

Towards Linguistic Neural Representation Learning and Sentence Retrieval from Electroencephalogram Recordings

Jinzhao Zhou

jinzhao.zhou@uts.edu.au

Yu-Cheng Chang

fred.chang@uts.edu.au

Yiqun Duan

yiqun.duan-1@uts.edu.au

Yu-Kai Wang

yukai.wang@uts.edu.au

Ziyi Zhao

ziyi.zhao-2@student.uts.edu.au

Thomas Do

thomas.do@uts.edu.au

Chin-Teng Lin

chin-teng.lin@uts.edu.au

ABSTRACT

Decoding linguistic information from non-invasive brain signals using EEG has gained increasing research attention due to its vast applicational potential. Recently, a number of works have adopted a generative-based framework to decode electroencephalogram (EEG) signals into sentences by utilizing the power generative capacity of pretrained large language models (LLMs). However, this approach has several drawbacks that hinder the further development of linguistic applications for brain-computer interfaces (BCIs). Specifically, the ability of the EEG encoder to learn semantic information from EEG data remains questionable, and the LLM decoder's tendency to generate sentences based on its training memory can be hard to avoid. These issues necessitate a novel approach for converting EEG signals into sentences. In this paper, we propose a novel two-step pipeline that addresses these limitations and enhances the validity of linguistic EEG decoding research. We first confirm that word-level semantic information can be learned from EEG data recorded during natural reading by training a Conformer encoder via a masked contrastive objective for word-level classification. To achieve sentence decoding results, we employ a training-free retrieval method to retrieve sentences based on the predictions from the EEG encoder. Extensive experiments and ablation studies were conducted in this paper for a comprehensive evaluation of the proposed approach. Our evaluation results demonstrate that our EEG encoder achieves up to 55.15% top-20 classification accuracy with unseen EEG signals. Visualization of the top prediction candidates reveals that our model effectively groups EEG segments into semantic categories with similar meanings, thereby validating its ability to learn patterns from unspoken EEG recordings. Additionally, using the predicted classification results, our retrieval method attains a recall@5 of up to 55.55% and a BLEU-1 score of 30.44% for sentence-level evaluation. Despite the exploratory nature of this work, these results suggest that our method holds promise for providing more reliable solutions for converting EEG signals into text.

KEYWORDS

Electroencephalogram, Brain-Computer-Interface, Multimodal Retrieval, Multimodal Understanding

1 INTRODUCTION

Decoding linguistic information from brain signals has traditionally relied on intracranial approaches, which offer promising prospects

for restoring communication abilities in individuals with paralysis or spinal cord injuries [29, 31, 48]. In contrast, the use of non-invasive brain signals such as EEG in linguistic decoding has only recently begun to attract research attention, due to their superior temporal resolution, portability, and safety [2, 3, 19].

For its ability to measure surface neural activity with high temporal resolution and detect a diverse range rhythmic patterns, EEG signals can capture electrical activity in the sensorimotor cortex, which is known to produce μ rhythms rich in information during speech production [39]. This inherent connection between EEG signals and speech has led to various successful approaches in decoding EEG into linguistic units such as syllables [4], phonemes [10, 11], and words [20, 30], despite limitations due to the scale of available datasets and subject variability.

On the other hand, decoding sentences from EEG signals during unspoken reading tasks presents several unique challenges. Firstly, unspoken speech elicits less discriminative brain activity compared to spoken speech, making it harder to distinguish between different neural responses [34]. Second, there exists significant data sparsity, as the number of semantic categories is large while the dataset size remains relatively small [32]. Third, the noisiness of thought during reading further complicates the decoding task. For instance, participants may not focus on every word equally, often paying less attention to grammatical words and more to words that contain crucial or interesting information from the sentence [15]. Existing methodologies for decoding reading sentences from EEG signals have predominantly relied on a framework that pairs an EEG encoder with a pretrained large language model (LLM) decoder, training and decoding sentences by the machine translation approach [18, 46]. However, recent analyses suggest that when training the EEG encoder with an LLM using a machine translation objective, the encoder's ability to genuinely learn to capture semantic EEG patterns remains unclear. Instead, the overpowered LLM decoder may generate sentences simply based on its training memory regardless of the EEG input [23]. These empirical findings underscore the necessity to validate the efficacy of learning EEG encoders from EEG reading tasks and highlight the limitations of using overly powerful pretrained LLMs for converting EEG signals into sentences.

To overcome the aforementioned limitations, we aim to develop a novel approach for EEG-to-sentence conversion, which seeks to eliminate the bias introduced by the training memory of an overpowered LLM while enabling the assessment of the semantic

information an EEG encoder learns from text-reading EEG data. To achieve this, we propose EEG-to-Text Retrieval (ETER), which consists of an EEG encoder and an unbiased sentence retrieval method. In particular, We first train our conformer-based EEG encoder using a masked contrastive learning loss to learn semantic EEG representations. Then we combine our EEG encoder with a classification head to predict a semantic keyword set (SK) for each input EEG signal. In the next stage, we employ a beam search retriever (BSR) to find the most relevant sentences based on the SK sets generated by our EEG encoder. Our two-step EEG-to-text retrieval method introduces two key features: it allows for a transparent evaluation of the EEG learning efficacy through word-level classification outputs and provides a training-free method to leverage these word-level results for sentence-level outputs. Extensive experiments and ablation analyses reveal that our EEG encoder effectively learns semantic EEG representations, achieving high accuracy in word-level classification. Additionally, results from the retriever demonstrate that the predictions from the first stage enable the retrieval of the correct ground-truth sentences. These findings underscore the feasibility of using a retrieval-based method for converting EEG signals into coherent sentences.

The main contributions of this paper are summarized as follows:

- We propose a novel retrieval-based approach for EEG-to-text conversion tasks. The ETER method leverages the output of a word-level EEG classifier to retrieve the most relevant sentence, thus eliminating the over-reliance on generative LLM decoders.
- We demonstrate the learning of effective semantic EEG representations using a Conformer-based EEG encoder trained with a masked contrastive objective. Visualization of the top prediction candidates further proves its capability to predict EEG signals as semantically related words.
- We designed a beam search retrieval method to efficiently retrieve relevant sentences from the prediction results of our Conformer-based EEG encoder. Despite the imperfect prediction from the EEG encoder, our retrieval method remains a viable solution for finding the correct sentence.
- We conduct extensive experiments to thoroughly validate the performance of the proposed ETER approach. Additionally, ablation studies confirm the vocabulary scalability of the method and validate our design choices, presenting a novel avenue for developing a linguistic BCI system.

2 RELATED WORKS

Linguistic unit or word decoding from brain signals Due to EEG's capacity to capture neural activities associated with speech production, pioneering works on linguistic decoding using EEG mainly focus on the decoding of linguistic units such as syllables or phonemes [11, 17, 43]. For instance, [7] proposed to extract autoregressive coefficients as features for imagined syllable classification with a k-nearest neighbor (KNN) classifier. [13] leveraged the Hilbert spectrum to extract features and classify the syllables using a Bayesian classifier.

