bezerra, S. Ramos, J. das, A. J. Ferreira-Martins, R. Chaves, and A. Canabarro, "Quantum machine learning for drowsiness detection with eeg signals," *Process safety and environmental protection/Transactions of the Institution of Chemical Engineers. Part B, Process safety and environmental protection/Chemical engineering research and design/Chemical engineering research & design*, vol. 186, p. 1197–1213, Jun 2024. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2024PSEP..186.1197L/abstract

[22] M. Tangermann, K. R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. R. Müller-Putz, G. Nolte, G. Pfurtscheller, H. Preissl, G. Schalk, A. Schlögl, C. Vidaurre, S. Waldert, and B. Blankertz, "Review of the bci competition iv," *Frontiers in Neuroscience*, vol. 6, Jan 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3396284/

[23] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2018. [Online]. Available: https://arxiv.org/abs/1802.03426

Dual-TSST: A Dual-Branch Temporal-Spectral-Spatial Transformer Model for EEG Decoding

Hongqi Li* *Member, IEEE,*, Haodong Zhang, Yitong Chen

***Abstract*—The decoding of electroencephalography (EEG) signals allows access to user intentions conveniently, which plays an important role in the fields of human-machine interaction. To effectively extract sufficient characteristics of the multichannel EEG, a novel decoding architecture network with a dual-branch temporal-spectral-spatial transformer (Dual-TSST) is proposed in this study. Specifically, by utilizing convolutional neural networks (CNNs) on different branches, the proposed processing network first extracts the temporal-spatial features of the original EEG and the temporal-spectral-spatial features of time-frequency domain data converted by wavelet transformation, respectively. These perceived features are then integrated by a feature fusion block, serving as the input of the transformer to capture the global long-range dependencies entailed in the non-stationary EEG, and being classified via the global average pooling and multi-layer perceptron blocks. To evaluate the efficacy of the proposed approach, the competitive experiments are conducted on three publicly available datasets of BCI IV 2a, BCI IV 2b, and SEED, with the head-to-head comparison of more than ten other state-of-the-art methods. As a result, our proposed Dual-TSST performs superiorly in various tasks, which achieves the promising EEG classification performance of average accuracy of 80.67% in BCI IV 2a, 88.64% in BCI IV 2b, and 96.65% in SEED, respectively. Extensive ablation experiments conducted between the Dual-TSST and comparative baseline model also reveal the enhanced decoding performance with each module of our proposed method. This study provides a new approach to high-performance EEG decoding, and has great potential for future CNN-Transformer based applications.**

***Index Terms*—EEG decoding, feature fusion, transformer, convolutional neural network, signal processing.**

I. Introduction

BRAIN-COMPUTER/MACHINE interfaces (BCIS/BMIS) have garnered much attention over the past decades due to their outstanding ability to convert the users' brain activity into machine-readable intentions or commands [1]. Among various BCI modalities, noninvasive electroencephalograph (EEG) has the advantages of adequate temporal resolution, non-surgical electrode placements, and low cost, thus leading to its widest application in the fields of rehabilitation engineering [2], [3], cognitive science [4], neuroscience, and psychology [5].

Various brain paradigms, such as the motor imagery (MI), event-related P300, and steady-state visual evoked potentials (SSVEP), have been extensively studied by researchers [5], and a complete EEG-based BCI system generally consists of the user intention decoding process of the signal acquisition, preprocessing, feature extraction, classification, and a final application interface for the control signal convert. To gain an accurate interpreting of the sampled EEG, two main categories of recognition methods, traditional machine learning (ML) algorithms [6] and advanced deep learning (DL) techniques [7], [8], have been innovatively investigated. Traditional ML methods usually involve feature extraction and feature classifi-cation, where the former procedure uses algorithms of the common spatial pattern (CSP), Filter bank CSP, fast Fourier transform (FFT), wavelet transform, etc. As for the feature classification, the supervised learning approach (e.g., linear discriminant analysis (LDA), support vector machine (SVM)) and the unsupervised methods (e.g., K nearest neighbor (KNN)) have been shown to be effective. However, since features are manually extracted from the raw non-stationary EEG with low signal-to-noise ratio, the specific expertise is generally required, leading to the process being time-consuming and complicated. Worse more, the useful information may also be lost due to insufficient expert experience. The DL methods, on the other hand, allow end-to-end models that are composed of multiple processing layers to learn the data representation automatically, thereby minimizing the need for human manual intervention and domain-specific preprocessing, and have already achieved excellent even the state-of-the-art (SOTA) performance in several domains such as computer vision [9], [10], and natural language processing [11].

Specifically for the EEG decoding, the convolutional neural networks (CNNs)-based ConvNet has reached comparable classification results to the traditional ML methods [12]. A compact network called EEGNet has been proposed in [13], which utilized depth wise and separable convolutions to build an EEG-specific model that capable of learning features across various tasks. Moreover, a long short-term memory (LSTM) based recurrent neural network (RNN) has been developed by Tortora et al. in [14] for decoding the gait events from EEG, where the network's ability to handle the time-dependent information was fully leveraged. However, despite these commendable advances, CNNs and RNNs are not perfect in pro-

Manuscript received Aug 21, 2024. This work was supported in part by the Natural Science Basic Research Program of Shaanxi Province under Grant 2024JC-YBQN-0659, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515110252, in part by the Basic Research Programs of Taicang under Grant TC2023JC16, in part by the Fundamental Research Funds for the Central Universities under Grant D5000210969 (Corresponding author: Hongqi Li.)

H. Li is with the School of Software, Northwestern Polytechnical University, Xi'an 710072, China, and also with the Research & Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen 518063, China, and also with the Yangtze River Delta Research Institute of Northwestern Polytechnical University, Taicang 215400, China (e-mail: lihongqi@nwpu. edu.cn)

H. Zhang and Y. Chen are with the School of Software, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: zhang_haodong@mail. nwpu.edu.cn, chenyt@mail.nwpu.edu.cn).

cessing EEG signals. More specifically, while CNNs are good at learning local features, it is difficult to obtain long-term dependencies across the whole data scale. RNNs, on the other hand, are also prone to difficulties in capturing long-term dependencies in long sequence data. Therefore, to address these shortcomings, the research processing for sequence signals is gradually shifting to the self-attention mechanism, which allows each element in a sequence to be processed taking into account the relationship with all other elements, thus capturing richer contextual features. Moreover, the multi-feature analysis of EEG has also increasingly attracted attention considering that the sampled signals contain multi-dimensional features of the temporal, spectral, and spatial domains.

A. Related Work

One of the most famous models based on the self-attention mechanism is the Transformer model, which has recently been attempted to EEG decoding. Sun et al. [15] introduced a novel approach by integrating the multi-head attention mechanisms with CNNs for motor imagery tasks, and the various positional embedding techniques were used to improve the classification accuracy. As a result, the introduced five Transformer-based models have significantly outperformed existing models. Similarly, a compact hybrid model of CNNs and Transformers, named EEG Conformer, was developed to decode EEG signals by capturing both the local and global features, which was excelled on three public datasets and potentially established a new baseline for EEG processing [16]. The proposed AD-FCNN in [17] utilized the convolutions at two different scales to capture comprehensive spatial details in EEG data, and the features were fused through a self-attention mechanism. Moreover, Arjun [18], Al-Quraishi [19], and Mulkey [20] first converted the EEG data into time-frequency images, and then used Vision Transformers based on the idea of computer vision field. In the realm of pretrained models, different models of the BERT [21], GPT and Swin Transformer model [22] have been designed to transform the EEG into textual and visual formats for the further processing. Particularly, a Transformer-like recognition approach of Speech2EEG has been proposed to leverage the pretrained speech processing networks for the robust EEG feature aggregation, thereby boosting EEG signal analysis capabilities [23]. Given these above advancements, the promising potential of Transformers in EEG decoding has been well demonstrated.

On the other hand, with the development of DL models, a single feature can no longer satisfy the requirement of the performance improvement for increasingly complex models in EEG decoding, and therefore, multi-feature analysis methods gradually occupy the mainstream of EEG analysis methods. In 2019, Tian et al. [24] crafted a multi-view DL strategy that first transforming the raw data into representations in the frequency and time-frequency domains, then independently extracting the features, which were finally merged to perform classification tasks efficiently. The data from multiple frequency bands was used in [25] to create multi-view representations, where the spatial discrimination patterns of the views were learned by CNN, temporal information was aggregated by a variance layer, and the resultant features were classified by a fully connected layer. Recently, a multi-domain CNN model of TSFCNet was developed for MI decoding, which significantly outperformed the other traditional methods by extracting multi-scale features from the time domain and capturing the additional spatial, frequency, and time-frequency features [26]. Earlier in this year, Liang et al. [27] developed an EISATC-Fusion model to leverage the multi-scale EEG frequency band information combined with the attention mechanism and temporal convo¬lutional networks (TCN) for an integrated feature extraction process. In addition, a lightweight multi-feature attention CNN was proposed in [28] to extract the information from frequency, localized spatial domains, and feature maps to enhance the precision of EEG analysis, where a hybrid neural network of SHNN was designed to autonomously extract the spatial, spectral, and temporal features from EEG [29]. In conclusion of these mentioned studies, the research community has tended to extract the temporal-spatial-spectral features simultaneously, which helps to improve the understanding and decoding effects of specific EEG signals. However, since the EEG signals are first collected and expressed in temporal domain, while the frequency/time-frequency/spatial features are represented or converted by various approaches, how to efficiently extract and integrate the features from different dimensions and establish a more robust extraction process still remain challenging.

B. Contribution and Overview

As mentioned earlier, the application of Transformer-based and multi-feature analysis in EEG decoding has just emerged and is in a phase of continuous development, and there are few attempts to naturally combine the two. Since the convolutional networks-based models are able to automatically learn more discriminative local features from raw EEG data, while the attention-based Transformer adepts to describe the long-range dependencies, the combination of these two modules is envisioned to benefit each other for a more comprehensive interpretation of human user EEG data.

Driven by this insight, in present work, a novel decoding architecture model with dual-branch temporal-spectral-spatial transformer, termed as Dual-TSST, is proposed to extract multi-dimensional features hidden in EEG while considering their global correlations. Specifically, the proposed architecture mainly consists of three parts of the feature extraction, feature fusion, and classification modules. The first feature extraction module is composed of two branches of convolutional neural networks to receive multi-view inputs from raw EEG and to extract the inherent temporal-spectral-spatial features. These obtained features are fed into the feature fusion module to be jointly concatenated and then to learn their global relationships by a Transformer, and a classifier composed of multilayer perceptron and global pooling layers is finally used to achieve the results output. The main contributions of this study are summarized below.

1) We propose a natural fusion and collaboration architecture based on the classical CNNs and emerging Transformer, which is highly generalizable to a wide range of EEG decoding tasks. Specifically, the developed network enables to extract abundant powerful features without handcraft while allowing

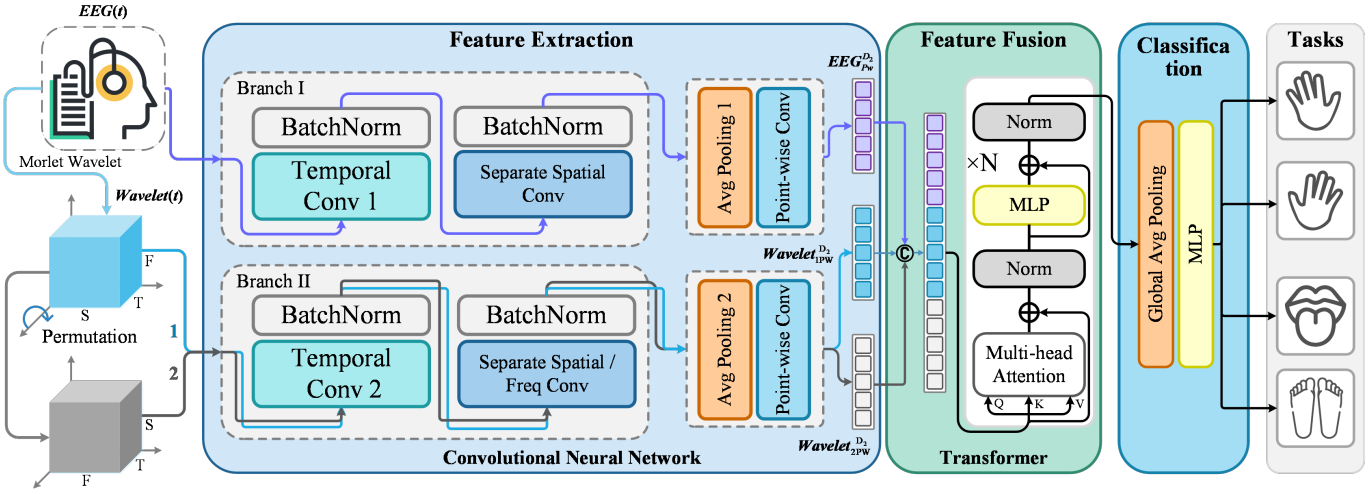


Fig. 1 The proposed Dual-TSST framework, including feature extraction of two CNN branches, feature fusion with Transformer, and classification modules.

long-range correlation among features being considered and processed concurrently.

