# Pretraining Large Brain Language Model for Active BCI: Silent Speech


## Abstract

This paper explores silent speech decoding in active brain-computer interface (BCI) systems, which offers more natural and flexible communication than traditional BCI applications. To overcome the reliance on external stimuli in passive BCI settings, we collected a new silent speech dataset of over 120 hours of electroencephalogram (EEG) recordings from 12 subjects, capturing 24 commonly used English words for pretraining and decoding. Following recent trend of pretraining large models with self-supervised paradigms to enhance representation learning, we propose the Large Brain Language Model (LBLM) pretrained for decoding silent speech for active BCI. To train LBLM, we propose Spectro-Temporal Predictive (STP) pretraining, a novel self-supervised approach for learning EEG representations. Unlike existing EEG pretraining in a single domain, our proposed STP method employs autoregressive modeling in the spatio-temporal domain, capturing both temporal transitions and frequency variations from the EEG signal. After pretraining, we finetune the LBLM backbone on downstream tasks, including word-level and semantic-group classification, using a spatio-temporal classifier to integrate learned representations across EEG channels. Extensive experiments demonstrate significant performance gains of LBLM over baseline models. In a challenging cross-session setting, our model achieves 42.8% accuracy in semantic classification and 39.6% in word-level classification on average, outperforming baseline methods by 4.6% and 7.3%, respectively. Our results highlight the feasibility of silent speech decoding in active BCI systems, paving the way for more natural and practical brain-to-text communications.



## CCS Concepts

• **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

## Keywords

Active Brain-Computer-Interface, EEG Silent Speech, Large Brain Language Model, Auto-regressive Pretraining



Author's Contact Information:





ACM 1557-735X/2018/8-ART111
https://doi.org/XXXXXXX.XXXXXXX





**ACM Reference Format:**

. 2018. Pretraining Large Brain Language Model for Active BCI: Silent Speech. *J. ACM* 37, 4, Article 111 (August 2018), 14 pages. https://doi.org/XXXXXXX.XXXXXXX


## 1 Introduction

Brain-computer interface (BCI) applications have been growing rapidly in recent years [1]. Among them, non-invasive electroencephalography (EEG)-applications remain the most competitive choice in passive BCI applications due to the low-risk protocol and affordability [2–4]. Building upon non-invasive EEG, active BCIs allow users to voluntarily control and interact with surrounding environment, positioning BCIs as potential alternatives to manual or touchless input devices [5, 6]. For instance, current motor imagery (MI)-based active BCI systems translate imagined limb movements into directional control signals [7, 8]. However, MI-based active BCI remains counterintuitive and limits user's ability to express more complex ideas. In contrast, language is inherently more natural and versatile, allowing for more efficient and expressive communication. Consequently, silent speech decoding in active BCI systems holds significant potential for assisting users with speech impairments or those who need to communicate in quiet environments.

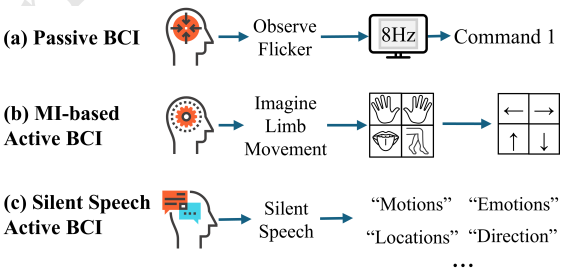



**Figure 1: Comparison of BCI paradigms for control. (a) Passive BCI: users focus on external stimuli, generating specific brain signal patterns that are mapped to predefined commands. (b) Motor Imagery (MI)-based active BCI: users voluntarily imagine limb movements, producing neural activity that is classified and translated into directional commands. (c) Silent speech active BCI: users generate neural activity through silent articulations, which is decoded into words to enable more natural communication.**

In neuroimaging studies, language-related neural activity has been identified using surface brain monitoring techniques, such as EEG[9–11]. These findings have spurred efforts to extract semantic information from EEG signals and translate them into text. For example, EEG2text[12] and DeWave[13] employed EEG-language alignment training schemes to bridge the modality gap between EEG and text, ultimately generating coherent sentences. Furthermore, spectrum alignment methods showed the potential to decode a more diverse range of vocabulary from EEG signals when text

is presented to subjects through audio or visual stimuli[14–17]. However, two key limitations remain to hinder the progression of language-based BCI research:

**(1) The difference between passive and active approaches in language processing:** Passive BCI paradigms for text generation rely on external stimuli, such as auditory cues during listening or visual cues during reading, to create brain responses. In contrast, silent speech active BCI detects language-related neural patterns without external stimulation [18–23]. Due to this fundamental difference, models trained on passive EEG data are not suitable for active BCI applications. Furthermore, existing datasets for imagined or silent speech are typically small and often consist of non-meaningful utterances [24] or have a limited vocabulary set [25, 26], which restricts their applicability in real-world scenarios and limits their potential for training large-scale brain models.