To decode higher-level semantics, numerous studies have dedicated efforts to word-level classification using EEG signal [5, 20, 33, 33, 45, 51]. However, most of these studies have trained and

evaluated their models on a very limited dataset, typically comprising only 4 to 10 words or a narrow set of directional words. As a result, recent research has sought to expand the output vocabulary scope to a more practical scale, either at the word-level [28] or the pre-word level [49]. [16] used a large-scale word-level EEG dataset collected during listening, they enhance word-level classification accuracy through contrastive learning to align E/MEG signals with speech.

End-to-End Decoding from EEG to sentence The recent trend in EEG-based sentence decoding on the other hand predominantly employs end-to-end machine translation approaches. For instance, EEG-to-Text [46] pioneered open-vocabulary decoding of EEG signals into sentences, establishing an initial performance benchmark. In their work a Transformer-based EEG encoder is used to transform EEG signals into EEG representations while a pre-trained LLM model takes these EEG representations as input and generate sentences. Building upon this, DeWave [15] advanced decoding performance by introducing discrete codex and achieved text decoding directly from raw EEG waves. Subsequent innovations such as BELT [9, 52] and Curriculum Contrastive [18] introduced contrastive learning to enhance encoding quality. Additionally, NuSpeech [50] leveraged the end-to-end speech decoding model Whisper [36] to achieve commendable performance.

However, these end-to-end methods are prone to issues where a newly initialized EEG encoder combined with a powerful pre-trained language decoder leads to the decoder merely memorizing and generating the training text without truly utilizing information from the EEG modality. Consequently, this may result in the EEG encoder failing to learn to capture EEG patterns. Diverging from these LLM-based approaches, our work first develops an effective EEG encoder for word-level classification and explores the feasibility of using a training-free, unbiased retrieval method to achieve sentence-level output. This approach eliminates potential limitations such as implicit teacher forcing evaluation or test sentence leakage.

3 EEG-TO-TEXT RETRIEVAL

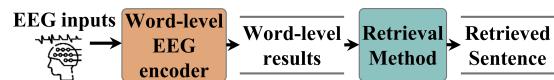


Figure 1: The overall structure of EEG-to-text retrieval approach. Our two-step approach consists of an word-level EEG encoder that encodes and prediction word-level results from EEG signals. Subsequently, a retrieval method is applied to find the most relevant sentence utilizing the word-level results.

In this section, we present our ETER approach, a two-step EEG-to-text retrieval method that identifies the most relevant sentence a participant reads based on word-level EEG classification results. The general pipeline of our approach is depicted in Figure 1. To achieve word-level decoding, we developed a Conformer-based EEG encoder. To learn semantic EEG representation, we guide the

EEG representation space using word representations extracted from a large language model and train the encoder with a masked contrastive objective. It is important to note that the language model is used solely to provide word representations for guiding the distribution of the learned EEG representation and is not utilized during testing. Subsequently, a classification head is added to the EEG encoder to fine-tune it for word-level prediction. Finally, we designed a beam search retrieval method to find relevant sentences based on the prediction results of the EEG encoder. Figure 2 gives an overview of the proposed method.

3.1 Preprocessing

To perform word-level EEG representation learning and classification, we first preprocess the dataset's vocabulary. Grammatical words such as "the," "a," "an," and "is" constitute a significant portion (40-60%) of English text in general [24, 26]. From a sample balance perspective, these grammatical words dominate the training and testing samples, potentially leading the EEG encoder or classifier to overemphasize on these words, which do not contain critical information about the sentence. Furthermore, previous neurobiological studies in reading comprehension has identified that "semantic strong" words elicit higher and more distinguishable neural patterns compared to "semantic moderate" words [25]. Therefore, during preprocessing, we remove EEG-word pairs containing these grammatical words from the dataset.

Additionally, we perform word lemmatization on the remaining vocabulary. The lemmatization step serves two purposes. First, we hypothesize that during reading comprehension, different forms of the same word will elicit similar neural patterns, as they convey the same meaning. So the EEG signals for these similar words can be seen as the same category. Second, this lemmatization process also increases the sample size for each word in the vocabulary and reduces the sparsity of the word-level training dataset.

For preprocessing the EEG signals, they are first transformed into word-level EEG embeddings using frequency-domain transformation following the same preprocessing pipeline in previous works [22, 47]. First, the EEG recordings are segmented according to the eye-tracking fixation on each word. Then, the segmented EEG signals are denoised and band-pass filtered into eight frequency bands: theta1 (4-6Hz), theta2 (6.5-8Hz), alpha1 (8.5-10Hz), alpha2 (10.5-13Hz), beta1 (13.5-18Hz), beta2 (18.5-30Hz), gamma1 (30.5-40Hz), and gamma2 (40-49.5Hz). The Hilbert transform is then applied to each channel. Finally, word-level EEG embeddings are obtained by averaging the frequency band power within each frequency band. In the remainder of this paper, we denote the word-level EEG embedding as \mathbf{e} . For the corresponding word of the EEG embedding, we use the embedding layer of a distilled BERT model [14] to convert it into word representation, denoted by \mathbf{w} for brevity. To enhance word-level EEG representation learning and classification performance, we apply standard normalization to the word-level EEG embeddings. Specifically, we compute the mean and standard deviation of \mathbf{e} for each subject and use these values for applying standard normalization. Empirically, we found that this normalization stabilizes the training process and improves performance, likely by suppressing noise and reducing inter-subject variations to some extent.

3.2 EEG Encoder

We train an EEG encoder for encoding and classifying EEG signals. We first tokenize \mathbf{e} into frequency tokens and then feed them to a Conformer encoder. The Conformer encoder outputs the same number of tokens as input, we use a global pooling layer to aggregate the information across all frequency bands into the final EEG representation \mathbf{h} .

3.2.1 Frequency-wise EEG tokenization. After preprocessing, the word-level EEG embedding has the shape of $\mathbf{e} \in \mathbb{R}^{N \times D}$. Here, N denotes the number of channels, and D is the number of frequency bands (in our case $D = 8$). To tokenize the EEG, we split \mathbf{e} into non-overlapping frequency bands across all channels $\{\mathbf{e}^{(i)}\}_{i=1,\dots,D}$. Since these EEG channels are distributed spatially on a participant's head so we employ spatial operations here to capture and aggregate frequency responses in a specific scalp area. As depicted in Figure 3, we use a spatial encoder to transform $\mathbf{e}^{(i)}$ into EEG token. The spatial encoder consist of a lightweight convolutional network. The spatial encoder comprises a lightweight convolutional network and a positional embedding layer. The convolutional network processes the channel dimension to produce embeddings that consolidate spatial information from specific frequency bands. Concurrently, the positional embedding layer is used to encode the positional information of the frequency bands, indicating which frequency range is contained within the input $\mathbf{e}^{(i)}$.

3.2.2 Conformer for EEG encoding. The detailed architecture of our EEG encoder is depicted in Figure 3. We use the conformer blocks [21] to build our EEG encoder for capturing both spectrum dependencies across EEG frequency bands and spatial relationships among channels [42, 52]. To aggregate the encoded EEG representations across all frequency bands, we used a global adaptive pooling layer to the output of the last conformer block and outputs \mathbf{h} as the final EEG representation for each word.