2) The designed Dual-TSST mainly comprises dual-scale convolution networks, wherein one is used to better extract the temporal feature from the raw EEG, and the other enables to acquire the time-frequency/time-spatial information from the converted EEG signals. These features are in the same scale to be concatenated and effectively jointly fused by a fusion patch. A self-attention mechanism is applied to adaptively enhance the flexibility of the feature fusion.

3) Dual-TSST has undergone extensive experiments on multiple public datasets to demonstrate the model structure and superior performance, the compared results with state-of-the- art models proved the effectiveness of the proposed method.

The rest of this paper is organized as follows. Section II introduces the design ideas and specific structural principles of Dual-TSST. Section III presents the used datasets with related data preprocessing, and experimental setups. The comparable results and visualized model effects are presented in Section IV, while a detailed discussion and conclusion is given in the final Section V.

II. Approach of Dual-TSST Network

The EEG signals are notable for their exceptional temporal resolution while encompassing extensive spectral and spatial properties. With the goal of processing EEG with multiple features involved being considered adequately and efficiently, a generalized network adheres to machine learning principles has been proposed, in which the advanced deep learning techniques are utilized to perform feature extraction, feature fusion, and the final classification step by step.

A. Overall Model Architecture

Some traditional practice of EEG decoding generally uses exclusive raw EEG or solely time-frequency images derived from the transformed data, which may lead to leakage of contained information during the conversion process. Instead, as illustrated in Fig. 1, a dual-branch model named Dual-TSST, capable of processing diverse views of EEG, is designed to start with both the given raw and converted EEG signals. For the first module of the feature extraction, two branches based on convolutional neural networks are applied to sufficiently extract potential characteristics from the temporal, frequency, and spatial domain. Branch II, in particular, is uniquely designed to simultaneously analyze the wavelet-transformed time-frequency EEG data in two separate flip-flop formats, collecting the time-frequency-space features comprehensively. To reduce the model complexity and computational load, the depth wise separable convolutions and average pooling layers are employed in this module.

The acquired features from branch I and branch II are then synergistically integrated and serve as the input of patch embedding for the feature fusion part, where a Transformer module is exploited to learn global relationships among the extracted properties. Ultimately, for the classification module, a global average pooling (GAP) layer and multilayer perceptron (MLP) module is used to analyze the inputted features and deliver the final classification outcomes.

B. Source of Data

1) Data Input: The original EEG data can be represented as $\boldsymbol{EEG}(t) \in \mathbb{R}^{ch \times T}$, where *ch* is the number of electrodes indicating spatial dimensions, and *T* represents the time samples of the EEG data. Initially, to convert the two-dimensional time-domain data into three-dimensional time-frequency domain, the Morlet wavelet transform [30] provided by MNE-Python is applied, for which the process is described by

$$\boldsymbol{W}(a,t) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} \boldsymbol{EEG}(\tau) \boldsymbol{\Psi} \left(\frac{\tau - t}{a} f_o \right) d\tau \qquad (1)$$

where $\boldsymbol{W}$(a,t) represents the transformed outcomes, a is the scale parameter related to frequency and sampling rate, fo means the central frequency, τ is the time variable, $\boldsymbol{\Psi}(t)$ is the wavelet function as

$$\boldsymbol{\Psi}(t) = \frac{1}{\sqrt{\sigma\sqrt{\pi}}} e^{-t^2/\sigma_t^2} e^{i2\pi f_o t} \qquad (2)$$

and σ_t is the wavelet's temporal standard deviation.

For the Morlet wavelet transformation, we set the frequency *freq* to match the frequency range used in the original data

filtering. Noting that the number of cycles (i.e., n_cycle) determines the width of the wavelet in the transformation, and is related to the temporal standard deviation σ_t. Larger n_cycle results in a wider wavelet, leading to lower time but higher frequency resolution. Conversely, a smaller n_cycle results in a narrower wavelet, enhancing both the time resolution and frequency resolution. To achieve a balance, n_cycle is set to be half of the frequency, i.e., n_cycle = *freq*/2, and the sampling rate for the time resolution remains the same as that of the original EEG signal. We use $\boldsymbol{Wavelet}(t) \in \mathbb{R}^{ch \times T \times F}$ to represent the transformed time-frequency data.

Then, both original $\boldsymbol{EEG}(t)$ and transformed time-frequency Wavelet data $\boldsymbol{Wavelet}(t)$ are subjected to Z-Score normalization, which preserves the data's dimensional shape while ensuring consistency in analysis and can be represented as:

$$\boldsymbol{x}^{'} = \frac{\boldsymbol{x} - \mu}{\sigma} \tag{3}$$

where $\boldsymbol{x}$ and $\boldsymbol{x}^{'}$ represents the input and output data, μ and σ are the calculated mean and standard deviations.

2) Data Augmentation: To mitigate the challenge of limited EEG data availability in decoding, several data augmentation strategies can be applied. Here the Segment and Reassemble (S&R) mechanism is adopted. More specifically, each EEG sample from the same category and its corresponding time-frequency data are divided into a predetermined number of fixed segments (labeled *R*). These segments are subsequently reconnected in various random orders that respect the original temporal sequence. This technique not only diversifies the training dataset but also enhances the model's ability to generalize from limited data samples. Following the guidelines set forth in references [16], [31], we generated augmented data in each epoch, matching the batch size, thereby ensuring consistent model training across different data permutations.

C. Feature Extraction based on Dual CNN Branches

1) Branch I for original $\boldsymbol{EEG}(t)$: As shown in Fig. 2, the shape of the inputted 2D EEG data is $[ch \times T_{B1}]$. To extract features in the temporal dimension, the time convolution is first used, resulting in a 3D feature map of $\boldsymbol{EEG}_{TC}^{D_1}$, with the shape $[D_1 \times ch \times T_{B1}1]$. Here, to capture local details in the temporal dimension as much as possible, the time convolution kernel size is set to be relatively small. The relevant process can be summarized as:

$$\boldsymbol{EEG}(t)_{TC}^{D_1} = TimeConv(\boldsymbol{EEG}^{'}) \tag{4}$$

Then, separable spatial convolution compresses the spatial dimension and extracts features from the electrodes, changing the feature map shape to $[D_1 \times 1 \times T_{B1}1]$. It should be noticed that, to ensure the performance, Batch Norm layers (see in Fig. 1) are added after the time convolution and separable spatial convolution, and the ELU activation function is also employed. The above data flow can be expressed as follows:

$$\boldsymbol{EEG}_{SSC}^{D_1} = ELU(BN(SSConv(\boldsymbol{EEG}_{TC}^{D_1}))) \tag{5}$$

where BN means the batch normalization function, and $SSConv$ indicates the related separable spatial convolution.

After that, an average pooling layer is used to extract features while reducing the data in the temporal dimension. With the enhanced generalization and noise suppressing ability, a feature map of $\boldsymbol{EEG}_{AP}^{D_1}$is derived, as the shape of $[D_1 \times 1 \times T_{B1}2]$. Finally, pointwise convolutions are applied for channel fusion and increasing the channel dimension to some extent, as enhancing the data's information content and expressive power. The final feature map $\boldsymbol{EEG}_{Pw}^{D_2}$ is in shape of $[D_2 \times 1 \times T_{B1}2]$. The entire data flow of these mentioned operations can be represented by the following process:

$$\boldsymbol{EEG}_{Pw}^{D_w} = PWConv(AP(\boldsymbol{EEG}_{SSC}^{D_1})) \tag{6}$$

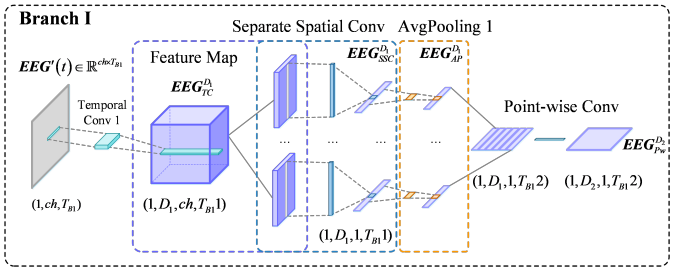


Fig. 2 Structure and data flow of Branch I for feature extraction module.

2) Branch II for converted $\boldsymbol{Wavelet}(t)$: Branch II is designed for processing the time-frequency EEG data. As illustrated in Fig. 1, to capture multidimensional features, distinct inputs in different viewpoints (i.e., Input 1 and Input 2) are fed into such branch. Here, the Input 2 is obtained by subtly transposing Input 1 by 90 degrees. Unlike branch I of processing original EEG in a single line, branch II is actually perform simultaneous processing of multi-inputs. Specifically, the time convolution is applied first with a followed Batch normalization, which can be expressed as:

$$\boldsymbol{Wavelet}_{i}{}_{TC}^{D_1} = TimeConv(\boldsymbol{Wavelet}_i^{'}), i = 1, 2 \tag{7}$$

where $\boldsymbol{Wavelet}_i (i = 1, 2)$ represents the branch inputs, and $\boldsymbol{Wavelet}_{i}{}_{TC}^{D_1}$ is the relevant output.

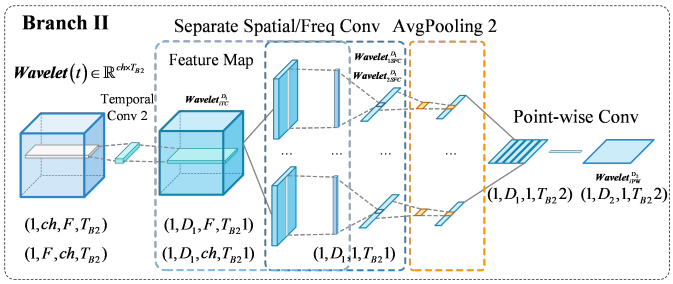


Fig. 3 Structure and data flow of Branch II for feature extraction module.

Given the differences in temporal resolution and information content between the time-frequency $\boldsymbol{Wavelet}(t)$ and the original $\boldsymbol{EEG}(t)$, a different time convolution of scale is used, which has shapes $[ch \times F \times T_{B2}]$ and $[F \times ch \times T_{B2}]$, respectively. Indeed, as shown in Fig. 3, such the choice is aimed to balance the features derived from the original and time-frequency data that will produce feature maps $[D_1 \times F \times T_{B2}1]$ and $[D_1 \times ch \times T_{B2}1]$.