**(2) Lack of self-supervised pretraining paradigm for EEG silent speech:** While multimodal EEG alignment methods have proven effective for bootstrapping EEG representation learning[17, 27], they rely on pairing EEG signals with external modalities, such as visual or auditory stimuli presented alongside silent speech. In contrast, unimodal self-supervised methods leverage unlabeled EEG data to learn rich representations without external signals, offering greater flexibility, particularly in language-related EEG tasks where labeled data is often scarce. However, due to the relatively smaller dataset sizes and higher inter-session/subject variability in EEG silent speech compared to other EEG tasks [28, 29], pretraining large models using simple self-supervised learning paradigms may lead to data inefficiency, poor generalization, and overfitting issues [30–32].

In this study, we address the challenges in silent speech active BCI by introducing the pretrained large brain language model (LBLM) for capturing neural patterns during silent speech. To achieve this goal, we collected a large-scale silent speech dataset with over 120 hours recordings from 12 subjects. Our dataset focuses on a vocabulary of 24 commonly used English words from 6 semantic groups. To pretrain the LBLM model, we introduce Spectro-Temporal Predictive (STP) pretraining paradigm to learn temporal-spectral representations of EEG signals. This approach explicitly models both the temporal dependencies of EEG signals and their corresponding spectral components that are crucial for language decoding. Specifically, our pretraining strategy consists of two stages where we first initialize the LBLM backbone through a masked EEG modeling (MEM) objective, and then use the STP objective to enforces future prediction of both future EEG waves and frequency characteristics which go beyond the simple the recovery of masked wave segments. Using the self-supervised pretraining method, we pretrain our LBLM model with over 22 million parameters on the collected dataset using a transformer-style backbone enhanced with a layer-gating mechanism. For downstream classification tasks, a spatio-temporal classifier is developed to integrate and select backbone representations from different channels. In summary, the main contribution of this paper are:

- **Self-supervised pretraining paradigm for silent speech**: We propose a self-supervised pretraining method for EEG backbone models without labeled data or external modalities. This pretraining paradigm encourages models to capture temporal-spectral patterns from EEG signals, enabling them to effectively perform downstream tasks such as signal recovery and classification.
- **Large EEG Language Model (LBLM) for Silent Speech Decoding**: We propose LBLM, a 22M-parameter model for feature extraction from language-related EEG signals. Extensive experiments evaluate the effectiveness of fine-tuning LBLM on word-level and semantic-group classification tasks. Our results show that LBLM achieves state-of-the-art performance in cross-session evaluations. These findings highlight the feasibility of silent speech decoding from EEG signals.
- **Large-Scale Silent Speech EEG Dataset**: We collected a 120-hour EEG dataset from 12 participants for silent speech decoding, with 16 sessions recorded form each participant. The dataset will be made publicly available to support active BCI research and neural pattern analysis.

## 2 Related Works

### 2.1 EEG Semantic Decoding

Non-invasive neuroimaging research have significantly enhanced our understanding of how the human brain processes semantic concepts[9–11, 33]. Building on these discoveries, there has been growing interest in developing deep learning models that can capture and decode linguistic activity from EEG[14, 16, 34]. Earlier works on this field mainly focused on decoding non-meaningful syllables[26, 35, 36] or directional words[37–39], primarily using simple classification models, where convolutional neural networks (CNNs) or Transformer-based architectures were directly trained for classification tasks. For more complex sentence decoding tasks, while more sophisticated model architectures are employed, training methods still follow an end-to-end approach without pretraining. For example, [40] improved the structure of a CNN-based EEG encoder and enhanced training by aligning EEG spectrograms with those of audio signals. Similarly, recent EEG-to-Text methods[12, 13] utilized complex transformer-style EEG encoders paired with a pretrained language model to generate sentences. In these approaches, we found only the language model is pretrained, while the EEG encoder is trained from scratch to align EEG tokens with text tokens. Although these methods generate coherent sentences, end-to-end training on complex EEG signals often leads to an undertrained encoder, failing to capture sufficient semantic information and making precise EEG decoding[41, 42].

Recently, it has shifted toward training models with self-supervised learning techniques to improve the generalizability of learned EEG representations. Models such as Larbram[43], EEGPT[44], and NeuroLM[45] leveraged large-scale EEG datasets and masked reconstruction pretraining to enhance the representations learned by the models. These methods have shown improvements, outperforming task-specific models that were not pretrained. However, these large models are typically applied to classification tasks with a limited number of categories (e.g., 2 to 6 classes), where EEG signals within each class are more distinct. Examples include distinguishing normal from abnormal EEG signals in the TUAB dataset[46].