In a Conformer block, two feed-forward networks (FFN^1 and FFN^2), a multi-head self-attention (MHA) module, a convolution (CN) module are stacked together using residual connections. We applied a 1/2 weigh for the two FFN layers. The convolution module is depicted in Fig. 4, which is in turn comprised of two pointwise convolution layers and a depthwise convolution layer. The first pointwise convolution layer of the convolution module uses the gated linear unit (GLU) as the activation function. A batch normalization layer and a swish activation function were also used after the depthwise convolution layer. Overall, the Conformer blocks take the EEG embeddings \mathbf{e} as input and output the continuous EEG representation \mathbf{h} .

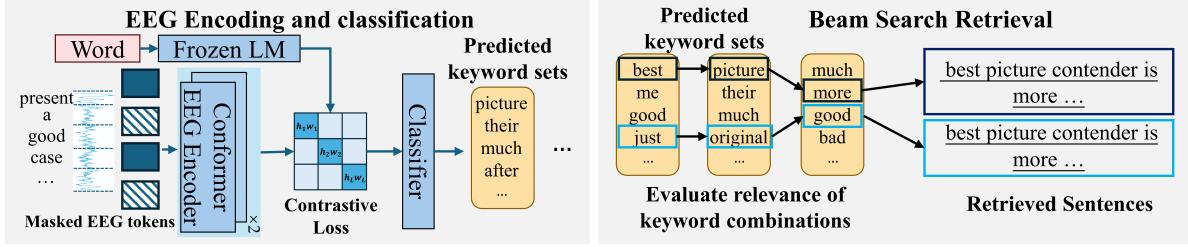


Figure 2: After segmenting and tokenizing EEG signals at the word level, an EEG encoder comprising two Conformer blocks learns semantic patterns from the EEG signals. We implement masked contrastive learning, leveraging a frozen language model to provide language supervision to the EEG representations. The EEG encoder predicts a keyword set independently for each input EEG segment. Finally, a sentence retriever utilizes these predicted keyword sets to identify the most pertinent sentence from the dataset corpus.

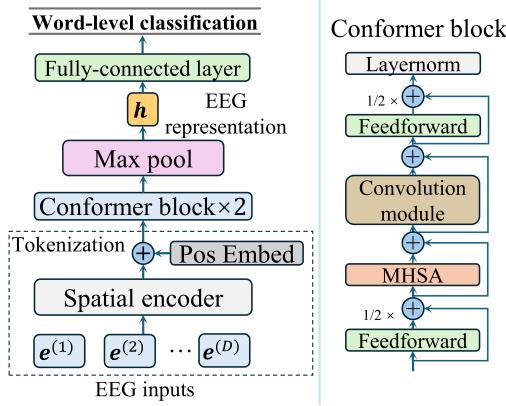


Figure 3: The architecture of the proposed EEG encoder. We first tokenize the low-level representations of EEG segments using a spatial encoder on the channel dimension and add positional embedding to indicate the frequency range of the token. Then we use 2 conformer blocks to further process these tokens. The Conformer block encodes input tokens using a multi-head self-attention (MHSA) and a convolution module. For classification, we use an adaptive max pooling layer to aggregate the output of all EEG tokens into the final EEG representation h for each word. For classification, a fully-connected layer will be used to the classification distribution of the input EEG signal.

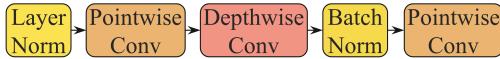


Figure 4: The detailed structure of the convolution module used in the Conformer blocks.

3.2.3 Masked contrastive training. To train the EEG encoder, we employ the masked contrastive learning objective, as depicted in Figure 5. This self-supervised approach aligns EEG representations h with word representations w , enabling the EEG encoder

to extract semantic information from EEG signals. This alignment ensures that EEG representations are not only closely related to its groundtruth word category but also to words with similar meanings. To further enhance the robustness of the EEG representations, we apply random masking to the input EEG tokens with a masking ratio η . Notably, we do not apply masking to the word embeddings to avoid introducing unnecessary noise into the learning process. The masked contrastive training loss function is defined by \mathcal{L}^{ct} as follows:

$$\mathcal{L}^{ct} = -\frac{1}{M} \sum_{i=1}^M \log \frac{\exp \text{sim}(h_i, w)_i / \tau}{\sum_{j=1}^M \exp \text{sim}(h_i, w_j) / \tau} \quad (1)$$

, where M is the training sample size of the dataset, τ is the temperature parameter that scales the logits, and $\text{sim}(\cdot, \cdot)$ denotes the dot product similarity measure. We employ a frozen, pretrained BERT model [14] as the text encoder to generate word representations and guide the learning of EEG representations. In our experiments, we empirically determined that a mask ratio of $\eta = 0.1$ and a temperature parameter of $\tau = 0.3$ yield optimal classification performance.

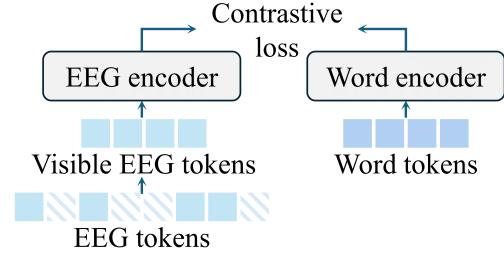


Figure 5: Our masked contrastive training scheme. Following CLIP [35] and FLIP [27], we perform contrastive learning on pairs of EEG segments and words. We randomly mask out EEG tokens with a small masking ratio and encode only the visible EEG tokens. Here, the slash-shadowed blocks mean the EEG token is masked and is not visible to the EEG encoder. We do not mask word tokens in our paper.

3.2.4 Word-level classification. While training an EEG encoder with a self-supervised objective provides a robust foundation for learning semantic representations, it alone is insufficient for effective EEG classification. To address this limitation, we introduce a supervised learning phase that augments the self-supervised training with an additional classification head. We use a fully-connected layer with softmax activation function as the classification head using the EEG representation \mathbf{h} . This layer maps the language-aligned EEG representations to specific word categories, leveraging supervised loss \mathcal{L}^{sup} (Equation 2) to refine the encoder’s predictions.

$$\mathcal{L}^{sup} = -\frac{1}{M} \sum_{i=1}^M y_i \log(p(\hat{y}_i | \mathbf{h}_i)) \quad (2)$$

, M denotes the number of training samples, y_i represents the one-hot encoded target word for the i -th sample, \hat{y}_i is the predicted word. In addition to the fully-connected layer classifier, we apply regularization techniques such as dropout and weight decay to prevent overfitting and ensure that the model generalizes well to unseen EEG samples.

3.3 Sentence Retrieval Method

In this section, we introduce the retrieval method designed to achieve EEG-to-text conversion based on the results from our word-level classification model. Our word-level EEG encoding and classification approach, as introduced previously, provides a solid and transparent measure of how well the encoder captures linguistic patterns from EEG data by allowing the direct evaluation using accuracy metrics. However, achieving high top-1 accuracy in linguistic EEG classifications remains a significant challenge under a large vocabulary as reported in previous works [6, 12, 38].