Similarly, separable spatial and frequency convolutions are employed for feature extraction and dimension compression in the spatial and frequency dimensions, resulting in feature

maps $\boldsymbol{Wavelet}_1{}_{SSC}^{D_1}$ and $\boldsymbol{Wavelet}_2{}_{SFC}^{D_1}$, both with shapes $[D_1 \times 1 \times T_{B2}1]$. The detailed operation of these two processing are:

$$\boldsymbol{Wavelet}_1{}_{SSC}^{D_1} = ELU(BN(SSConv(\boldsymbol{Wavelet}_1{}_{TC}^{D_1}))) \quad (8a)$$

$$\boldsymbol{Wavelet}_2{}_{SFC}^{D_1} = ELU(BN(SFConv(\boldsymbol{Wavelet}_2{}_{TC}^{D_1}))) \quad (8b)$$

Subsequently, an average pooling layer is used to suppress noise, extract features, and reduce data volume, resulting in data with shape $[D_1 \times 1 \times T_{B2}1]$. Finally, pointwise convolutions are applied to achieve channel fusion and dimension elevation, producing the feature maps $\boldsymbol{Wavelet}_1{}_{PW}^{D_2}$ and $\boldsymbol{Wavelet}_2{}_{PW}^{D_2}$, each with the shape $[D_2 \times 1 \times T_{B2}1]$. Similarly, the hyper parameters D1 and D2 are set to be the same of branch I. The entire data flow of these descriptions is as follows:

$$\boldsymbol{Wavelet}_i{}_{PW}^{D_2} = PWConv(AP(\boldsymbol{Wavelet}_i{}_{SSC/SFC}^{D_1})) \quad (9)$$

D. Feature Fusion based on Transformer

Three representative feature characteristics can be acquired from the above feature extraction process with Branch I and Branch II. To better integrate them, we reshape these outputs to be $\boldsymbol{EEG}_S^{D_2}$, $\boldsymbol{Wavelet}_1{}_S^{D_2}$, and $\boldsymbol{Wavelet}_2{}_S^{D_2}$ with shapes of $[T_{B1}2 \times D_2]$, $[T_{B2}2 \times D_2]$, and $[T_{B2}2 \times D_2]$, respectively. Such dimensional conversion is employed to suit the data need of the succeeding Transformer, which is applied to learn the cross-channel context information and the appropriate Encoder accepts inputs shaped as [SeqLength × FeatureSize]. The reshaped feature maps are horizontally concatenated to form a unified dataset $\mathbf{EW}_{Fusion}$, which represents a fusion of the original EEG and time-frequency Wavelet data:

$$\mathbf{EW}_{Fusion} = \text{Concat}(\boldsymbol{EEG}_S^{D_2}, \boldsymbol{Wavelet}_1{}_S^{D_2}, \boldsymbol{Wavelet}_2{}_S^{D_2}) \quad (10)$$

The new feature $\mathbf{EW}_{Fusion}$, which takes on the shape $[D_2 \times T_{B1}2 \times T_{B2}2 * 2]$, is then processed using a multi-head attention mechanism within a complete Transformer Encoder. This setup captures the detailed correlations within the input sequence, thereby obtaining comprehensive global characteristics across time, space, and frequency dimensions from the combined EEG and time-frequency data.

An encoding approach akin to those in Vision Transformers is adopted, which involves parameterizable position encodings initialized with random values as:

$$\mathbf{P} = Parameter(\mathbf{P}_{init}) \quad (11)$$

$$\mathbf{X}_P = \mathbf{EW}_{Fusion} + \mathbf{P} \quad (12)$$

where $\mathbf{P}$ is the position encoding matrix, $\mathbf{P}_{init}$ is its initial value, determined by random numbers, and $\mathbf{X}_P$ represents the encoded feature matrix with shape of $[D_2 \times T_{B1}2 \times T_{B2}2 * 2]$, which is subsequently mapped to the Query (Q), Key (K), and Value (V) spaces through linear transformations, with learnable weighting matrices $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, and $\boldsymbol{W}_V$ as:

$$\boldsymbol{Q} = \boldsymbol{X}_P\boldsymbol{W}_Q, \boldsymbol{K} = \boldsymbol{X}_P\boldsymbol{W}_K, \boldsymbol{V} = \boldsymbol{X}_P\boldsymbol{W}_V \quad (13)$$

$$Attention(Q, K, V) = Softmax(QK^T/\sqrt{D_2})V \quad (14)$$

where D_2 is the dimensionality of patches within the data.

The Transformer Encoder applies multi-head attention to parallelize the computation on data, thereby enhancing the expressivity and efficiency of the model and improving its generalizability. Multi-head attention (MHA) includes several self-attention layers, where each head generates an attention output, and the outputs from all heads are concatenated to form the final multi-head attention, as depicted in the following:

$$MHA(Q, K, V) = Concat(head_1, ...head_h)W^O \quad (15)$$

$$Head_i = SelfAttention(QW_i^Q, KW_i^K, VW_i^W) \quad (16)$$

where h denotes the number of heads, W^o is the weight matrix that integrates information captured by different heads of $head_i$.

After the multi-head attention mechanism, as can be seen in Fig. 1, a series of residual connections and layer normalization are performed to facilitate the information flow and stabilize the training process. The output is further processed using the MLP, followed by additional layer normalization, and residual connections, culminating in the final outputs from the multiple Encoder layers.

E. Classification Module

The outcome of the Transformer Encoder maintains the same dimensional structure as its input. To effectively distill this complex data, global average pooling (GAP) is employed, which simplifies the feature map by averaging out the features over the entire spatial extent of each channel. This process extracts pivotal global information that is crucial for the next stage of processing.

Following the pooling, the data is routed to an MLP module with two linear layers. The Softmax function, which normalizes the linear outputs to form a probability distribution over the predicted output classes, aids in the transformation of the pooled features into an M-dimensional vector. The model's performance is evaluated using a cross-entropy loss function, which is essential for classification tasks and is mathematically represented as:

$$l = -\frac{1}{N_b}\sum_{i=1}^{N_b}\sum_{j=1}^{N_c} ylog(\hat{y}) \quad (17)$$

where N_b is the batch size indicating the number of samples processed per training iteration, N_c denotes the total number of categories in the classification task, y is the true label of the data, and $\hat{y}$ is the predicted probability for each class. Briefly, the function can effectively measures the difference between the predicted probabilities and the actual distribution, guiding the model towards more accurate predictions through training.

III. Dataset and Experimental Setup

To evaluate the proposed method, we utilized three public datasets. Specifically, two BCI competition datasets in MI [32], sourced from MOABB (Mother of All BCI Benchmarks) project [33], and one widely used emotional SEED dataset [34] are included. This section gives the relevant introduction and several required necessary procedures.

A. Datasets

Dataset I: **BCI Competition IV 2a** - This dataset comprises EEG recordings from 9 subjects that performing four distinct MI tasks, i.e., imagery movements of the left hand, right hand, both feet, and the tongue. Data were collected by 22 positioned Ag/AgCl electrodes according to international 10-20 system. To ensure signal quality, a 250 Hz sampling rate was utilized and the recorded data was filtered between 0.5 Hz and 100 Hz. The dataset includes two sessions, where the first session serves as the training set, and the second as the test set. Each session consists of six runs, with 48 trials per run that distributed evenly across task categories. In our study, we set the time window for each trial of this dataset between 2 and 6 seconds, and filtered the data using a frequency range from 0 to 40 Hz.

Dataset II: **BCI Competition IV 2b** - This dataset features by the data from 9 subjects engaged in left and right hand MI tasks. The recording data were captured from three electrodes of C3, Cz, and C4, with a sampling frequency of 250 Hz. A band-pass filter of 0.5-100 Hz and a notch filter at 50 Hz have been used. Each subject participated in five sessions, where the initial two collected data without visual feedback and the subse¬quent three sessions included online feedback. Moreover, the dataset designates the initial three sessions (400 trials in total) for training and the final two (i.e., 320 trials) for testing. Noting that in our study each trial is allocated a time window from 3 to 7.5 s, with data similarly filtered within the 0 to 40 Hz range.

Dataset III: **SEED** - Provided by BCMI Lab from Shanghai Jiao Tong University, this dataset consists of EEG data from 15 subjects who viewed clips from Chinese films edited to evoke various emotions (e.g., positive, negative, neutral). The films last for 4 minutes, with data processed using 1-seconds or 4s sliding windows across 62 channels, and downsampled to 200 Hz. Each subject underwent three experimental sessions, with data filtered through a 0-75 Hz band-pass filter. In addition, five-fold/ten-fold cross-validation techniques were involved in training. Also, a band-pass filter ranging from 0.5 Hz to 50 Hz was utilized on the SEED dataset, and the continuous data from each experiment was segmented into 1-second windows.

B. Experiment Setting

We constructed the developed model using Python 3.11 and PyTorch 2.0, and conducted training on a Nvidia GeForce RTX 4090 GPU using the Adam optimizer. The Adam optimizer was configured with a learning rate of 0.0001 and a weight decay of 0.0012, with β_1 and β_2 values at 0.5 and 0.999, respectively. Throughout the training, the epoch value was set to 1000, with a batch size of 32. The critical hyperparameters D1 and D2 were set to 40 and 120. On Datasets I and II, the data augmentation parameters R were designated as 8 and 9. Since the data scale is enough, no data augmentation was applied in Dataset III. The learning rate was adjusted with Cosine Annealing [35], which can be explained by the following formula:

$$lr = lr_{min} + 1/2(lr_{max} - lr_{min})(1 + cos\frac{T_{cur}}{T_{max}}\pi) \quad (18)$$

where lr is the current learning rate, lr_{max} and lr_{min} are the related maximum and minimum values, respectively. T_{cur} is the current training epoch, and T_{max} is the total number of training epochs in a cycle. The learning rate decreases to lr_{min} at the end of a cycle. For the experiments, T_{max} was set to 32 to allow better model convergence and generalization during training.

C. Choice of Model Parameters

Table I illustrates the input shapes, kernels, strides, and output configurations for each layer in the feature extraction, emphasizing how each layer contributes to the final outputs.

TABLE I
DIFFERENT FEATURES EXTRACTED ON EEG BY BASIC TRANSFORMER MODELS

Module		Layer*	Input shaped	Kernel	Stride	Output
Branch I		TC	(ch, T)	(1,30)	(1,1)	(D_1, ch, T_1)
		SSC	(D_1, ch, T_1)	$(ch, 1)$	(1,1)	$(D_1, 1, T_1)$
		AP	$(D_1, 1, T_1))$	(1,120)	(1,12)	$(D_1, 1, T_2)$
		PWC	$(D_1, 1, T_2)$	(1,1)	(1,1)	$(D_2, 1, T_2)$
Branch II	1	TC	(ch, F, T)	(1,125)	(1,1)	(D_1, F, T_1)
		SFC	(D_1, F, T_1)	$(F, 1)$	(1,1)	$(D_1, 1, T_1)$
		AP	$(D_1, 1, T_1)$	(1,64)	(1,32)	$(D_1, 1, T_2)$
		PWC	$(D_1, 1, T_2)$	(1,1)	(1,1)	$(D_2, 1, T_2)$
	2	TC	(F, ch, T)	(1,125)	(1,1)	(D_1, ch, T_1)
		SSC	(D_1, ch, T_1)	$(ch, 1)$	(1,1)	$(D_1, 1, T_1)$
		AP	$(D_1, 1, T_1)$	(1,64)	(1,32)	$(D_1, 1, T_2)$
		PWC	$(D_1, 1, T_2)$	(1,1)	(1,1)	$(D_2, 1, T_2)$

* TC: Time Convolution, SSC: Separate Spatial Convolution, AP: Average Pooling, PWC: Point-wise Convolution, SFC: Separate Frequency Convolution

Specifically, from the model structure, it is apparent that the final feature size outputted by each branch is primarily governed by the kernel size and stride of the Average Pooling layer. For Branch I, a relative small convolution kernel is set to capture more granular features along the temporal dimension. However, despite richer details can be extracted, it may result in a larger feature map size. Using a larger Pooling Kernel size helps control the map size and the receptive field of the features. Meanwhile, it helps to reduce the computational requirements and enhance the model's generalization capabilities while maintaining substantial contextual information. In contrast, a larger convolution kernel set in Branch II aims to capture broader features along the time-frequency dimension, and a following smaller Pooling Kernel Size may facilitate more intensive feature extraction. Indeed, balancing the convolution kernel sizes and pooling parameters between different branches enhances the model's flexibility, which helps to better adapt to the model's intrinsic structure and allow the model to learn features of different scales from different data types, thus improving model performance. Here, to balance the features obtained while enhancing the model performance, we set a larger Pooling Kernel size P_1 of 120 with a stride of $P_1/10$ for Branch I, and a smaller Pooling Kernel size P_2 of 64 with a stride of $P_2/2$ for Branch II.

Moreover, the Transformer Encoder was configured with 4 blocks, and the multi-head attention mechanism was set with 10 heads. Finally, the model's performance was evaluated

using classification accuracy and the Kappa value, and the Kappa value is defined as:

$$Kappa = \frac{P_o - P_e}{1 - P_e} \tag{19}$$

where P_o is the proportion of correctly classified samples to the total number of samples, i.e., overall classification accuracy, and P_e represents the probability of chance agreement, i.e., the correctness of random guesses.

Besides, we also used the Wilcoxon Signed-Rank Test to analyze the potential statistical significance.