In contrast, decoding silent speech is more difficult since EEG signals for different words are often more similar, especially within the same semantic groups. To explore more effective patterns, the proposed STP pretraining paradigm predicts future EEG signals as well as frequency components in a next-token-prediction setting, which requires a deeper understanding of the EEG signals and their semantic relationships, allowing for more accurate decoding performance.

## 2.2 Time Series Prediction

In EEG decoding, where temporal correlations are complex, training models on future sequence prediction has the potential to help them capture long-range dependencies and multi-frequency components, thereby enhancing the encoding of neural patterns. For a broader view, time series prediction has evolved from classical statistical methods like ARIMA [47] and Holt-Winter [48] to deep learning models, such as TCN [49], N-BEATS [50], and DeepAR [51]. While early methods struggled with non-linear dependencies, deep learning models captured richer temporal patterns. However, these models were still limited by task-specific training without generalizability. Transformer-based models, such as Temporal Fusion Transformer (TFT) [52], Informer [53], and PatchTST [54], are widely used to model long-range dependencies. TFT integrates attention mechanisms for multi-horizon forecasting, Informer enhances efficiency with sparse self-attention, and PatchTST uses patch-based tokenization for better long-term dependency modeling. Additionally, self-supervised pretraining has improved generalization in forecasting models like TimesFM[55] and MOIRAI[56, 57], which model future states for stronger generalization.

In our work, we build upon this concept by training our LBLM backbone model to predict future time series in the EEG domain, thereby enhancing its representational power. However, unlike existing time-series prediction models, we adopt an autoregressive GPT-style architecture for LBLM backbone. In this setup, future segments are predicted as a series of tokens, each representing a prediction for a specific time window, rather than generating the entire multi-variable time series with a single prediction head. This design also offers better flexibility for downstream classification tasks.

# 3 Method

In this section, we present the proposed training paradigm and the detailed backbone architecture for LBLM. We consider the input EEG data to be a multivariate time series with $L$ timesteps: $\mathbf{x} = (x_1, \cdots, x_L)$, where each $x_t$ at time step $t$ is a $M$-dimensional vector representing signals from $M$ channels. To learn effective EEG representations useful for silent speech decoding, we construct our EEG backbone with Layer-Gated Conformer blocks and pretrain our backbone via two concecuitve self-superivsed learning stages.

As shown in Figure 2, we first pretrain the backbone using masked EEG modeling (MEM) method, where the model learns contextualized representations by reconstructing randomly masked-out EEG signals. We then refine the backbone through STP pretraining, an auto-regressive multiview prediction task, enforcing the model to anticipate future EEG states in both temporal and spectral domain from more meaningful representations. Finally, we finetune the pretrained backbone with a spatio-temporal classification head for both semantic- and word-level classification tasks.

## 3.1 EEG input patching

The aim of our LBLM model backbone is to understand the temporal dynamics of EEG data with non-stationary nature. To avoid the information bottlenecks introduced by Vector Quantized (VQ) encoders [43, 44], we use channel-wise EEG patches as input tokens to the backbone. In particular, we divide EEG input $\mathbf{x}$ into a sequence of overlapping patches $\mathbf{x}_p$ with the patch length of $P$ and an overlapping stride of $S$. Consequently, each input EEG channel will be segmented into $N = \lfloor \frac{L-P}{S} + 1 \rfloor$ patches. This patching strategy enables the model to capture local temporal patterns within each patch while maintaining sequential correlations across consecutive patches, thereby supporting a more expressive representation of its dynamics

## 3.2 LBLM Backbone

## 3.3 Temporal and Subject Embedding

We enhance the backbone's ability to capture the inherent temporal order of patches and subject-specific information by incorporating temporal and subject embeddings into the input tokens. For sequential ordering, we add the temporal embedding to each token, allowing the model to encode temporal patterns across the sequence. This addition ensures that the model can differentiate tokens based on their position in time. For subject embedding, we multiply the embedding with each token, enabling the model to capture subject-specific characteristics by modulating the token's representation. This approach helps the model adapt to individual subject differences while preserving temporal dependencies within EEG data.

## 3.4 Layer-Gated Conformer Backbone

Our backbone model map the input patch embeddings to a latent space of dimension $d$ through Layer-Gated Conformer (LGConformer) blocks. We design the LGConformer block by adding a layer-gating connection from the input token to the output token of the conformer block[58]. Each conformer block contains four modules, two feed-forward layer, a convolution module, and a multi-head self-attention (MHSA) layer. The MHSA layer transforms the input tokens $\mathbf{x}_p$ into output tokens $\mathbf{e}_p$ by the following process:

$$\mathbf{e}_p = \text{Softmax}\left(\frac{\mathbf{x}_p^\top W^Q (\mathbf{x}_p^\top W^K)^\top}{\sqrt{d}}\right)\mathbf{x}_p^\top W^V \quad (1)$$

where linear layer projections $W^K$, $W^V$, and $W^Q$ transform the input patches into key, value, and query matrices. The normalization term $\sqrt{d}$ is used to scale the dot product in the attention mechanism. The convolution module (presented in Figure 8), comprised of two pointwise convolution layers and a depthwise convolution layer. The use of a convolution module inside a Transformer block allows the conformer block to capture local patterns more effectively within each EEG token.