To address this limitation, we leverage a characteristic that emerged from our masked contrastive learning approach. After training, our model can generate top- k word predictions with similar meanings from the input EEG signals. This capability is crucial as it mitigates the challenges of achieving precise top-1 classification by aggregating semantically related words. This aggregation enhances the robustness and accuracy of our retrieval method, allowing for more reliable decoding of EEG signals into meaningful sentences. We denote the group of top- k prediction words as a keyword set (KS), denoted by S^k . Here, k denotes the number of top prediction words. Building upon this, we design our retrieval method to leverage the S^k from each “EEG word” to identify the most relevant sentence from the reading corpus. We denote the sequence of KSs in a sentence as $\mathcal{S} = \{S_i^k\}_{i=1, \dots, L}$, where L denotes the number of KS predicted for the sentence.

3.3.1 Beam search retrieval method. We depict the proposed beam search retrieval method (BSR) in Figure 6. The BSR method is designed to leverage a large search space that considers all k candidates in S^k , while reducing exponential memory consumption. BSR begins by constructing keyword combination queries from the first n KSs. Each query contains one candidate from a S^k , and will be scored according to its relevance to sentences in the dataset corpus. The scoring method will be explained in Section 3.3.2. The score for each query measures the relevance of this query to the dataset corpus. After scoring, we apply re-ranking to the queries and only

keep the best m combination queries for the next evaluation round. In the next round, the $(n+1)^{th}$ KS will be added to the queries to produce further combination queries. This iterative method ensures that at each step, we maintain the most promising combinations, incrementally building up to the final sentence retrieval. Mathematically, this iterative beam search process can be described as follows:

$$\begin{aligned} \mathbf{q}^0 &:= \{\emptyset\} \\ \mathbf{q}^l &= \underset{\substack{\mathbf{q}' \subseteq \mathcal{B}_l \\ |\mathbf{q}'|=m}}{\operatorname{argmax}} H(\mathbf{q}', C) \end{aligned} \quad (3)$$

, here \mathbf{q}^0 denotes the initial combination query set before the interactive search. It is an empty set as there is no relevant query is kept at the start. \mathbf{q}^l denotes the retained combination queries after the l^{th} iteration. We use \mathcal{B}_l to denote the new combinations obtained when adding the l^{th} KS (S_l^k) in this iteration. $H(\mathbf{q}', C)$ denotes a scoring method between the combinations \mathbf{q}' and sentences from the dataset corpus C . We set $|\mathbf{q}'| = m$ to limits the beam width of the searching. We calculate the candidate query set at $l > 0$ by:

$$\mathcal{B}_l = \{q \circ y | q \subseteq \mathbf{q}^{l-1}, y \in S_l^k\} \quad (4)$$

, where \circ denotes the concatenation operation. We borrow the process depicted in Figure 6 as an illustrative example. Assume we have a total of $L = 3$ KS in the sentence. Figure 6(a) shows all KSs from stage 1. In this example, none of the KS predicted the ground truth word as its top-1 prediction. However, the correct word can be found within the top- k prediction set. Figure 6(b) illustrates the ground truth words and ground truth reading sentence for reference. The BSR method, as shown in Figure 6(c), compares a number of combinations to the dataset corpus, distinguishing relevant combinations from irrelevant ones. In our example, the relevant combinations are [may, become, star], [during, time, work], and [during, time, much]. Using these relevant combinations, our model is able to identify the closest sentences from the dataset, including the ground truth sentence “During this time, he worked...” and returns this as the retrieval result.

3.3.2 Scoring Method. We use the Aho-Corasick algorithm [1] as the training-free scoring method. In particular, the Aho-Corasick algorithm efficiently finds all occurrences of the combination query within a sentence from the corpus by constructing a finite state machine. Thus, we denote the calculation of $H(\mathbf{q}, C)$ by:

$$H(\mathbf{q}, C) = \sum_{top-m} \max |q \cap c|, c \in C, q \in \mathbf{q} \quad (5)$$

, where $|q \cap c|$ denotes the number of occurrences of a query within a sentence c . We score a query using its average occurrence match with the sentence to allow the tolerance of “wrong keywords” in the query.

4 EXPERIMENT

4.1 Dataset

In this study, we utilize the Zurich Cognitive Language Processing Corpus (ZuCo) dataset [22] for training and evaluating the proposed method. The ZuCo dataset contains EEG data recorded during unspoken reading tasks involving 12 participants. It includes

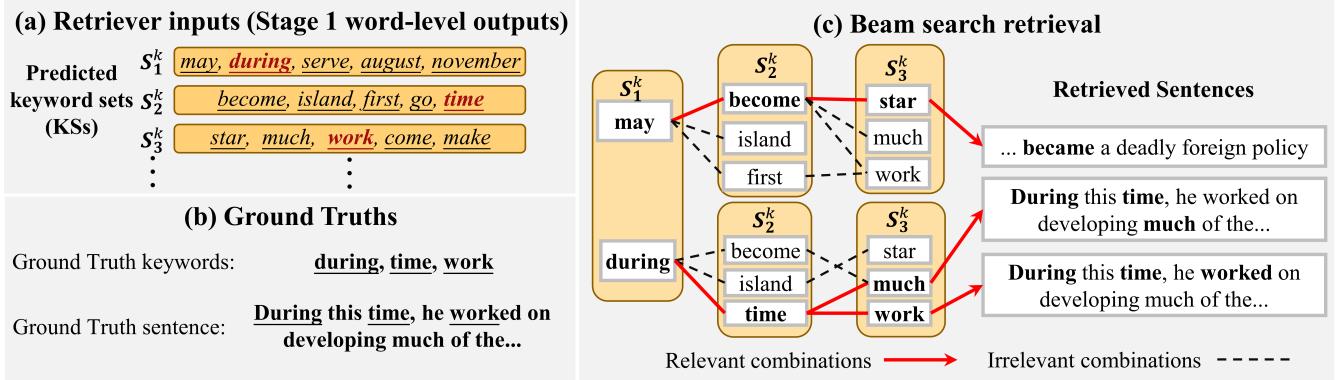


Figure 6: Illustration of the BSR method for retrieving reading sentences using word-level classification results from previous stage. (a) An example of the retriever’s input. Although stage-1 model fails to predict the correct word in its top-1 prediction, it predicts correct words within the top-k KSs. (b) Ground truth words reading sentence for reference. (c) The BSR method retrieves the ground truth sentence using these KSSs iteratively.

data from 105 EEG channels, with EEG waves denoised and filtered into eight frequency bands after segmentation. For our experiments, we use data from reading comprehension tasks, specifically Task 1 and Task 3 to evaluate the performance of our ETER method. Task 1 focuses on sentiment comprehension from movie reviews [41], while Task 3 involves understanding and extracting entities’ relation from Wikipedia biography articles. As discussed in Section 3.1, we removed all EEG-word pairs containing grammatical words from the dataset and performed lemmatization on the remaining words, merging words with the same lemmatized root form. Additionally, we observed a sharp decrease in sample numbers for words outside the top-100 most frequently occurring words in the remaining EEG-word pairs. As depicted in Figure 7, most long-tailed cases have fewer than 30 samples in the entire dataset, with some extreme cases having only one sample. This imbalance problem results in significant sparsity in the training dataset. Making it impossible to develop any effective word-level models on the full vocabulary of the dataset. To address this issue, we selected only the 100 most frequently appearing words from the dataset for training our EEG encoder. Although this selection may limit the system’s ability to scale, it provides relatively stable performance and serves as a reliable solution for our current needs. We have also conducted experiments involving a larger vocabulary in ablation studies, which will be discussed in Section 4.6.2.