IV. Results

In this section, we compared the relevant results of proposed model against a variety of innovative state-of-the-art methods, where ablation experiments also being involved to demonstrate the contribution of each model part. Finally, we illustrated the interpretability of the extracted features visually, which helps in understanding underlying principles of the model. In general, for the selected datasets, the classification effect of existing deep learning has made great progress compared to machine learning. To avoid redundancy and ensure the persua¬siveness of the comparison, we mainly review the latest methods of the relevant datasets in the past three years, as well as some well-established deep learning techniques (e.g., ConvNet [12], EEGNet [13], FBCNet [36], EEG Conformer [16]). Among used methods, DRDA [37] offers a sophisticated end-to-end domain adaptation approach tailored for EEG-based motor imagery classification tasks, and DAFS [38] merges small sample learning with domain adaptation, enhancing domain-specific classification efficacy in MI-EEG tasks by leveraging source domain insights. EEG-ITNet [39] features an interpretable CNN framework that relies on inception modules and dilated causal convolutions, whereas IFNet [40] is a streamlined interactive convolutional network focusing on the interplay among various frequency signals to boost EEG feature depiction. MANN in [41] integrates multiple attention mechanisms with transfer learning for EEG classification, and incorporates domain adaptation techniques to enhance its efficacy, while a multi-scale hybrid convolutional network of MSHCNN [42] leverages convolu-tions across different dimen¬sions to distinctly extract temporal and spatial features from EEG data. In addition, several other latest models such as TSFCNet [26], Speech2EEG [23], EISATC-Fusion [27], FSA- TSP [43], FTCN [46] have also been used for the evaluation.

A. Head-to-head Comparison Results

Table II lists the comparison results of different algorithms applied in Dataset I. Specifically, the proposed Dual-TSST outperformed the existing SOTA methods in terms of overall average classification accuracy and the Kappa metrics, notably across subjects S1, S5, S6, and S7. In particular, compared with classical EEG decoding techniques like ConvNet and EEGNet, the average accuracy under current model has improved by 8.14% ($p<0.05$) and 6.17% ($p<0.05$), respectively, also with an obviously corresponding rise in Kappa values. Such the results underscore Dual-TSST's enhanced capability for global feature extraction, as opposed to those local feature focus seen in ConvNet and EEGNet. Moreover, for the most test subjects, Dual-TSST achieved superior results over the FBCSP-inspired FBCNet and domain adaptation methods like DRDA and DAFS ($p<0.05$), although being slightly inferior on S2 and S4. Compared to the models that introduced attention mechanisms into deep learning networks, such as Conformer, ADFCNN, and M-FANet, the developed Dual-TSST also showed better performance in most subjects' accuracy, the average classifi¬cation accuracy, and the Kappa value. Among all the compared methods, the SHNN model excels in subjects S8 and S9, while being less accurate in S6. Overall, our proposed dual-TSST framework delivers varied improvements in the classification accuracy among different subjects within the Dataset I, and leads in terms of the average accuracy, and the Kappa metrics.

For the binary classification Dataset II, as we see from Table III, several additional models have been supplemented for the evaluation. Consequently, near the similar effects have been observed with those in Dataset I, where the dual-TSST not only surpasses conventional deep learning models such as ConvNet and EEGNet but significantly outperformed other advanced methods like DRDA, SHNN, Conformer, and ADFCNN, in almost all metrics ($p<0.05$). In head-to-head comparisons with other leading techniques of MANN, TSFC-Net, MSHCNN, and EISATC-Fusion, Dual-TSST consistently achieved superior average accuracy and Kappa values. Furthermore, the standard deviation values of the proposed dual-TSST

TABLE II
Comparison Results Of Different Methods On Dataset I [Avg Acc: The Average Accuracy(%)

Year	Methods	S1	S2	S3	S4	S5	S6	S7	S8	S9	Avg Acc	Std	Kappa
2017	ConvNet [12]	76.39	55.21	89.24	74.65	56.94	54.17	92.71	77.08	76.39	72.53	13.43	0.6337
2018	EEGNet [13]	85.76	61.46	88.54	67.01	55.90	52.08	89.58	83.33	86.87	74.50	14.36	0.66
2021	FBCNet [36]	85.42	60.42	90.63	76.39	74.31	53.82	84.38	79.51	80.90	76.20	11.28	0.6827
2021	DRDA [37]	83.19	55.14	87.43	75.28	62.29	57.15	86.18	83.61	82.00	74.74	12.22	0.6632
2022	SHNN [29]	82.76	68.97	79.31	65.52	58.62	48.28	86.21	**89.66**	**89.87**	74.26	13.93	0.6648
2022	DAFS [38]	81.94	64.58	88.89	73.61	70.49	56.60	85.42	79.51	81.60	75.85	9.86	0.678
2022	EEG-ITNet [39]	84.38	62.85	89.93	69.10	74.31	57.64	88.54	83.68	80.21	76.74	10.82	-
2023	IFNet [40]	88.47	56.35	91.77	73.78	69.72	60.42	89.24	85.42	88.72	78.21	12.73	-
2023	Conformer [16]	88.19	61.46	93.40	**78.13**	52.08	65.28	92.36	88.19	88.89	78.66	14.42	0.7155
2024	ADFCNN [17]	87.15	61.45	**93.75**	75.69	75.34	65.27	88.54	82.29	85.06	79.39	10.23	-
2024	M-FANet [28]	86.81	**75.00**	91.67	73.61	76.39	61.46	85.76	75.69	87.15	79.28	**8.84**	0.7259
2024	**Dual-TSST**	**91.32**	59.38	93.40	69.44	**77.79**	**68.75**	**94.44**	85.76	85.76	**80.67**	11.76	0.7413

TABLE III
COMPARISON RESULTS OF DIFFERENT METHODS ON DATASET II [AVG ACC: THE AVERAGE ACCURACY(%)]

Year	Methods	S1	S2	S3	S4	S5	S6	S7	S8	S9	Avg Acc	Std	Kappa
2017	ConvNet [12]	76.56	50.00	51.56	96.88	93.13	85.31	83.75	91.56	85.62	79.37	16.27	0.5874
2018	EEGNet [13]	75.94	57.64	58.43	98.13	81.25	88.75	84.06	93.44	89.69	80.48	13.63	0.6096
2021	DRDA [37]	81.37	62.86	63.63	95.94	93.56	88.19	85.00	95.25	90.00	83.98	11.94	0.6796
2022	MANN [41]	82.81	60.36	59.06	97.50	91.88	86.38	84.06	93.44	86.88	82.54	12.95	0.6510
2022	SHNN [29]	83.33	61.76	58.33	97.30	91.89	88.89	86.11	92.11	91.67	83.49	13.10	0.6697
2023	Conformer [16]	82.50	65.71	63.75	**98.44**	86.56	90.31	87.81	94.38	**92.19**	84.63	11.49	0.6926
2023	TSFCNet [26]	76.25	70.00	83.75	97.50	72.81	86.56	88.44	92.50	89.69	86.39	8.81	0.7234
2023	MSHCNN [42]	**86.80**	**77.94**	65.97	97.97	93.24	88.88	86.80	82.89	86.80	85.25	8.67	-
2023	Speech2EEG [23]	80.70	62.04	71.74	96.09	94.51	84.06	84.06	**95.65**	87.76	84.07	10.80	-
2024	ADFCNN [17]	79.37	72.50	82.81	96.25	**99.37**	84.68	93.43	95.31	86.56	87.81	**8.39**	-
2024	EISATC-Fusion [27]	75.00	72.86	**86.56**	96.88	97.81	84.38	**94.06**	93.75	86.88	87.58	8.54	0.7515
2024	**Dual-TSST**	85.63	66.20	84.06	98.13	98.44	**90.94**	89.06	93.13	92.19	88.64	9.17	**0.7718**

TABLE IV
COMPARISON RESULTS OF DIFFERENT METHODS ON DATASET III

Year	Methods	Avg Acc	Std	Kappa
-	SVM	80.80	12.87	-
2021	BiHDM [44]	93.12	6.06	-
2022	RGNN [45]	79.37	10.54	-
2022	EeT [47]	96.28	4.39	-
2023	FSA-TSP [43]	93.55	5.03	-
2023	Conformer [16]	95.30	-	0.9295
2024	FTCN [46]	89.13	4.49	-
2024	**Dual-TSST**	**96.65**	**1.93**	**0.9488**

in Table III is 9.17, which is relatively lower to most of the compared methods. Such the result underscores the model's robust generalization ability to deliver steady strong results for diverse subjects.

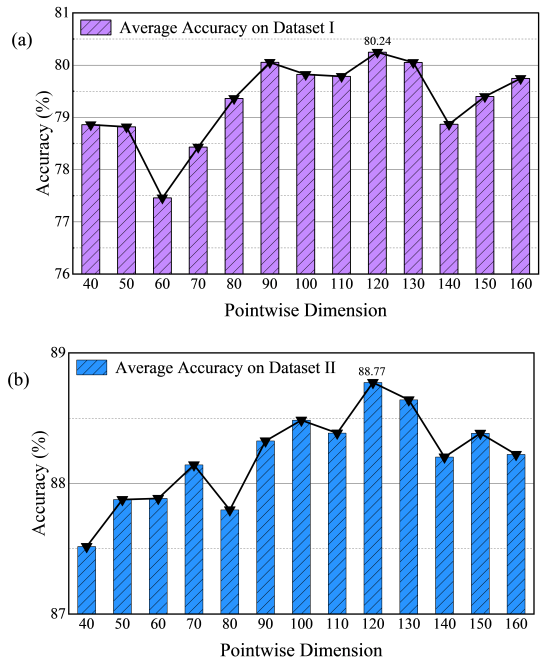


Fig. 4 The influence of pointwise dimension on model performance.

To further evaluate the robustness and generalization ability of the model, we extended our analysis with the challenging emotion Dataset III of SEED, which presents a different type of task and requires the model to adapt to new patterns. As listed in Table IV, the model continues to outperform the traditional machine learning algorithms and majority of the compared SOTA methods, indicating a commendable level of adapta¬bility of the designed model to effectively capture and interpret complex patterns associated with widely used EEG paradigms.

B. Parameter Sensitivity

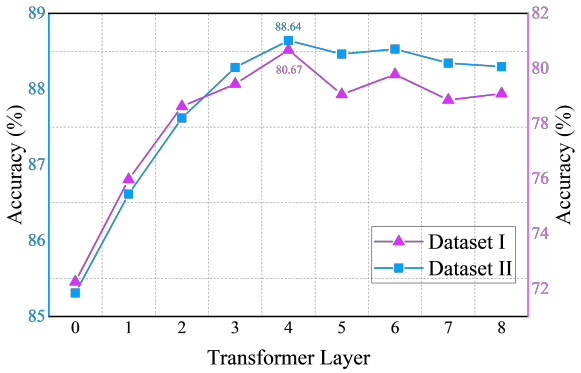


Fig. 5 The influence of the Transformer layer number on the average accuracy

Obviously, for DL models, the internal hyper-parameters of the network significantly affect its performance. The critical hyperparameters of our constructed model mainly include the dimensionality used for channel fusion and upscaling through pointwise convolution, the number of Transformer encoder layers and Transformer Heads.

First, here the pointwise dimension refers to the dimension parameter D_2, which is used in Dual-TSST for feature fusion and dimensionality increase through the pointwise convolution. To study the effects of this dimensional parameter, a range of [40, 160] with an interval of 10 has been designated, and Fig. 4 gives the resultant average accuracy. As shown in Fig. 4, with an increase in D_2, the accuracy in Dataset I shows an overall trend of decreasing first, then rising, and finally maintaining a mild fluctuation. Similarly, the average accuracy corresponding to Dataset II initially increases with D_2 and then oscillates. Interestingly, the optimal dimensional parameter for both is found at D_2 = 120, which avoids the complexity associated with too high dimensions and effectively enhances the model expressive capabilities.

The number of Transformer layers refers to the stack levels of Transformer Encoders, which essentially define the depth of the model that determines the complexity and hierarchy

of the information the model can learn. Generally, the deeper models typically enhance the model's representational ability and fit the data better. However, as the number of layers increases, issues such as the overfitting and gradient explosion may occur, along with an increase in computational costs. The accuracy trends with changes in number of Transformer layers are illustrated in Fig 5, where we see that the introduction of the Transformer (from zero to one layer) may lead to a marked performance improvement. Besides, while initial increases in layers (i.e., from 0 to 4) enhance performance for both datasets, the accuracy associated with all datasets begins to decline after the fourth layer. This may indicate that the model has reached its learning saturation or is beginning to overfit the noise within the data. For Dataset I, the peak accuracy with Dual-TSST exceeded the lowest by 4.36% ($p<0.01$), and for Dataset II, the corresponding value is 2.03% ($p<0.05$). These results suggest that while increasing the number of layers can enhance performance up to a certain limit, excessively high numbers may hinder training and increase the risk of overfitting.