To improve training stability, we augment the Conformer block with a layer-gating mechanism that adaptively regulates information flow across layers. Specifically, we connect the input token

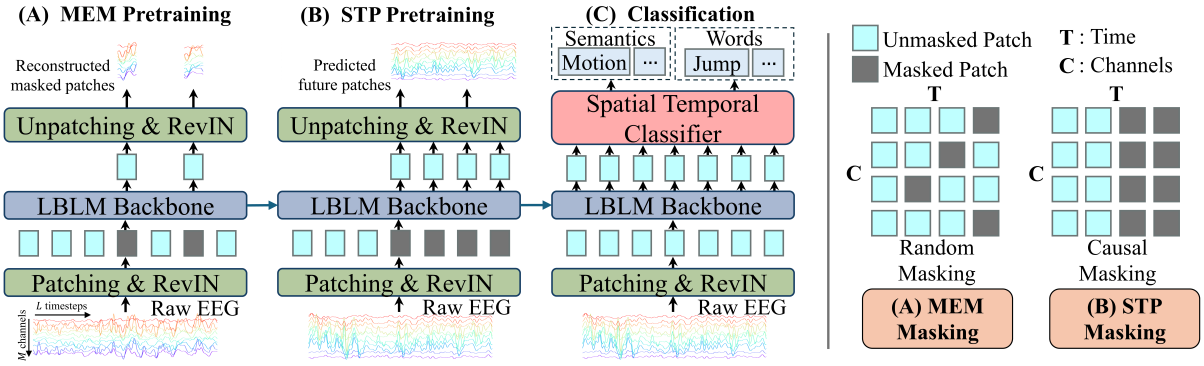



**Figure 2: Overview of the proposed self-supervised pretraining framework for the LBLM, which consists of three stages (A, B, and C). In Stage A, masked EEG modeling (MEM) pretraining is used to warm up the model weights by reconstructing randomly masked EEG patches. In Stage B, future EEG patches are completely masked out to help the model learn non-trivial representations. In Stage C, a spatio-temporal classifier is added to aggregate and refine the representations learned by the backbone model. We finetune the whole model for effective semantic-level and word-level classification tasks.**

from the previous layer $\mathbf{e}^{l-1}$ with the output token $\mathbf{e}^{l}$ after the Conformer transformation. The updated layer representation is computed as:

$$\mathbf{e}^{l} = g(\mathbf{e}^{l-1})\mathbf{e}^{l} + (1 - g(\mathbf{e}^{l-1}))\mathbf{e}^{l-1} \tag{2}$$

Here, $g(\mathbf{e}^{l-1})$ is a gating function implemented by a submodule comprising a zero convolution layer, a feed-forward layer, and a sigmoid function, which generates token-wise gating values between 0 and 1. The zero convolution layer follows the design from [59] where we use a $1 \times 1$ convolution with both weight and bias initialized to zero. This mechanism ensures a more stable update of the early layer during the beginning of training. In later training stages, it also allows information from the previous layers to be progressively blended to the next layer in a controlled manner.

During pretraining, each EEG channel is processed independently, improving data efficiency and enhancing generalization. On the other hand, the integration of channel-wise information will be performed by our spatio-temporal classifier in Section 3.6.

## 3.5 Spectro-Temporal Predictive Pretraining

We propose STP prtraining to enhance upon MEM and fully leverage unlabeled recordings from silent speech experiments. Recent self-supervised EEG approaches [43**?** ] only train a backbone model to reconstruct the masked encoded tokens. However, simply masking contiguous regions often leads to trivial interpolation or averaging from adjacent inputs due to the strong temporal continuity of EEG signals and overlapping window strides. Such shortcuts undermine the high-level feature extraction needed for robust EEG understanding. To overcome this limitation, our approach combines masking with a time-series forecasting objective. Given a lookback window of length $L$, the model must predict the subsequent timesteps (see Figure 4b for sliding window sampling). By explicitly forcing the network to anticipate future EEG segments, STP discourages naive interpolation and encourages deeper modeling of temporal dependencies, leading to richer spatiotemporal representations for downstream silent speech decoding. During training, we use the Huber loss [60] to reduce sensitivity to noise and sudden changes of amplitude in EEG signals, which is computed as follows:

$$L_H(y, \hat{y}, \delta) = \begin{cases} \frac{1}{2}(\mathbf{x}_p - \hat{\mathbf{x}_p})^2 & \text{if } |\mathbf{x}_p - \hat{\mathbf{x}_p}| \le \delta, \\ \delta\left(|\mathbf{x}_p - \hat{\mathbf{x}_p}| - \frac{1}{2}\delta\right) & \text{if } |\mathbf{x}_p - \hat{\mathbf{x}_p}| > \delta \end{cases} \tag{3}$$

where $\mathbf{x}_p$ and $\hat{\mathbf{x}_p}$ denot the ground-truth and predicted patch values, respectively, and $\delta$ is a hyperparameter that controls the transition between quadratic and linear loss terms. Beyond capturing temporal dependencies in the raw EEG waveform, we also train the backbone to reconstruct frequency amplitude and phase (as shown in Figure 4a), thereby encouraging a richer spectro-temporal understanding of the signal. This design aims to enhance the model's capacity for long-range dependency modeling and improve generalization, leading to the following pretraining loss function:

$$L_{total} = L_H^w + \lambda_1 L_H^a + \lambda_2 L_H^p \tag{4}$$

Here, $L_H^w$, $L_H^a$, and $L_H^p$ computed from the residuals between the actual and predicted waveforms, amplitude spectrums, and phase spectrums, respectively. The amplitude and phase losses, $L_a$, $L_p$, each have weighting coefficients $\lambda_1$, $\lambda_2$ that balance their contributions within the total objective. Because amplitude and phase jointly characterize the underlying EEG signal, optimizing them together leads to a richer representation. Consequently, instead of competing with each other, these losses reinforce each other by capturing complementary information of the data, guiding the training process toward better convergence.

## 3.6 Spatio-Temporal Classifier

For the downstream classification task, we design a classifier that leverages the feature learned by the LBLM backbone. Drawing from prior work on EEG decoding [61–65], we incorporate multi-scale temporal filtering and spatial aggregation as essential components for capturing spatio-temporal patterns. As illustrated in Figure 3, our Spatio-Temporal Classifier begins with a spatial convolution

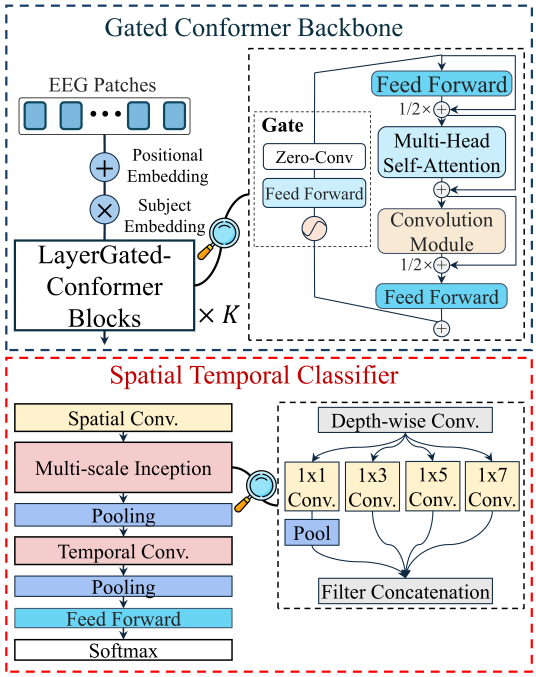
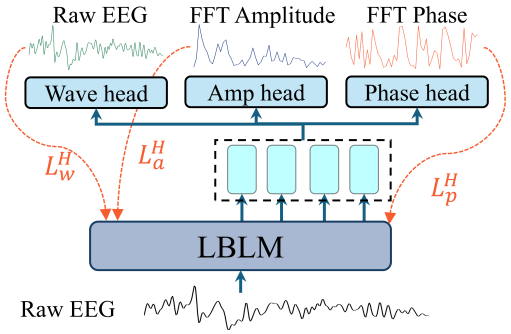
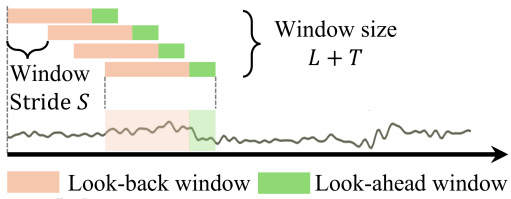