4.2 Metrics

To ensure a thorough evaluation of our approach, we utilize a range of evaluation metrics for both the EEG classifier and the retrieval method. Firstly, we evaluate the effectiveness of the EEG encoder through classification accuracy assessment. In the context of sentence retrieval, we employ the retrieval metrics including recall@5 and precision@5 to evaluate the ability of our system to retrieve relevant sentence based on the results from EEG classification. Additionally, we calculate the BLEU metric [8] to quantify the relevance between the retrieved sentences and the target sentence.

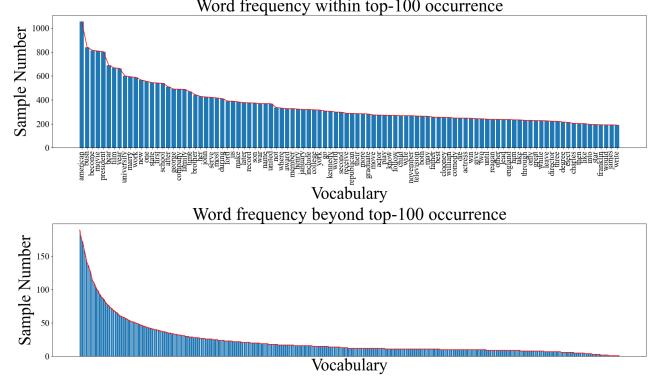


Figure 7: Samples number of the lemmatized vocabulary of the Zuko dataset. (top) sample number of words within the top 100 most occurring words. (bottom) sample number of words outside the top 100 most occurring words. A sharp decrease in sample number can be observed (red curve). EEG-word preprocessing. We preprocess and lematize the original sentence from the dataset. EEG-word pairs are extracted for training the word-level EEG encoder.

4.3 Implementation Details

We train a Conformer encoder with 2 Conformer blocks. We set the embedding dimension to 512 with 8 attention heads with the feed-forward dimension size of 1024. During training, we set the coefficient for training loss as $\alpha = 0.5$ and $\beta = 0.5$ respectively. We optimize the parameters of the Conformer models using AdamW optimizer with an initial learning rate of $1e^{-4}$ and a weight decay of 0.05. The learning rate warms up over the first 500 steps to $1e^{-2}$ and linearly decays to $1e^{-6}$. In all experiments, we set the batch size to 256 and train the model for 100 epochs. Training is performed on a single A40 GPU with 48 GBs of memory.

4.4 Word-level classification performance

We train and evaluate our EEG encoder and its ablative versions using the ZuCo dataset to demonstrate its ability to learn semantic representations from unspoken EEG signals. For the baseline, we use a random model that predicts a uniform distribution over the EEG segments. Our initial model is a conformer EEG encoder trained solely with the supervised learning loss \mathcal{L}^{sup} , without subject-baseline removal. We then assess the performance gains by incorporating subject-baseline removal (+bm.) and masked contrastive loss (+MCT). As shown in Table 1, our model predicts the correct word from EEG with a top-20 accuracy of 55.15% and a top-10 accuracy of 36.4%. This indicates that for more than half of the unseen EEG samples across different subjects, the ground truth words rank significantly higher than others within a 100-word vocabulary. Compared to the random baseline, our model achieves nearly three times higher accuracy. Furthermore, we observe that the addition of baseline removal and masked contrastive training improves the top-20 accuracy by 7.86% and 7.62%, respectively. These results highlight the incremental improvement provided by these methods in learning linguistic EEG patterns during reading.

Table 1: Word-level classification accuracy (%) on unseen EEG segments

Method	Top-1	Top-5	Top-10	Top-15	Top-20
Random model	1.08%	5.07%	9.57%	14.27%	19.19%
base model (Ours)	5.31%	16.36%	26.11%	32.72%	39.67%
+ bm.	6.48%	21.48%	31.24%	40.38%	47.53%
+ MCT	8.66%	24.90%	36.40%	46.28%	55.15%

¹ bm. denotes baseline removal using standard normalization for each participant.

² MCT denotes masked contrastive training.

4.5 Sentence-level retrieval performance

We evaluate the performance of the second-stage retrieval method using the sentiment movie review corpus from Task 1 and the Wikipedia biography corpus from Task 3. For these evaluations, we impose constraints on the number of available words within the sentences, requiring at least 5 or 7 Ks, as shown in Table 2. We compare the proposed BSR method using various scoring methods including the Aho-Corasick method, Levenshtein distance, and Term Frequency-Inverse Document Frequency (TF-IDF). When using Levenshtein distance, we compute the edit distance between the query and the compared sentences while for the TF-IDF method, we calculate the cosine similarity between the bag-of-word representations of the query and the comparison sentence retrieved from the corpus. As presented in Table 2, our experiment demonstrates the superior performance of using the Aho-Corasick-based scoring with our BSR method to accurately retrieve relevant sentences from the corpus based on input keyword sets. In the sentiment movie review corpus, our method achieves a recall@5 metric of 37.5% for sentences containing over 5 keyword sets. For sentences containing over 7 keyword sets in both corpora, we achieve a recall@5 of over 50%. Moreover, our method demonstrates the highest performance in retrieving relevant sentences, as evidenced by the BLEU metrics,

surpassing a BLEU-1 score of 40% on both reading corpora for sentences with over 7 keyword sets. Since These results are achieved without requiring any training in the retrieval method, it showcase the plausibility of the proposed ETER method for EEG-to-text conversion.

Aside from the quantitative results, Table 3 presents a qualitative assessment of the proposed ETER method. For qualitative comparison with a generative LLM decoder, we additionally fine-tuned a T5 model [37] to generate ground truth sentences using lists of keywords as input. We show that our approach effectively retrieved top-ranking sentences in the first example case. In contrast, the T5 model produced sentences outside the training dataset, which is largely based on its pre-training memory. In the last example, although our model failed to find the correct sentence, it still managed to successfully identify keywords like 'best', enabling retrieval of similarly sentiment sentences from the corpus. This underscores the efficacy of our retrieval-based method in transcribing EEG signals into text given an imperfect word-level EEG classifier.

4.6 Ablations and Discussions

4.6.1 EEG encoder’s architecture. We delve deeper into assessing the effectiveness of different encoder architectures using the same training paradigm as the proposed method. Specifically, we interchange Conformer blocks with Transformer blocks [44] or Emformer blocks [40], in our EEG encoder architecture and evaluate their word-level classification performance. Results are presented in Table 4. To begin with, both the Conformer and Emformer, with their ability to capture local patterns across channels, exhibit notably superior performance compared to the general Transformer encoder. This observation underscores the significance of leveraging structures that exploit local patterns, justifying our design choice of EEG encoder. Notably, the Conformer encoder achieves better performance than the Emformer-based encoder in our assessments, indicating that the convolutional layers in the conformer architecture allow the model to learn hierarchical features, which can be crucial for understanding complex signals such as EEG. Additionally, our results show that the introduction of a reconstructive term does not yield a consistence enhancement to the performance, further validating our choice of a masked contrastive learning scheme without a reconstructed decoder.