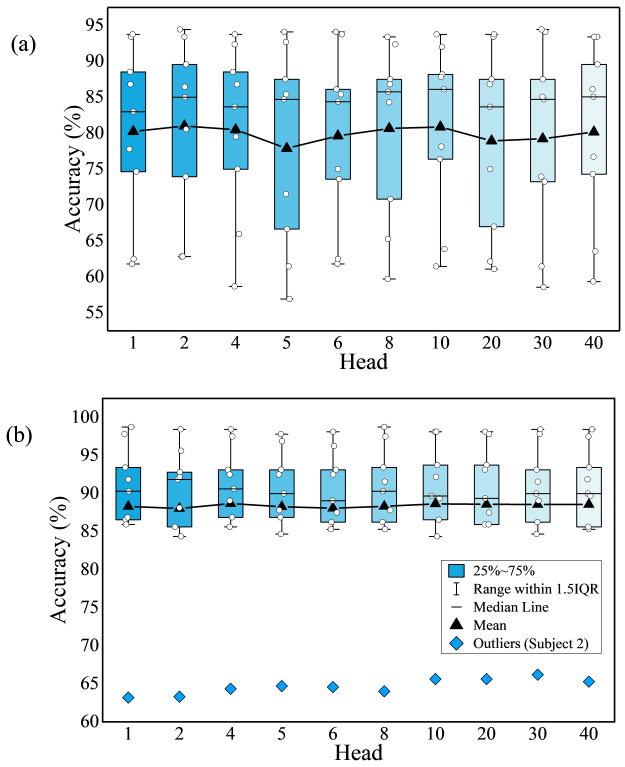


Fig. 6 The influence of the number of Multi-head attention heads on the accuracy for different datasets of (a) Dataset I, (b) Dataset II.

In the Transformer model, each involved Head can be seen as an independent self-attention mechanism, while multi-head attention allows the model to concurrently attend to different semantic information, thereby capturing diverse relationships and features in the input sequence. More specifically, each head learns different weights to better encode information in various contexts, and thus enhancing the richness and expressive power of the representation. However, too many heads can also lead to overfitting or an increase in computational complexity. In this study, the influence of the heads number has been studied, for which the results are depicted in Fig. 6. Noting that since the accuracy of Subject 2 of Dataset II is far from others, it is listed separately.

As illustrated in Fig. 6, the average accuracy on Dataset I varies significantly with the number of heads, while on Dataset II, the fluctuation seems to be smaller. Overall, the highest accuracy for both Dataset I and II is achieved when the Head count was 10, showing an improvement to the lowest accuracy of 2.97% ($p>0.05$) and 0.61% ($p>0.05$), respectively. Since the increase is not substantial, we conclude that changes in the number of Transformer heads do not significantly impact the model performance.

C. Ablation Experiment

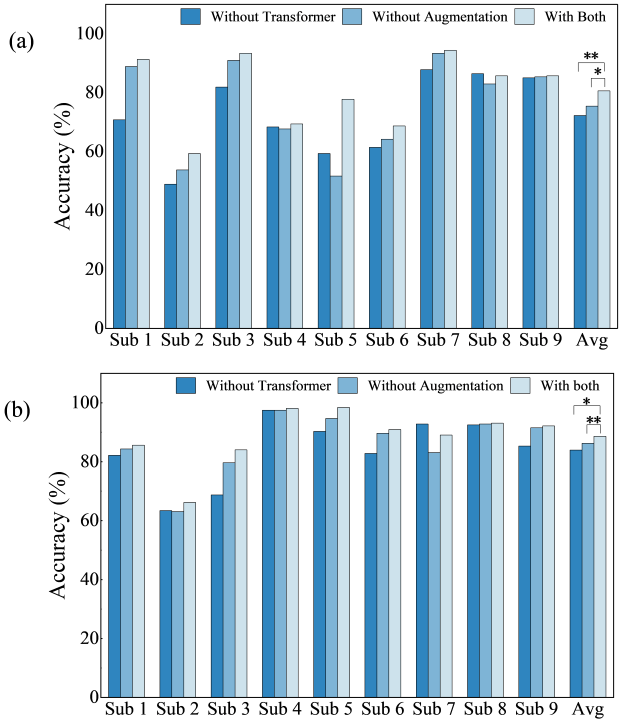


Fig. 7 Ablation experiments on data augmentation and Transformer application on datasets of (a) Dataset I, (b) Dataset II.

The Dual-TSST model comprises multiple modules, and we introduced the data augmentation measures into the proposed framework. To determine the specific effects of each functional modules, ablation experiments were conducted on both Dataset I and Dataset II to assess the impact of data augmentation, the Transformer module, different branches, and various inputs.

Initially, we conducted ablation experiments on the data augmentation and Transformer modules. As illustrated in Fig. 7, when it is without the Transformer module, we note an obvious decrease in the accuracy across most of the specific different subjects and the average results for the used datasets. However, an increase in performance is observed for Subject 7 of Dataset II, which possibly indicating overfitting when the module was used. Overall, for the tested two datasets, reintegrating the Transformer improved the overall average accuracy by 8.41% ($p<0.01$) and 4.68% ($p<0.05$), respectively, underscoring its critical role in boosting accuracy.

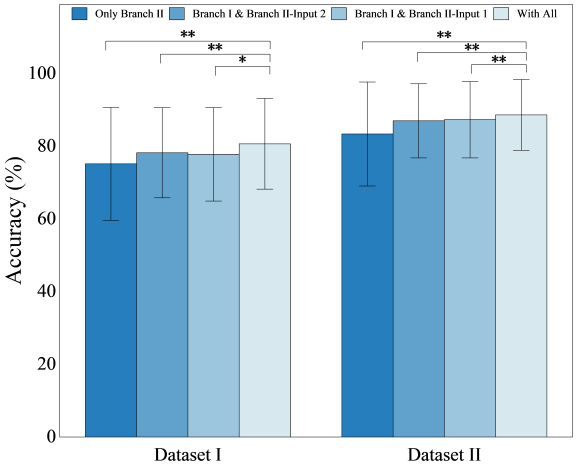


Fig. 8 Ablation experiments on different branches and inputs on both datasets.

The operation of data augmentation is envisioned to expand the data scale, aid the model in capturing more complex patterns, and mitigate overfitting tendencies. Meanwhile, it also introduces additional variability and disturbances. Across the datasets I and II, the application of data augmentation strategies led to a 5.21% ($p<0.05$) and 2.37% ($p<0.01$) increase in average accuracy, which indicates that such the module has proven to enhance model performance significantly.

We further conducted experiments by removing Branch I and Branch II (with Input 1 or Input 2), where the results of the remaining parts are reported in Fig. 8. It was observed that removing Branch I (i.e., only Branch II) significantly impacted the overall performance on both datasets ($p<0.05$), because Branch I provides the majority of the temporal features to the model. Besides, removing the input from Branch II had some impact, but not as significant as that from Branch I. Overall, on both datasets, performance using two branches was superior to using just one. Within Branch II, using two inputs also showed an improvement over using just one input. In addition, the improved error bar range of the model with all branches implies the enhanced robustness.

D. Visualization

To further intuitively demonstrate the effectiveness of the designed branches and self-attention mechanism, a comparative study of low-dimensional visualizations, using t-SNE [48], was conducted for one typical subject (i.e., Subject 7 of Dataset I). Fig. 9 reports the relevant results with/without prominent components (e.g., Branch I or Branch II of feature extraction part, Transformer modules). Specifically, for the test data, as in Fig. 9 (a), when it with Branch II only, the features of focused categories are closely mixed. In contrast, as shown in (b) and (c), the distance between classes becomes larger with the help of Branch I, even if only a part of Branch II is involved. The results of inter-category distance is being more evident with all developed branches, thus illustrating the capacity of our model.

Moreover, as we can see from Fig. 9 (e), without the Transformer, the t-SNE visualization of the training set reveals several well-separated clusters, indicating a clear distribution of categories in low-dimensional space. However, when applied to the testing set, the model exhibits a significant reduction in category separation (see Fig. 9 (f)). In particular, a substantial overlap of features between the feet and tongue, left- and right-hands can be apparently observed. This overlap indicates that while efficiently learning the properties of each category on the training data, the model exhibits poor generalization when exposed to unknown data, failing to discriminate between comparable classes. Conversely, the introduction of Transfor¬mer results in a dramatic improvement. As in t-SNE visualiza¬tion, the training set displays highly distinct and well-separated clusters, with each category occupying a clear, even non-overlapping region in low-dimensional space. This implies that the transformer module significantly improves the model's capacity to describe diverse properties, resulting in a better defined distribution of categories. Importantly, the t-SNE visualization of the test set also exhibits considerable improvement, where the distinctions between hand features and other categories become more prominent. Especially in categories prone to confusion (such as left and right hand), the Transformer module significantly lowers overlap, highlighting its vital role in strengthening the model's generalization performance and capacity to differentiate between comparable categories.

To further exhibit the impact of the integrated Transformer modules, graphical confusion matrix was used to present the classification performance across the specific categories. For each dataset, the results of one subject with/without related part are depicted in Fig. 10. As it can be seen, the results clearly demonstrate that the model without the Transformer module

faces considerable challenges in capturing the discriminative features. For instance, the confusion matrix reveals that 23.61% of left-hand features were erroneously classified as right-hand features, whereas a notable 19.44% misclassification rate of tongue features being recognized as feet (see in Fig. 10 (a)). Such results suggest that the model struggles with discerning subtle feature differences, which may lead to a generalization inadequacy, particularly when dealing with categories that exhibit similar features. Instead, upon incorporating the Trans¬former, a marked improvement in the model's classification capabilities is observed (mere the value of 5.56% and 6.94% of corresponding index are found). Also for hand imaginary recognition of Fig. 10 (c) and (d), following the implementation of the Transformer, the model for subject 4 of dataset II has a particularly satisfactory classification accuracy of 98.75%, a notable increase of near 5%. These improvements suggests that the Transformer module bolsters the model's feature extraction capabilities and enhances its generalization ability and robust¬ness across different datasets.

V. Disscussion and Conclusion

The statistical distribution of non-stationary EEG data varies across different subjects and recording sessions, making it challenging for BCI researchers to design a classifier with high accuracy and generalization capability. Borrowing the idea of machine learning with data flow of feature extraction, feature fusion, and classification, a novel efficient DL-framework that fully incorporates the CNN and Transformer is suggested in this study to handle the data processing of EEG signals.

Fig. 9 The comparison of the features for Subject 7 of Dataset I of with/without the Transformer module by t-SNE visualization.

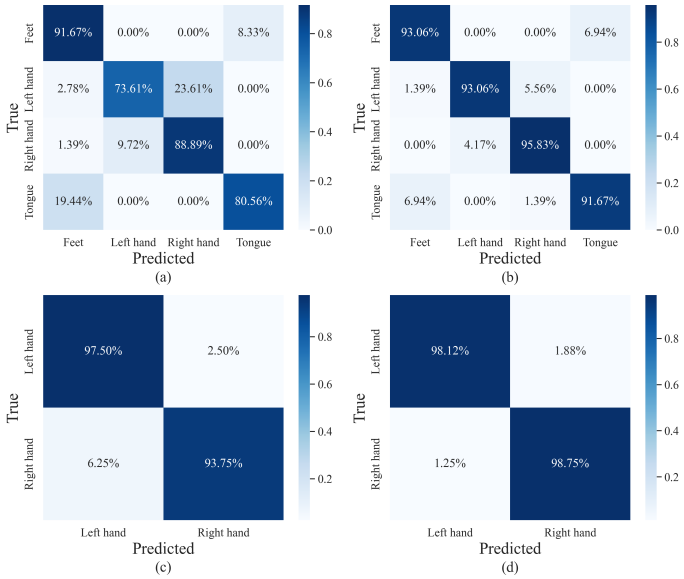


Fig. 10 Confusion matrices for (a) Subject 7 of Dataset I without Transformer, (b) Subject 7 of Dataset I with Transformer, (c) Subject 4 of Dataset II without Transformer, (d) Subject 4 of Dataset II with Transformer.