**Figure 3: Detailed architecture of the proposed LBLM model. Our LBLM model consists of a Layer-Gated Conformer (LGConformer) backbone for EEG feature extraction and a Spatio-Temporal (ST) classifier for pattern recognition. EEG signals are first segmented into overlapping patches, with positional embeddings added to encode temporal order and subject embeddings multiplied to incorporate subject-specific variations. The gated mechanism, a specialized skip connection, adaptively regulates information flow by controlling how much newly processed information is integrated with the input token. This helps prevent early collapse, stabilizes training, and enables more effective selection of input features for classification. The ST classifier integrates features across EEG channels and extracts multi-scale temporal patterns for improved classification performance. The LGConformer backbone is pretrained independently, while the ST classifier is trained during fine-tuning to refine the learned representations for downstream tasks.**

that aggregates information across all EEG channels to extract topographical patterns relevant to silent speech. Then, two temporal modules: a multi-scale inception block and a temporal convolution block, is used to capture both short- and long-range temporal dependencies. The multi-scale inception block enhances subject adaptability by processing short-term features at multiple receptive field sizes simultaneously, using convolutional kernels of size $1 \times 1$, $1 \times 3$, $1 \times 5$, and $1 \times 7$. The subsequent temporal convolution block integrates longer-term temporal context by further combining and refining these features. Finally, all representations are pooled and passed through a feedforward layer with a softmax activation to generate classification probabilities.




**(a) Illustration of the STP pretraining.**




**(b) Illustration of the time window segment strategy.**

**Figure 4: Figure (a) The backbone model predicts the Fourier amplitude and phase in addition to EEG waves during the pretraining phases. Figure (b) denotes the ways of processing time-series signals into window slides.**

## 4 Experiments

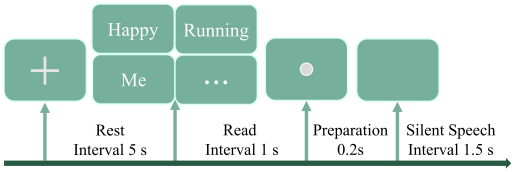



**Figure 5: Experiment design. Each trial consisted of four segments: rest, read, preparation, and silent speech. The number in between indicate the duration of each segment.**

### 4.1 Data Collection

We recruited 12 healthy individuals (S1-S10) for our silent speech experiment. The experimental protocol is shown in Figure 5 and consists of four segments: rest, read, preparation, and silent speech production. The rest segment lasted for 5 seconds, showing an eye fixation cross (+) 1.5 seconds before the word cue appear. In the read segment, the participants were presented with a word cue and were asked to read the word presented on the screen. The read segments last for 1 second. At the end of the read segment, an audio cue signals the transition to the silent speech segment for a brief preparation period of 0.2 seconds. Afterwards, participants were given 1.5 seconds to silently produce the word. To reduce fatigue, participants are provided with extended breaks after every 20 trials. Additional details regarding data collection procedures, equipment,

preprocessing and analysis can be found in Appendix C. Altogether, we have collected 6000 trials from each participant with 16 EEG sessions. The duration for an entire session last for approximately 40 minutes and the duration of the whole dataset is around 120 hours in total.

## 4.2 Training and Evaluation

*Pretraining and Finetuning.* We divided the data from each subject into training, validation, and testing splits using a cross-session setup. Specifically, out of the 16 EEG recording sessions for each subject, 2 sessions were held out from the training data used in the pretraining and finetuning stages. For the classification task, one held-out session was used for validation and model selection, while the other was reserved exclusively for testing, and all reported results are based on the test session. In total, we utilized approximately 105 hours of unlabeled EEG recordings for the pretraining stage, and over 5,200 labeled silent speech segments for finetuning.

During pretraining, we train our model on the MEM and STP tasks outlined in Section 3.5. Reversible normalization (RevIN) is applied to normalize the input data and accurately reconstruct predictions in the original scale, while multi-band dataset mixing is used to enlarge the training set and enhance diversity during training.(See Appendix B.1 and B.2 for more details). For finetuning on semantic- and word-level classification, we employ cross-entropy loss, which is described in Appendix B.3. Implementation details as well as hyperparameter settings are described in Appendix A.

*Baselines and Metrics.* We compare our LBLM model against several widely used EEG classification models, including EEGNet [65], TCNet [62], EEGConformer [66], and STTransformer [67]. These baselines represent convolutional and Transformer architecture widely used to capture spatio-temporal features in EEG data. In addition to these external baselines, we perform internal comparisons focusing on the backbone architecture. Specifically, we compare our LayerGatedConformer backbone with Transformer ($\text{LBLM}_T$) and Conformer ($\text{LBLM}_C$) backbones. For fair comparisons, all models are trained and evaluated using the same data as our LBLM.

# 5 Results

## 5.1 Classification Performance

We evaluate the classification performance of the proposed EEG backbone model on two classification tasks: word-level classification and semantic-group classification. Results are shown in Table 1 and Table 2.