4.6.2 Vocabulary size. The ablation results on vocabulary size is depicted in Figure 8. This result highlight the scalability of our proposed method across varying vocabulary sizes. Notably, our approach consistently outperforms competing models, demonstrating robustness even with smaller vocabulary sizes. However, as the vocabulary size increases to include more than 200 words, we can observe a significant decline in decoding performance. This decline is primarily due to the inherent imbalance and increased scarcity in the dataset as shown in Figure 7, where words with lower frequencies lack sufficient training data. Despite these challenges, our proposed Conformer model maintains competitive performance, achieving a top-10 accuracy of 28.4% with a 200-word vocabulary. This result compares favorably to recent classifications of listening EEG data, which reported a top-10 accuracy of $31.4 \pm 1.59\%$ with a 203-word vocabulary size [12]. This comparison underscores the efficacy of our approach, particularly with the exclusion of

Table 2: Sentence Retrieval Performance

Reading Corpus	Scoring Method	Recall@5 ($T \geq 5$)	Precision@5 ($T \geq 5$)	Recall@5 ($T \geq 7$)	Precision@5 ($T \geq 7$)	BLEU-1 ($T \geq 5$)	BLEU-4 ($T \geq 5$)	BLEU-1 ($T \geq 7$)	BLEU-4 ($T \geq 7$)
(Task 1) Sentiment	TF-IDF	2.94%	0.73%	0.20%	0.06%	6.20%	0.00%	9.58%	0.00%
	L.D.	18.75%	4.06%	50.00%	10.00%	14.60%	1.85%	18.03%	3.47%
	A.C.	37.50%	22.90%	50.00%	19.64%	30.37%	20.68%	47.92%	41.81%
(Task 3) Wikipedia	TF-IDF	3.03%	1.51%	6.75%	2.98%	11.33%	3.80%	9.11%	5.96%
	L.D.	7.69%	2.10%	15.00%	4.58%	16.21%	4.72%	21.35%	8.20%
	A.C.	15.38%	5.55%	55.55%	20.37%	29.12%	14.08%	49.86%	36.27%

¹ L.D. denotes Levenshtein distance.

² A.C. denotes Aho–Corasick score.

Table 3: Visualization of the retrieved sentences

GT Sentence	hepburn win an emmy award in 19xx for her lead role in love among the ruin and be nominate for four other emmy and two tony award during the course of her more than 70 year act career
T5 Generation	Among those attending the Kennebunkport, Maine wedding ceremony were Isabel Stillman Rockefeller (daughter of Percy Rockefeller), Hope Lincoln, Mary Keck...
ETER Query	[‘award’, ‘war’, ‘her’]
ETER Retrieval	hepburn win an emmy award in 19xx for her lead role in love among the ruin and be nominate for four other emmy and two tony award during the course of her more than 70 year act career
GT Sentence	henry ford july 30 19xx april 7 19xx be the founder of the henry ford motor company which later become cadillac and ford motor company
T5 Generation	During this period, McNamara helped to plan the 1945 bombing of Tokyo.
ETER Query	[‘found’, ‘company’, ‘later’, ‘henry’, ‘ford’, ‘john’, ‘become’, ‘then’]
ETER Retrieval	henry ford july 30 19xx april 7 19xx be the founder of the henry ford motor company which later become cadillac and ford motor company
GT Sentence	like the best of godard’s movie it be visually ravish penetrate impenetrable
T5 Generation	He was a member of the Executive Committee of the United States Golf Association (USGA) from 1928–1935, serving successively as Secretary, Vice President and President.
ETER Query	[‘other’, ‘best’, ‘not’]
ETER Retrieval	it’s the best film of the year so far the benchmark against which all other best picture contender should be measure
GT Sentence	adam receive one electoral vote in the presidential election of 19xx
T5 Generation	During this time, he was prescribed Ritalin for hyperactivity; years later, his wife Annette had been prescribed Ritalin for hyperactivity;
ETER Query	[‘adam’, ‘january’, ‘elect’]
ETER Retrieval	although adam lose in both the popular and electoral vote in the presidential election of 19xx ...

Table 4: Word-level classification results from various architectures and encoding block selection.

Encoding block	Top-1	Top-5	Top-10	Top-15	Top-20
Transformer	5.66%	19.11%	29.23%	35.89%	40.01%
+ reconstruction	5.16% (-0.60)	18.03% (-1.08)	26.60% (-2.63)	32.51% (-3.38)	36.51% (-3.50)
Emformer	7.74%	21.57%	32.64%	41.38%	48.83%
+ reconstruction	7.95% (+0.21)	24.27% (+2.70)	36.39% (+3.75)	45.75% (+4.37)	53.75% (+4.92)
Conformer	8.66%	24.90%	36.40%	46.28%	55.15%
+ reconstruction	8.48% (-0.18)	24.72% (-0.18)	36.95% (+0.55)	45.97% (-0.31)	54.32% (-0.83)

grammatical words from the vocabulary, which enhances the focus on meaningful content words and improves overall performance.

However, it also suggests that with a more balanced dataset, our proposed method has the potential to achieve even higher word-wise

performance. This improvement would, in turn, enhance retrieval accuracy for sentence-based BCI systems on a larger scale.

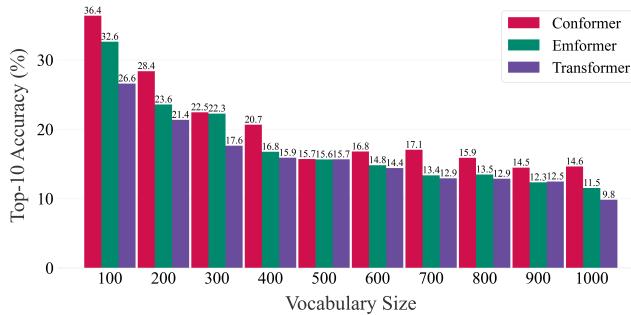


Figure 8: Ablation on different vocabulary size on different encoder architectures.

4.6.3 Visualization of word-level results. Figure 9 illustrates the top-10 keyword set predicted by the encoder on unseen EEG samples from the test set. After training, our model shows a strong capability of encoding EEG signals to similar concepts. For example, when predicting ‘university’, our model also considers ‘graduate’, ‘school’, and ‘college’ to be in the same keyword group, indicating the model has assigned these concepts into a close semantic representation space. We consider the clustering of meaningful concepts to support that our encoder has learned useful representation from the brain signal and has aligned these linguistic representations with language modality in the subspace.