The proposed dual-TSST model first leverages a dual-branch CNN structure, which accepts data from diverse perspectives, to dig the comprehensive representation of entailed features. In this configuration, Branch I is tasked with extracting spatio¬temporal features from raw EEG, while Branch II handles the spatio-temporal-frequency features from wavelet-transformed data. The Transformer module further explores the long-range global dependencies and synthesizes all the diverse features into a cohesive feature set, which is finally classified by the classification module. For the proposed dual-TSST framework, minimal yet critical preprocessing with band-pass filtering, as in [16], [26], is needed to EEG signals, which avoids the specific sophisticated data preprocessing steps. In essence, such the proposed DL model does not require the extra expert knowledge but can automatically extract the comprehensive spatio-temporal-frequency features, which is conducive to the identification from multifaceted data sources.

Experimentally, the framework was evaluated through two BCI Competition IV Datasets of 2a, 2b, and one widely used emotional Dataset of SEED, where the superior performance compared with state-of-the-art methods has been achieved. In general, extensive parameter sensitivity and ablation studies affirm that each component significantly contributes to the model's effectiveness, particularly highlighting the substantial impact of pointwise dimension and Transformer. Particularly, the number of Transformer layers, which also being termed as the depth in other related studies, directly influences the model's classification result, highlighting the importance of such the module introduction. More importantly, our specific results of Fig. 5 reveal the instructive suggestion for subsequent layer configuration of future transformer-based EEG decoding. Conversely, the number of heads in the multi-head attention setup showed marginally impact on the final performance, and such insensitivity may aid in the lightweight iterative design of future models. Moreover, the effects of the related branches and data augmentation have also been intuitively presented, thus clarifying the rationality of the developed framework.

Whereas the approach proposed in this study boosts the

model capability to extract more discriminative EEG features, it still has several limitations. First, a notable limitation of our current model is its structural complexity. The majority of the parameters in current model originates from the comprehensive Transformer module and the fully connected layers for the classification, which coincides with the prior research of EEG Conformer [16]. Although depthwise separable convolutions were applied to mitigate this issue, current model still maintains a higher parameter count. Second, as the deep learning model, the number of samples for the focused data is expected to be large enough. While the data augmentation strategies can be strategically adopted as current work, one should maintain the quest for more effective source data [40]. Since it is expensive, not-friendly, and impractical to always collect a larger number of recording data, several advanced methods, such as the transfer learning based domain adaptation, which uses knowledge from source subjects to improve the performance of a targeted one [38], ought to be applied toward all accessible data. Moreover, only the subject-specific based experiments are conducted in this study, while more cross-subjects validation should be focused to further investigate the generalizability of the model [17], [26].

To sum up, the developed innovative architecture leverages the distinct strengths of related data types to enhance the accuracy and robustness of the decoding process, while also improving network interpretability through obeying ML-based processing flow. Moving forward, our objectives will focus on optimizing the model's architecture and reducing its parameter footprint, alongside exploring online potential applications.

REFERENCES

[1] J. R. Wolpaw *et al.*, "Brain-computer interface technology: a review of the first international meeting," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 8, no. 2, pp. 164-173, Jun. 2000.

[2] H. Li, L. Bi, X. Li, and H. Gan, "Robust predictive control for EEG-based brain–robot teleoperation," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 8, pp. 9130-9140, Aug. 2024.

[3] H. Li, L. Bi, and J. Yi, "Sliding-mode nonlinear predictive control of brain-controlled mobile robots," *IEEE Trans. Cybern.*, vol. 52, no. 6, pp. 5419-5431, Jun. 2022.

[4] H. Li, L. Bi, and H. Shi, "Modeling of human operator behavior for brain-actuated mobile robots steering," *IEEE Trans. Neural. Syst. Rehabil. Eng.*, vol. 28, no. 9, pp. 2063-2072, Sep. 2020.

[5] R. Abiri, S. Borhani, E. W. Sellers, Y. Jiang, and X. Zhao, "A comprehensive review of EEG-based brain–computer interface paradigms," *J. Neural Eng.*, vol. 16, no. 1, Feb. 2019, Art. no. 011001.

[6] S. Aggarwal and N. Chugh, "Review of machine learning techniques for EEG based brain computer interface," *Arch. Comput. Method Eng.*, vol. 29, no. 5, pp. 3001-3020, Aug. 2022.

[7] S. Gong, K. Xing, A. Cichocki, and J. Li, "Deep learning in EEG: advance of the last ten-year critical period," *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 2, pp. 348-365, Jun. 2022.

[8] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, May. 2015.

[9] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE Trans Neural Netw. Learn. Syst.*, vol. 33, np. 12, pp. 6999-7019, Dec. 2022.

[10] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Comput.*, vol. 29, no. 9, pp. 2352-2449, Sep. 2017.

[11] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE Trans Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 604-624, Feb. 2021.

[12] R. T. Schirrmeister *et al.*, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human. Brain Mapp.*, vol. 38, no. 11, pp. 5391-5420, Aug. 2017.

[13] V. J Lawhern *et al.*, "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, Jul. 2018, Art. no. 056013.

[14] [S. Tortora *et al.*, "Deep learning-based BCI for gait decoding from EEG with LSTM recurrent neural network," *J. Neural Eng.*, vol. 17, no. 4, Jul. 2020, Art. no. 046011.

[15] J. Sun, J. Xie, and H. Zhou, "EEG classification with transformer-based models," in *Proc. IEEE 3rd Glob. Conf. Life Sci. Technol., (LifeTech).*, pp. 92-93, 2021.

[16] Y. Song, Q. Zheng, B. Liu, and X. Gao, "EEG conformer: convolutional transformer for EEG decoding and visualization," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 710-719, Dec. 2023.

[17] W. Tao *et al.*, "ADFCNN: attention-based dual-scale fusion convolutional neural network for motor imagery brain–computer interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 32, pp. 154-165. 2024.

[18] A. Arjun, A. S. Rajpoot, and M. Raveendranatha Panicker, "Introducing attention mechanism for EEG signals: emotion recognition with vision transformers," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC).*, pp. 5723-5726, Nov. 2021.

[19] M. S. Al-Quraishi *et al*., "Decoding the user's movements preparation from EEG signals using vision transformer architecture," *IEEE Access*, vol. 10, pp. 109446-109459, Oct. 2022.

[20] M. A. Mulkey *et al.*, "Supervised deep learning with vision transformer predicts delirium using limited lead EEG," *Sci. Rep.*, vol. 13, no. 1, May. 2023, Art. no. 7890.

[21] A. Nogales *et al*., "BERT learns from electroencephalograms about Parkinson's disease: transformer-based models for aid diagnosis," *IEEE Access*, vol. 10, pp. 101672-101682, Jan. 2022.

[22] B. Wang, X. Fu, Y. Lan, L. Zhang, and Y. Xiang, "Large transformers are better EEG learners," *arXiv: 2308.11654*.

[23] J. Zhou, Y. Duan, Y. Zou, Y. -C. Chang, Y. -K. Wang, and C. -T. Lin, "Speech2EEG: leveraging pretrained speech model for EEG signal recognition," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 2140-2153, Apr. 2023.

[24] X. Tian *et al.*, "Deep multi-view feature learning for EEG-based epileptic seizure detection," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 10, pp. 1962-1972, Oct. 2019.

[25] R. Mane *et al.*, "A multi-view CNN with novel variance layer for motor imagery brain computer interface," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC).*, pp. 2950-2953, Jul. 2020.

[26] H. Zhi, Z. Yu, T. Yu, Z. Gu, and J. Yang, "A multi-domain convolutional neural network for EEG-based motor imagery decoding," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 3988-3998, Oct. 2023.

[27] G. Liang, D. Cao, J. Wang, Z. Zhang, and Y. Wu, "EISATC-fusion: inception self-attention temporal convolutional network fusion for motor imagery EEG decoding," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 32, pp. 1535-1545, Mar. 2024.

[28] Y. Qin, B. Yang, S. Ke, P. Liu, F. Rong, and X. Xia, "M-FANet: multi-feature attention convolutional neural network for motor imagery decoding," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 32, pp. 401-411, Jan. 2024.

[29] C. Liu *et al.*, "SincNet-based hybrid neural network for motor imagery EEG decoding," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 540-549, Mar. 2022.

[30] M. X. Cohen, "A better way to define and describe Morlet wavelets for time-frequency analysis," *NeuroImage*, vol. 199, pp. 81-86, Oct. 2019.

[31] F. Lotte, "Signal processing approaches to minimize or suppress calibra¬tion time in oscillatory activity-based brain–computer interfaces," *Proc IEEE*, vol. 103, no. 6, pp. 871-890, Jun. 2015.

[32] M. Tangermann *et al.*, "Review of the BCI competition IV," *Front. Neurosci.*, vol. 6, p.55, Jul. 2012.

[33] V. Jayaram and A. Barachant, "MOABB: Trustworthy algorithm bench¬marking for BCIs," *J. Neural Eng.*, vol. 15, no. 6, Dec. 2018, Art. no. 066011.

[34] W. L. Zheng and B. L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auton. Mental Develop.*, vol. 7, no. 3, pp. 162-175, Sep. 2015.

[35] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.

[36] R. Mane, *et al*. "FBCNet: A multi-view convolutional neural network for brain-computer interface," 2021, *arXiv:2104.01233*.

[37] H. Zhao, Q. Zheng, K. Ma, H. Li, and Y. Zheng, "Deep representation¬based domain adaptation for nonstationary EEG classification," *IEEE Trans Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 535-545, Feb. 2021.

[38] C. Phunruangsakao, D. Achanccaray, and M. Hayashibe, "Deep adversa¬rial domain adaptation with few-shot learning for motor-imagery brain¬computer interface," *IEEE Access*, vol. 10, pp. 57255-57265, Jan. 2022.
[39] A. Salami, J. Andreu-Perez, and H. Gillmeister, "EEG-ITNet: an explainable inception temporal convolutional network for motor imagery classification," *IEEE Access*, vol. 10, pp. 36672-36685, Apr. 2022.
[40] J. Wang, L. Yao, and Y. Wang, "IFNet: An interactive frequency convolutional neural network for enhancing motor imagery decoding from EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1900-1911, Jan. 2023.
[41] P. Chen, Z. Gao, M. Yin, J. Wu, K. Ma, and C. Grebogi, "Multiattention adaptation network for motor imagery recognition," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 52, no. 8, pp. 5127-5139, Aug. 2022.
[42] X. Tang, C. Yang, X. Sun, M. Zou, and H. Wang, "Motor imagery EEG decoding based on multi-scale hybrid networks and feature enhance¬ment," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1208-1218, Feb. 2023.
[43] M. Jiménez-Guarneros and G. Fuentes-Pineda, "Cross-subject EEG-based emotion recognition via semisupervised multisource joint distribu¬tion adaptation," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–12, 2023.
[44] Y. Li *et al.*, "A novel Bi-hemispheric discrepancy model for EEG emotion recognition," *IEEE Trans. Cogn. Develop. Syst.*, vol. 13, no. 2, pp. 354-367, 2021.
[45] P. Zhong, D. Wang and C. Miao, "EEG-based emotion recognition using regularized graph neural networks," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1290-1301, 2022.
[46] L. Yang *et al.*, "Electroencephalogram-based emotion recognition using factorization temporal separable convolution network," *Eng. Appl. Artif. Intell.*, vol. 133, 2024, Art. no. 108011.
[47] J. Liu *et al.*, "Spatial-temporal transformers for EEG emotion recognition," in *Proc. Int. Conf. Adv. Artif. Intell.*, 2022, pp. 116-120.
[48] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579-2605, 2008.