As shown in Table 1, our proposed LBLM model outperform all baselines and achieve the state-of-the-art average accuracy of 39.6% and 42.8% on the word-level and semantic-level classification tasks respectively. Especially in the more challenging word-level classification task, our model achieves a more significant improvement of +7.3% compared to the best baseline TCNet model. This findings align with previous works on pretrained EEG backbone that pretrainining yields highter accuracy improvements on more finegrained classification tasks [43]. In our own LBLM model, we observed that both MEM and STP pretraining brings improvements to the model's performance, where after MEM pretraining, the average accuracy is increased by 1.6% and 1.5% on both tasks respectively. On the other hand, further pretraining on the proposed STP method further enhancing the the averaged accuracy to 39.6% (+2.7%) and 42.8% (+2.9). Notable improvements can be observed on subject S09 and S10, where our pretraining paradigm increases accuracy from 47.8% to 66.2% (S09) and from 17.7% to 32.2% (S10). We attribute the improvement to the rick spatio-temporal representation learned by the SPT pretraining, which forces the LayerGatedConformer backbone to learn non-trivial transformation of the input EEG signals. Moreover, comparisons with ablated backbone architectures ($\text{LBLM}_T$ and $\text{LBLM}_C$) further support the effectiveness of the gating mechanism in the LayerGatedConformer block. In addition to improving training stability, the gating mechanism also facilitates feature selection across backbone layers during finetuning.

Compared to baseline convolutional and Transformer models, we find that larger Transformer architectures like EEGConformer and STTransformer struggle to generalize when finetuned on limited subject-specific data. In contrast, CNN-based models such as EEGNet and TCNet, which require fewer parameters, are easier to optimize and perform better in this setting. For example, TCNet achieves average accuracies of 32.3% and 37.3%, while EEGConformer reaches only 26.9% and 31.6%. Despite this advantage, CNNs are inherently limited in scalability due to rigid architectural constraints. Our LBLM model, by contrast, handles data scarcity more effectively and offers better scalability than other Transformer-based models.

## 5.2 Ablation Analysis

*Ablation on Pretraining Loss Selection.* We investigate how different loss components affect our model's performance on both pretraining tasks and downstream word-level classification. As shown in Table 5, incorporating the amplitude loss term ($L^aH$) and the phase loss term ($L^pH$) leads to increased error in masked reconstruction and future prediction tasks. Notably, the MSE for future prediction task in STP pretraining task rises from 0.23 to 0.34. This suggests that emphasizing amplitude components introduces additional complexity, potentially making the prediction of future EEG signals more difficult. However, despite the increase in reconstruction error, we observe consistent improvements in word-level classification accuracy as these terms are added. This indicates that learning frequency-related features, especially amplitude and phase dynamics, is beneficial for extracting more discriminative EEG representations, ultimately leading to better classification performance.

*Ablation on Downstream Classifier.* To evaluate the generalizability of our pretrained LBLM backbone across different classifiers, we compare our spatio-temporal (ST) classifier with alternatives used in prior work: a linear classifier (Pooling + FFN) from EEGConformer/STTransformer and a convolutional classifier (Depthwise + Separable Conv + FFN) from EEGNet. As shown in Table 4, pretraining consistently improves classification performance across all classifiers. The linear classifier shows only a marginal gain (21.93% vs. 21.19%), likely due to the lack of spatial modeling, which limits its ability to leverage the learned representations for silent speech recognition. In contrast, the convolutional classifier benefits more noticeably, improving from 33.81% to 35.45%. These results demonstrate that while different classifiers benefit to varying degrees,

**Table 1: Results on word-level classification**

| Model | Param. | S01 | S02 | S03 | S04 | S05 | S06 | S07 | S08 | S09 | S10 | S11 | S12 | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EEGNet | 8.54 K | 40.8 | 35.3 | 33.5 | **22.9** | 17.7 | 10.6 | 49.6 | 14.0 | 38.7 | 24.7 | 37.4 | 41.3 | 30.5 |
| TCNet | 78.62 K | 36.9 | 36.4 | 32.7 | 22.3 | 14.3 | 13.5 | 44.4 | **16.4** | 50.4 | 25.5 | 46.8 | 48.1 | 32.3 |
| EEGConformer | 0.75 M | 58.2 | 32.7 | 27.5 | 12.2 | 11.2 | 7.0 | 30.9 | 7.0 | 23.9 | 28.3 | 35.1 | 48.3 | 26.9 |
| STTransformer | 2.78 M | 60.3 | 16.4 | 30.9 | 21.3 | 13.2 | **17.4** | 35.1 | 15.6 | 31.2 | 23.4 | 46.2 | 53.8 | 30.4 |
| $LBLM_T$ | 22.36 M | 61.3 | 11.4 | 17.9 | 11.7 | 12.5 | 9.4 | 33.8 | 10.1 | 30.9 | 14.8 | 43.0 | 49.0 | 25.5 |
| $LBLM_C$ | 22.61 M | **76.6** | 18.4 | 31.4 | 19.2 | 13.8 | 9.9 | 42.6 | 10.1 | 33.5 | 15.0 | 48.6 | 56.4 | 31.3 |
| LBLM | 22.61 M | 75.8 | 33.5 | **36.1** | 22.3 | 16.9 | 11.9 | 47.8 | 10.1 | 47.8 | 17.7 | 47.8 | 55.6 | 35.3 |
| w/ MEM | 22.61 M | 76.1 | 35.6 | 31.7 | 21.0 | 14.5 | 10.4 | 42.9 | 10.1 | 65.5 | 24.2 | 50.6 | 59.7 | 36.9 |
| w/ MEM + STP | 22.61 M | 73.8 | **37.1** | 35.8 | 21.3 | **18.7** | 11.7 | **49.9** | 12.7 | **66.2** | **32.2** | **52.7** | **62.9** | **39.6** |