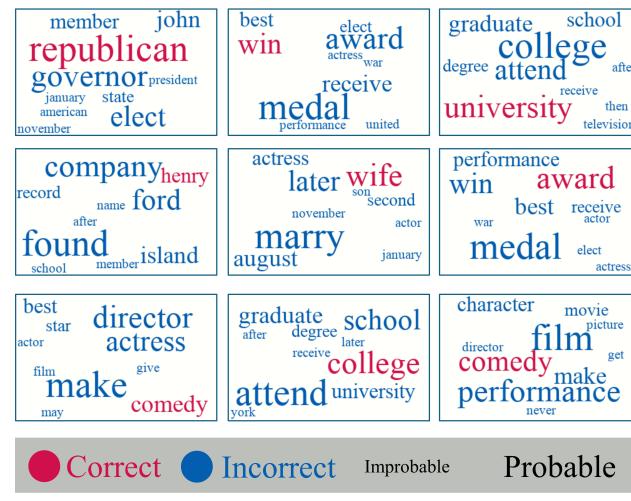


Figure 9: Visualization of top-10 prediction results from the single-word prediction using our EEG encoder. Text color indicates whether the predicted word is correct and text size is proportional to the likelihood of the model’s predictions.

4.6.4 Retrieval method and number of keyword sets. Lastly, we compare the performance of our proposed beam search retrieval method with a more straightforward greedy retrieval (GS) method and observe the retrieval performance of retrieval methods using a different number of KSs as input. The GS method is designed to use only the top-1 prediction from each keyword set for retrieving sentences from the corpus. The comparison results, presented in Figure 10, show that both strategies perform well on the training set as the number of keyword sets within a sentence increases. This observation suggests the potential of extracting full sentences from EEG signals using a high-accuracy EEG encoder. However, the GS method’s performance significantly decreases on the test set due to its limited capacity to explore a broader array of queries, constrained by the EEG encoder’s moderate top-1 prediction performance. This finding underscores the rationale for implementing a beam search-based retrieval strategy. On the other hand, we are aware that our EEG encoder is currently constrained to the top 100 highest-frequency words, limiting the number of keywords that can be decoded from a sentence and, consequently, the application scenarios of our method. Despite this limitation, the proposed retrieval method shows improved performance when more keywords are given, even if some noise is present. These findings suggest that our method is promising and has potential for future development, particularly if the vocabulary constraint can be addressed to allow for a broader range of keywords and more accurate sentence retrieval.

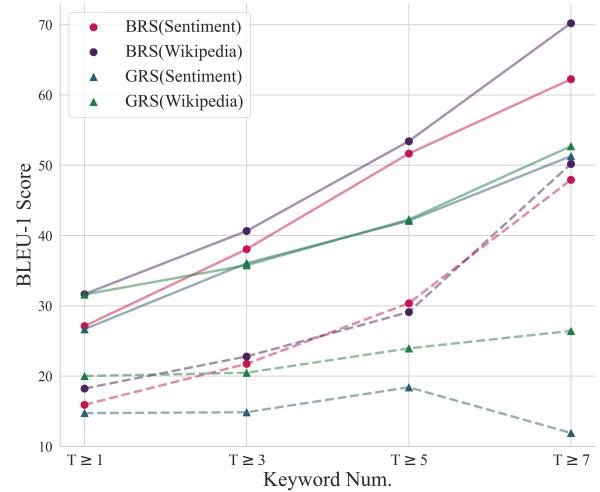


Figure 10: Comparison of retrieval relevance of BSR and GS Strategies Across Varying Numbers of EEG-Word Pairs in Sentences. Solid lines and dashed lines depict retrieval performance based on predicted keyword sets from training EEG segments and unseen EEG segments respectively.

5 CONCLUSION

This paper demonstrates the potential of combining an EEG encoder with a retrieval method to convert EEG signals into sentences, introducing a pioneering approach termed EEG-to-text retrieval(ETER).

This novel method employs a transparent EEG encoder, to learn semantic patterns from EEG data. By extracting keyword sets from unseen EEG segments, ETER enables the sentence retriever to identify the most relevant sentences from a corpus. Both quantitative and qualitative evaluations affirm the efficacy of our approach in acquiring meaningful semantic representations and retrieving relevant sentences. Our extensive experiments and ablation studies validate the approach's ability to learn patterns from unspoken EEG recordings both quantitatively and qualitatively, demonstrating that our method holds promise for providing more reliable solutions for converting EEG signals into text. Despite the achieved results, we recognize the substantial room for future improvement. Given the exploratory nature of this research, we only employed a simple retrieval method and tested it on a limited vocabulary set. Our future work will focus on exploring more diverse datasets to continuously improve the EEG encoder design and enhance retrieval methods to accommodate larger vocabularies, thereby improving sentence retrieval accuracy on a larger scale. Additionally, collecting more EEG data at the word level will be pursued to further advance research in linguistic EEG decoding.