BrainDecoder: Style-Based Visual Decoding of EEG Signals

Minsuk Choi
Department of Computer Science and Engineering
Waseda University
Tokyo, Japan
minsuk@fuji.waseda.jp

Hiroshi Ishikawa
Department of Computer Science and Engineering
Waseda University
Tokyo, Japan
hfs@waseda.jp

***Abstract*—Decoding neural representations of visual stimuli from electroencephalography (EEG) offers valuable insights into brain activity and cognition. Recent advancements in deep learning have significantly enhanced the field of visual decoding of EEG, primarily focusing on reconstructing the semantic content of visual stimuli. In this paper, we present a novel visual decoding pipeline that, in addition to recovering the content, emphasizes the reconstruction of the style, such as color and texture, of images viewed by the subject. Unlike previous methods, this "style-based" approach learns in the CLIP spaces of image and text separately, facilitating a more nuanced extraction of information from EEG signals. We also use captions for text alignment simpler than previously employed, which we find work better. Both quantitative and qualitative evaluations show that our method better preserves the style of visual stimuli and extracts more fine-grained semantic information from neural signals. Notably, it achieves significant improvements in quantitative results and sets a new state-of-the-art on the popular Brain2Image dataset.**

***Index Terms*—Deep Learning, Image Synthesis, EEG, Multimodal**

I. Introduction

Understanding neural representations in the brain and the information they encode is crucial for enhancing our knowledge of cognitive processes and developing brain-computer interfaces (BCIs) [1]. In particular, decoding and simulating the human visual system has emerged as a significant challenge. Recent advancements have led to substantial progress in visual decoding, allowing for the reconstruction of visual stimuli perceived by a subject during brain activity measurement. [2] [3] [4] [5] [6] [7] [8] [9]

Electroencephalography (EEG) is a technique for recording brain signals, widely used due to its non-invasive nature, cost-effectiveness, and high temporal resolution. Although it has notable limitations [10] such as relatively lower spatial resolution as well as susceptibility to physiological artifacts and individual differences, conducting research based on EEG remains crucial for practical applications. The technique's accessibility and ability to capture real-time brain activity make it invaluable.

This work was partially supported by JSPS KAKENHI Grant Number JP20H00615.

Previous research [11] [12] [13] [14] on EEG-based visual decoding has primarily focused on capturing high-level semantic content by aligning with the text or image embedding space of CLIP (Contrastive Language–Image Pretraining) [15]. While these approaches have successfully represented broad semantic categories, they often fall short in accurately reproducing stylistic details such as color and texture, revealing a gap between semantic understanding and detailed visual representation.

In this paper, we present ***BrainDecoder***, a novel method that aims to overcome this limitation by aligning EEG signals with both image and text embeddings as separate conditions in a pretrained latent diffusion model [16]. In the text-to-image generation literature, previous researches [17] [18] have demonstrated that incorporating image "prompts" along with the text ones enable image generation that preserves style and content. By aligning EEG signals with both image and text embedding spaces, we show it is possible to extract both style and semantic information. This dual approach enhances the model's ability to more accurately reconstruct the stylistic features of the images viewed by the EEG subject. Our qualitative and quantitative evaluations demonstrate that BrainDecoder outperforms the state-of-the-art by a large margin in both reconstruction details and generation quality, setting a new benchmark for EEG-based visual decoding.

II. Methodology

We introduce a novel framework for reconstructing images viewed by an EEG subject, as illustrated in Fig. 1. It consists of three main components: A) Aligning EEG signals with CLIP image space, B) Aligning EEG signals with CLIP text space, and C) Combining the CLIP-aligned EEG representations for visual stimuli reconstruction.

A. EEG Alignment in Image Space

Prior work [17] [19] [20] has demonstrated the ability of CLIP image embeddings to facilitate both semantic and stylistic transfer when the generator model is conditioned accordingly. Building on these findings, our approach aims to extract image-related information from EEG signals by aligning them with CLIP image embeddings. To achieve this, we process the EEG signals and their corresponding ground

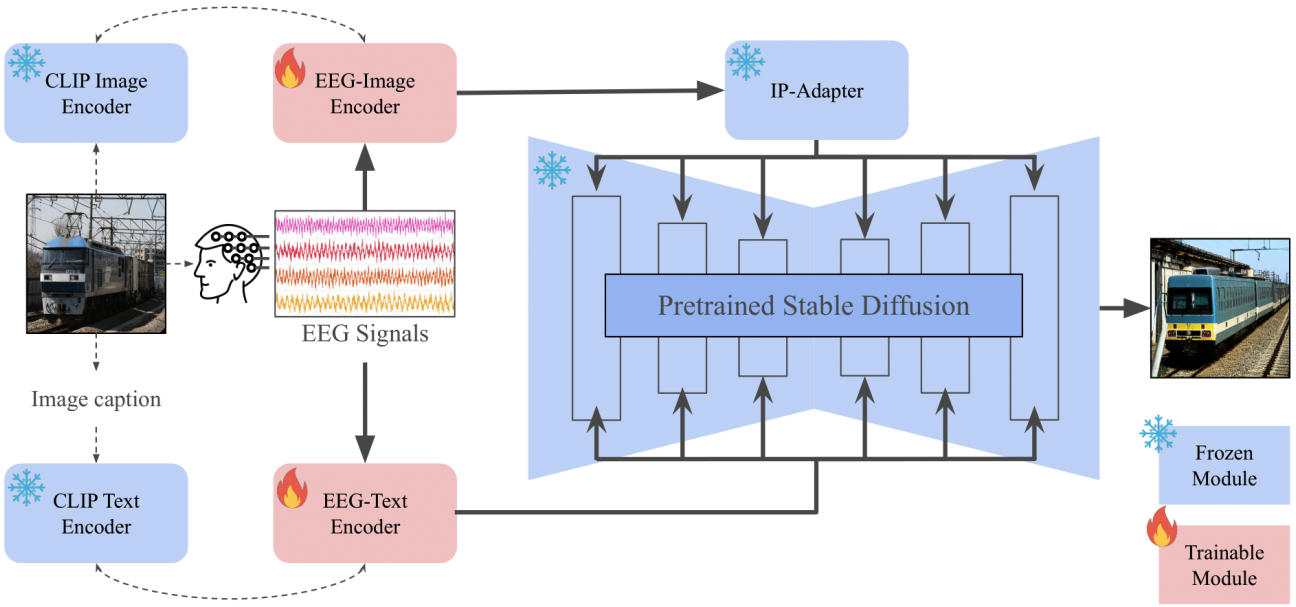


Fig. 1. The overall architecture of our proposed BrainDecoder framework. The modules in blue are frozen during training and only the modules in red are updated. The bold arrows are used during inference and the dotted lines are used during training.

truth images (i.e., the ones that the EEG subject was watching when the signal was taken) using an EEG encoder and a CLIP image encoder, respectively, and aim to correlate the outputs. Let F_{image} be an encoder that processes the input EEG signal x, and E_{image} the CLIP image encoder applied to the input image I. We call F_{image} the EEG-image encoder because it is trained to align with the image as encoded by E_{image}. We employ Mean Squared Error (MSE) as the loss function to measure the similarity between the EEG and image representations:

$$L_{\text{image}} = \text{MSE}(F_{\text{image}}(x) - E_{\text{image}}(I)). \tag{1}$$

To effectively encode the EEG data, we extend upon previous approaches [2] [21] [6] by utilizing an LSTM-based encoder architecture followed by fully connected layers.

B. EEG Alignment with Text Space

Recent approaches for visual brain signal decoding [11] [22] have sought to align brain signals with CLIP [15] text embeddings obtained from captions generated by pretrained image caption generators. However, since CLIP was trained on image-text pairs publicly available on the Internet with often short captions, those methods using longer generated captions, particularly with Stable Diffusion [16], have been less effective. Although Stable Diffusion allows up to 77 tokens as input, empirical evidence suggests that the effective token length of CLIP is considerably shorter [23]. Accordingly, we adopt a simpler labeling approach: we make the caption by appending the class label of the image to the text "an image of". We show empirically that this method improves performance over previous approaches and that more fine-grained information can be captured by the EEG-image encoder instead. We use the CLIP text encoder E_{text} that embeds the caption C to train the EEG-text encoder F_{text} that encodes the corresponding EEG signal x. As in the image alignment step, we use MSE as the loss function to quantify the similarity between the EEG and the text representations:

$$L_{\text{text}} = \text{MSE}(F_{\text{text}}(x) - E_{\text{text}}(C)). \tag{2}$$

Similar to the image processing pipeline, an LSTM-based encoder is used for EEG signal encoding.

C. Visual Stimuli Reconstruction

After training the EEG-image and EEG-text encoders, we leverage the resulting EEG embeddings to generate images. Our method uses a pretrained latent diffusion model (e.g., Stable Diffusion [16]), with the EEG embeddings from both encoders serving as distinct conditioning inputs. This is achieved through a decoupled cross-attention mechanism [17]. We hypothesize that by aligning the EEG signals in CLIP image space, the EEG encoder can capture detailed semantics and style that may not be easily conveyed through text alone. This approach is analogous to the way latent diffusion models incorporate both text and image prompts as conditioning factors. The reconstructed visual stimuli are defined as:

$$\hat{y} = \text{SD}(F_i(x), F_t(x)) \tag{3}$$

Here, $\hat{y}$ represents the reconstructed image, and SD denotes the pretrained Stable Diffusion, conditioned on the outputs of both the EEG-image encoder F_{image} and the EEG-text encoder F_{text}.

III. Experiments and Results

This section is divided into two main parts. We begin by detailing our experimental setup for training the EEG

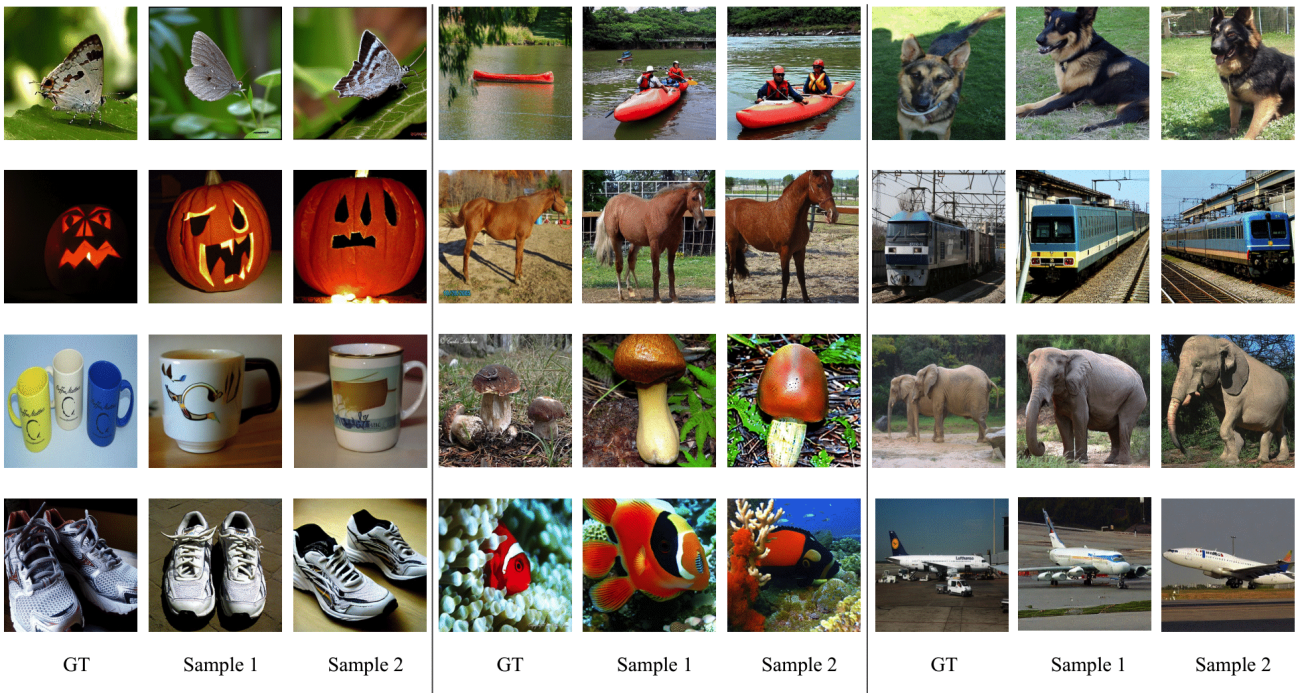


Fig. 2. Sample outputs. The images on the left show the ground truth visual stimuli shown during dataset collection. The following two images are sample outputs from our framework. Notably, the sample results show a high correspondence in semantics and style to the visual stimuli.

encoders. Following this, we present our findings and discuss various ablation studies.

A. Dataset

We utilize the Brain2Image [21] [2], an EEG-image pair dataset with 11,466 EEG recordings from six participants, for our experiments. These recordings were captured using a 128-channel EEG system as the participants were exposed to visual stimuli for 500 ms. The stimuli consisted of 2,000 images with labels spanning 40 categories, derived from the ImageNet dataset [24]. Each category included 50 easily recognizable images to ensure clarity in the participants' neural responses.

B. Implementation

For the EEG encoders, we extend from previous approaches [21] [2] [6] and use a 3-layered LSTM network with a hidden dimension of 512. The output of the network is then passed through a fully connected linear network with a BatchNorm [25] and LeakyReLU [26] activation function in between. Only the EEG encoders are trained in our framework, keeping the framework computationally efficient. We use the Adam [27] optimizer with a weight decay of 0.0001. The initial learning rate is set to 0.0003 and a lambda learning rate scheduler is used with a lambda factor of 0.999.

In order to align with CLIP image space, we follow the approach outlined in the IP-Adapter [17] framework, utilizing the CLIP-Huge model to process the images. For aligning EEG with CLIP text space, we process the captions using the CLIP-Large model which is used by Stable Diffusion 1.5. The captions are generated by concatenating "an image of" with the class label. For ablation studies, we employ the LLaVA-1.5-7b model [28] for layout-oriented caption generation and BLIP [29] for general caption generation. For visual reconstruction, we employ Stable Diffusion version 1.5, aligning our method with recent results for fair comparison and we employ a PNDM scheduler [30] with 25 inference steps.

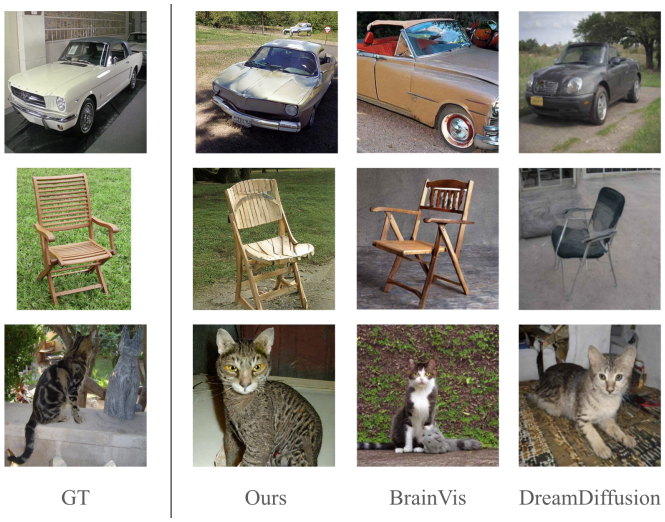


Fig. 3. Comparison of output images with the ground truth and outputs from other methods.

C. Evaluation Metrics

We employ the following metrics to objectively assess the performance of our framework. **ACC**: The N-way Top-

TABLE I
QUANTITATIVE RESULTS

Methods	*ACC* ↑	*IS* ↑	*FID* ↓	*SSIM* ↑	*CS* ↑
Brain2Image	-	5.07	-	-	-
DreamDiffusion*	45.8	-	-	-	-
BrainVis	45.5	-	-	-	0.602
EEGStyleGAN-ADA	-	10.82	174.13	-	-
Ours	**95.2**	**28.11**	**69.97**	**0.2277**	**0.7575**

*Results from DreamDiffusion were computed using data from subject 4

TABLE II
ABLATION STUDY RESULTS

Methods	*ACC* ↑	*IS* ↑	*FID* ↓	*SSIM* ↑	*CS* ↑
Only Text (LLaVA)	60.43	20.22	151.98	0.1797	0.6188
Only Text (BLIP)	68.91	24.0	127.19	0.1832	0.6541
Only Text (label)	72.61	26.43	105.6	0.1845	0.6610
Only Image	79.7	26.1	75.88	0.2239	0.7177
Original	**95.2**	**28.11**	**69.97**	**0.2277**	**0.7575**

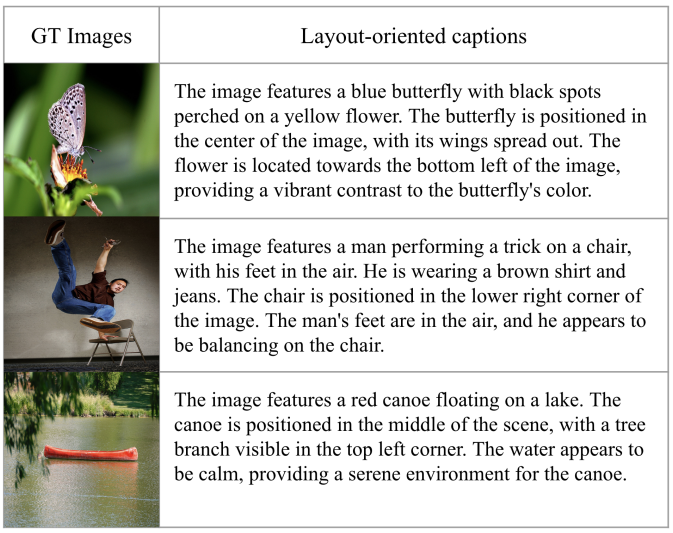


Fig. 4. Example layout-oriented captions generated with LLaVA.

K Classification Accuracy [8] [31] evaluates the semantic accuracy of the reconstructed images. We set $N = 50$ and $K = 1$. **IS**: The Inception Score [32] assesses the diversity and quality of the generated images. **FID**: Fréchet inception distance [33] measures the distance from the ground truth images. **SSIM**: The Structural Similarity Index Measure [34] evaluates the quality of images. **CS**: CLIP Similarity [35] [11] reflects how well the generated images capture the semantic and stylistic content of the ground truth images.

D. Results

Fig. 2 presents sample outputs of BrainDecoder. Beyond capturing the high-level semantics, our method demonstrates the ability to retain fine-grained visual features, including color and texture. Notably, there is also a resemblance in the color composition of the background in addition to the main object's color. This capability is further illustrated in the example of the electric locomotive class. The object's color is depicted as light blue—matching the visual stimuli—despite the range of potential color variations. This demonstrates the model's ability to recover nuanced visual attributes with a high fidelity.

This is further demonstrated in Fig. 3, where we compare our results with prior studies. Notably, in the second image, our method is able to reconstruct not only the wooden texture of the chair, but the grass in the background as well, which was absent in the results by other methods.

Table. I shows the quantitative results of BrainDecoder compared to baselines [2] [12] [11] [7]. We evaluate our methodology on 5 evaluation metrics in §III-C. BrainDecoder outperforms the state-of-the-art in both reconstruction fidelity and generation quality. Notably, BrainDecoder achieves a surprising 95.2% on the 50-way top-1 classification accuracy metrics, showing that the trained EEG encoders are able to extract rich information from the brain signals very well.

E. Ablation

We conduct an ablation study to understand the contributions of each component. Rows 3-5 of Table II show visual decoding with the EEG-image encoder yields a higher SSIM (0.2239) than with only the EEG-text encoder (0.1845). This supports our premise that aligning in CLIP's image space facilitates style transfer. Furthermore, the framework achieves the best performance when both encoders are used, indicating the complementary nature of the two encoders.

Additionally, we empirically show that captions generated by Vision Language Models (VLMs) are suboptimal for EEG-based visual decoding. We compare our label caption method with two VLMs: BLIP [29] and layout-oriented LLaVA [28]. A key challenge in image reconstruction from brain signals is preserving the visual layout. We hypothesize that associating EEG signals with detailed layout-oriented CLIP text embeddings might help. Using the LLaVA model, we generate layout-oriented captions following the instruction: "Write a description of the image layout. EXAMPLE OUTPUT: [object] is in the top left of the image, facing right." Fig. 4 shows example layout-oriented captions. Notably, rows 1-3 of Table II indicate that the simple label caption ("an image of [class name]") performs best, while layout-oriented captions (row 1) perform the worst. This further supports our premise that simple label captions are more effective for EEG encoders and complex prompts are harder for CLIP to fully interpret.

IV. CONCLUSION

Our research introduces ***BrainDecoder***, a novel approach to image reconstruction from EEG signals that preserves both stylistic and semantic features of visual stimuli. By aligning EEG signals with CLIP image and text embeddings separately, we bridge the gap between neural representations and visual content. Our analysis demonstrates significant improvements over existing models, offering a richer interpretation of neural signals through the dual-alignment strategy.

REFERENCES

[1] M. Orban, M. Elsamanty, K. Guo, S. Zhang, and H. Yang, "A review of brain activity and eeg-based brain-computer interfaces for rehabilitation application," in *Bioengineering*, 2022.

[2] I. Kavasidis, S. Palazzo, C. Spampinato, D. Giordano, and M. Shah, "Brain2image: Converting brain signals into images," *Proceedings of the 25th ACM international conference on Multimedia*, 2017.

[3] P. Tirupattur, Y. S. Rawat, C. Spampinato, and M. Shah, "Thoughtviz: Visualizing human thoughts using generative adversarial network," in *Proceedings of the 26th ACM International Conference on Multimedia*, ser. MM '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 950–958.

[4] S. Khare, R. Choubey, L. Amar, and V. Udutalapalli, "Neurovision: perceived image regeneration using cprogan," *Neural Computing and Applications*, vol. 34, pp. 1–13, 04 2022.

[5] T. Fang, Y. Qi, and G. Pan, "Reconstructing perceptive images from brain activity by shape-semantic gan," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13 038–13 048, 2020.

[6] P. Singh, P. Pandey, K. Miyapuram, and S. Raman, "Eeg2image: image reconstruction from eeg brain signals," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[7] P. Singh, D. Dalal, G. Vashishtha, K. Miyapuram, and S. Raman, "Learning robust deep visual representations from eeg brain recordings," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7553–7562.

[8] Z. Chen, J. Qing, T. Xiang, W. L. Yue, and J. H. Zhou, "Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 710–22 720.

[9] Y. Takagi and S. Nishimoto, "High-resolution image reconstruction with latent diffusion models from human brain activity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 453–14 463.

[10] C. Q. Lai, H. Ibrahim, M. Z. Abdullah, J. M. Abdullah, S. A. Suandi, and A. Azman, "Artifacts and noise removal for electroencephalogram (eeg): A literature review," *2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, pp. 326–332, 2018.

[11] H. Fu, Z. Shen, J. J. Chin, and H. Wang, "Brainvis: Exploring the bridge between brain and visual signals via image reconstruction," *arXiv preprint arXiv:2312.14871*, 2023.

[12] Y. Bai, X. Wang, Y.-p. Cao, Y. Ge, C. Yuan, and Y. Shan, "Dreamdiffusion: Generating high-quality images from brain eeg signals," *arXiv preprint arXiv:2306.16934*, 2023.

[13] Y.-T. Lan, K. Ren, Y. Wang, W.-L. Zheng, D. Li, B.-L. Lu, and L. Qiu, "Seeing through the brain: image reconstruction of visual perception from human brain signals," *arXiv preprint arXiv:2308.02510*, 2023.

[14] D. Li, C. Wei, S. Li, J. Zou, and Q. Liu, "Visual decoding and reconstruction via eeg embeddings with guided diffusion," *arXiv preprint arXiv:2403.07721*, 2024.

[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[16] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[17] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721*, 2023.

[18] Q. Wang, X. Bai, H. Wang, Z. Qin, and A. Chen, "Instantid: Zero-shot identity-preserving generation in seconds," *arXiv preprint arXiv:2401.07519*, 2024.

[19] P. Wang and Y. Shi, "Imagedream: Image-prompt multi-view diffusion for 3d generation," *arXiv preprint arXiv:2312.02201*, 2023.

[20] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[21] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, and M. Shah, "Deep learning human mind for automated visual classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6809–6817.

[22] M. Ferrante, F. Ozcelik, T. Boccato, R. VanRullen, and N. Toschi, "Brain captioning: Decoding human brain activity into images and text," *arXiv preprint arXiv:2305.11560*, 2023.

[23] B. Zhang, P. Zhang, X. Dong, Y. Zang, and J. Wang, "Long-clip: Unlocking the long-text capability of clip," *arXiv preprint arXiv:2403.15378*, 2024.

[24] J. Deng, R. Socher, L. Fei-Fei, W. Dong, K. Li, and L.-J. Li, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, vol. 00, 06 2009, pp. 248–255.

[25] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 448–456.

[26] B. Xu, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[28] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, 2023.

[29] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[30] L. Liu, Y. Ren, Z. Lin, and Z. Zhao, "Pseudo numerical methods for diffusion models on manifolds," in *ICLR*, 2022.

[31] G. Gaziv, R. Beliy, N. Granot, A. Hoogi, F. Strappini, T. Golan, and M. Irani, "Self-supervised natural image reconstruction and large-scale semantic classification from brain activity," *NeuroImage*, vol. 254, p. 119121, 2022.

[32] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, 2016.

[33] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[34] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity." *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[35] Y. Lu, C. Du, Q. Zhou, D. Wang, and H. He, "Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion," in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 5899–5908.