REFERENCES

- [1] Alfred V Aho and Margaret J Corasick. 1975. Efficient string matching: an aid to bibliographic search. *Commun. ACM* 18, 6 (1975), 333–340.
- [2] Soraya M Alarcao and Manuel J Fonseca. 2017. Emotions recognition using EEG signals: A survey. *IEEE transactions on affective computing* 10, 3 (2017), 374–393.
- [3] Phillip M Alday. 2019. M/EEG analysis of naturalistic stories: a review from speech to language processing. *Language, Cognition and Neuroscience* 34, 4 (2019), 457–473.
- [4] Milad Amani Arjestan, Mansour Vali, and Farhad Faradji. 2016. Brain computer interface design and implementation to identify overt and covert speech. In *2016 23rd Iranian Conference on Biomedical Engineering and 2016 1st International Iranian Conference on Biomedical Engineering (ICBME)*. IEEE, 59–63.
- [5] MR Asghari Bejestani, Mohammad Khani, VR Nafisi, F Darakeh, et al. 2022. EEG-based multiword imagined speech classification for persian words. *BioMed Research International* 2022 (2022).
- [6] Mohamad Amin Bakshali, Morteza Khademi, Abbas Ebrahimi-Moghadam, and Sahar Moghimi. 2020. EEG signal classification of imagined speech based on Riemannian distance of correntropy spectral density. *Biomedical Signal Processing and Control* 59 (2020), 101899.
- [7] Katharine Brigham and BVK Vijaya Kumar. 2010. Imagined speech classification with EEG signals for silent communication: a preliminary investigation into synthetic telepathy. In *2010 4th International Conference on Bioinformatics and Biomedical Engineering*. IEEE, 1–4.
- [8] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* (2020).
- [9] Charles Chau, Yiqun Duan, Yu-Cheng Chang, Thomas Do, Yu-Kai Wang, and Chin-teng Lin. 2023. BELT-2: Bootstrapping EEG-to-Language representation alignment for multi-task brain decoding. (2023).
- [10] Ciaran Cooney, Rafaella Folli, and Damien Coyle. 2018. Mel frequency cepstral coefficients enhance imagined speech decoding accuracy from EEG. In *2018 29th Irish Signals and Systems Conference (ISSC)*. IEEE, 1–7.
- [11] Ciaran Cooney, Rafaella Folli, and Damien Coyle. 2019. Optimizing layers improves CNN generalization and transfer learning for imagined speech decoding from EEG. In *2019 IEEE international conference on systems, man and cybernetics (SMC)*. IEEE, 1311–1316.
- [12] Alexandre Défossez, Charlotte Caucheteux, Jérémie Rapin, Ori Kabeli, and Jean-Rémi King. 2023. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence* 5, 10 (2023), 1097–1107.
- [13] Siyi Deng, Ramesh Srinivasan, Tom Lappas, and Michael D’Zmura. 2010. EEG classification of imagined syllable rhythm using Hilbert spectrum methods. *Journal of neural engineering* 7, 4 (2010), 046006.
- [14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- [15] Yiqun Duan, Charles Zhou, Zhen Wang, Yu-Kai Wang, and Chin teng Lin. 2023. DeWave: Discrete Encoding of EEG Waves for EEG to Text Translation. In *Thirty-seventh Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=WaLi8slhLw>
- [16] Alexandre Défossez, Charlotte Caucheteux, Jérémie Rapin, Ori Kabeli, and Jean-Rémi King. 2023. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence* 5, 10 (Oct. 2023), 1097–1107. <https://doi.org/10.1038/s42256-023-00714-5>
- [17] Michael D’Zmura, Siyi Deng, Tom Lappas, Samuel Thorpe, and Ramesh Srinivasan. 2009. Toward EEG sensing of imagined speech. In *Human-Computer Interaction. New Trends: 13th International Conference, HCI International 2009, San Diego, CA, USA, July 19–24, 2009, Proceedings, Part I* 13. Springer, 40–48.
- [18] Xiaocheng Feng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. Aligning Semantic in Brain and Language: A Curriculum Contrastive Method for Electroencephalography-to-Text Generation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* (2023).
- [19] Isabelle Gaudet, Alejandra Hüser, Phetsamone Vannasing, and Anne Gallagher. 2020. Functional brain connectivity of language functions in children revealed by EEG and MEG: a systematic review. *Frontiers in human neuroscience* 14 (2020), 62.
- [20] Erick F González-Castañeda, Alejandro A Torres-García, Carlos A Reyes-García, and Luis Villaseñor-Pineda. 2017. Sonification and textification: Proposing methods for classifying unspoken words from EEG signals. *Biomedical Signal Processing and Control* 37 (2017), 82–91.
- [21] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100* (2020).
- [22] Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific data* 5, 1 (2018), 1–13.
- [23] Hyeyeong Jo, Yiqian Yang, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hee Lee. 2024. Are EEG-to-Text Models Working? *arXiv preprint arXiv:2405.06459* (2024).
- [24] M Kontra. 1985. Frequency Analysis of English Usage: Lexicon and Grammar.
- [25] Marta Kutás and Steven A Hillyard. 1980. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science* 207, 4427 (1980), 203–205.
- [26] Geoffrey Leech, Paul Rayson, et al. 2014. *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.
- [27] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. 2023. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23390–23400.
- [28] Diego Lopez-Bernal, David Balderas, Pedro Ponce, and Arturo Molina. 2022. A state-of-the-art review of EEG-based imagined speech decoding. *Frontiers in human neuroscience* 16 (2022), 867281.
- [29] 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, et al. 2022. Generalizable spelling using a speech neuroprosthesis in an individual with severe limb and vocal paralysis. *Nature communications* 13, 1 (2022), 6510.
- [30] Kusuma Mohanchandra and Snehantha Saha. 2016. A communication paradigm using subvocalized speech: translating brain signals into speech. *Augmented Human Research* 1, 1 (2016), 3.
- [31] David A Moses, Sean L Metzger, Jessie R Liu, Gopala K Anumanchipalli, Joseph G Makin, Pengfei F Sun, Josh Chartier, Maximilian E Dougherty, Patricia M Liu, Gary M Abrams, et al. 2021. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine* 385, 3 (2021), 217–227.
- [32] Nicolás Nieto, Victoria Peterson, Hugo Leonardo Rufiner, Juan Esteban Kamienkowski, and Ruben Spies. 2022. Thinking out loud, an open-access EEG-based BCI dataset for inner speech recognition. *Scientific Data* 9, 1 (2022), 52.
- [33] Dipti Pawar and Sudhir Dhage. 2020. Multiclass covert speech classification using extreme learning machine. *Biomedical engineering letters* 10, 2 (2020), 217–226.
- [34] Timothée Proix, Jaime Delgado Saa, Andy Christen, Stephanie Martin, Brian N Pasley, Robert T Knight, Xing Tian, David Poeppel, Werner K Doyle, Orrin Devinsky, et al. 2022. Imagined speech can be decoded from low-and cross-frequency intracranial EEG features. *Nature communications* 13, 1 (2022), 48.
- [35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [36] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*. PMLR, 28492–28518.
- [37] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.

- [38] Ana-Luiza Rusnac and Ovidiu Grigore. 2022. Imaginary speech recognition using a convolutional network with long-short memory. *Applied Sciences* 12, 22 (2022), 11873.
- [39] Tim Saltuklaroglu, Andrew Bowers, Ashley W Harkrider, Devin Casenhisier, Kevin J Reilly, David E Jenson, and David Thornton. 2018. EEG mu rhythms: Rich sources of sensorimotor information in speech processing. *Brain and language* 187 (2018), 41–61.
- [40] Yangyang Shi, Yongqiang Wang, Chunyang Wu, Ching-Feng Yeh, Julian Chan, Frank Zhang, Duc Le, and Mike Seltzer. 2021. Emformer: Efficient memory transformer based acoustic model for low latency streaming speech recognition. In *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6783–6787.
- [41] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1631–1642.
- [42] Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. 2022. EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 31 (2022), 710–719.
- [43] Markus-Oliver Tamm, Yar Muhammad, and Naveed Muhammad. 2020. Classification of vowels from imagined speech with convolutional neural networks. *Computers* 9, 2 (2020), 46.
- [44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [45] Li Wang, Xiong Zhang, Xuefei Zhong, and Yu Zhang. 2013. Analysis and classification of speech imagery EEG for BCI. *Biomedical signal processing and control* 8, 6 (2013), 901–908.
- [46] Zhenhailong Wang and Heng Ji. 2022. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 5350–5358.
- [47] Zhenhailong Wang and Heng Ji. 2022. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 5350–5358.
- [48] Francis R Willett, Donald T Avansino, Leigh R Hochberg, Jamie M Henderson, and Krishna V Shenoy. 2021. High-performance brain-to-text communication via handwriting. *Nature* 593, 7858 (2021), 249–254.
- [49] Francis R Willett, Erin M Kunz, Chaofei Fan, Donald T Avansino, Guy H Wilson, Eun Young Choi, Foram Kamdar, Matthew F Glasser, Leigh R Hochberg, Shaul Druckmann, et al. 2023. A high-performance speech neuroprosthesis. *Nature* 620, 7976 (2023), 1031–1036.
- [50] Yiqian Yang, Yiqun Duan, Qiang Zhang, Renjing Xu, and Hui Xiong. 2024. Decode Neural signal as Speech. *arXiv preprint arXiv:2403.01748* (2024).
- [51] Shunran Zhao and Frank Rudicic. 2015. Classifying phonological categories in imagined and articulated speech. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 992–996.
- [52] Jinzhao Zhou, Yiqun Duan, Yu-Cheng Chang, Yu-Kai Wang, and Chin-Teng Lin. 2023. Belt: Bootstrapping electroencephalography-to-language decoding and zero-shot sentiment classification by natural language supervision. *arXiv preprint arXiv:2309.12056* (2023).

Received 20 July 2024; revised 12 March 2009; accepted 5 June 2009