**Table 2: Results on semantic-level classification**

| Model | Param. | S01 | S02 | S03 | S04 | S05 | S06 | S07 | S08 | S09 | S10 | S11 | S12 | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EEGNet | 8.54 K | 69.4 | **36.4** | **38.2** | 31.4 | 23.6 | 33.2 | 46.5 | 30.9 | 18.4 | **33.5** | 54.3 | 69.4 | 36.2 |
| TCNet | 78.62 K | 77.9 | 30.6 | 35.6 | **31.9** | 22.9 | 30.4 | 56.6 | **33.5** | 23.4 | 30.3 | 56.9 | 69.4 | 37.3 |
| EEGConformer | 0.75 M | 66.2 | 26.2 | 35.3 | 30.4 | 22.3 | 22.9 | 37.9 | 23.9 | 21.6 | 28.8 | 34.5 | 57.9 | 31.6 |
| STTransformer | 2.78 M | 65.2 | 25.7 | 27.0 | 28.8 | 20.0 | 22.3 | 36.9 | 26.2 | 39.7 | 27.3 | 57.9 | 74.0 | 31.9 |
| $LBLM_T$ | 22.36 M | 74.3 | 26.8 | 33.8 | 28.6 | 24.4 | 26.8 | 41.8 | 23.9 | 50.6 | 27.3 | 54.0 | 65.5 | 35.8 |
| $LBLM_C$ | 22.61 M | 76.1 | 27.8 | 35.1 | 29.1 | 25.7 | 26.2 | 53.2 | 25.5 | 65.7 | 27.8 | 54.0 | 66.0 | 39.2 |
| LBLM | 22.61 M | 75.8 | 27.8 | 34.8 | 28.3 | 24.9 | 27.3 | 47.0 | 25.7 | 62.1 | 30.6 | 57.4 | 64.4 | 38.4 |
| w/ MEM | 22.61 M | 76.1 | 27.8 | 36.1 | 28.8 | 26.0 | 29.6 | 47.3 | 26.0 | 72.2 | 29.4 | 57.1 | 81.3 | 39.9 |
| w/ MEM + STP | 22.61 M | **79.5** | 28.6 | 36.4 | 29.6 | **26.5** | **33.2** | **59.5** | 27.0 | **76.4** | 31.2 | **58.4** | **77.7** | **42.8** |

**Table 3: Ablation on Training Settings**

| Training | Masked Reconstruction | | Future Prediction | | Word Level |
|---|---|---|---|---|---|
| Loss | MSE | MAE | MSE | MAE | Acc. (%) |
| $L_H^w$ | 0.14 | 0.22 | 0.23 | 0.28 | 37.4 |
| $+L_H^a$ | 0.15 | 0.24 | 0.34 | 0.30 | 38.5 |
| $+L_H^a+L_H^p$ | 0.15 | 0.24 | 0.27 | 0.32 | 39.6 |

**Table 4: Word-level ablation results with different classifiers on our pretrained backbone.**

| Classifier | Training | Avg. Acc (%) |
|---|---|---|
| Linear | w/o Pretrain | 21.19 |
| Linear | w/ Pretrain | 21.93 |
| Convolution | w/o Pretrain | 33.81 |
| Convolution | w/ Pretrain | 35.45 |
| Spatio-Temporal (Ours) | w/o Pretrain | 35.28 |
| Spatio-Temporal (Ours) | w/ Pretrain | 39.59 |

the pretrained LBLM backbone consistently enhances downstream performance, with the greatest gains achieved when paired with our ST classifier.

## 5.3 EEG signal perdiction

Different from existing end-to-end trained EEG classifiers or existing EEG models pretrained via masked token reconstruction (e.g., labram), our pretrain model additional has the ability to reconstruct missing EEG segment or recover detached channels, or anticipate user's active speaking intent when predicting EEG states in the future. We compare the performance of the proposd LBLM after STP pretraining with a strong baseline, PatchTST [54], which is commonly adopted for long-term time series forecasting. The evaluation metrics include Mean Squared Error (MSE) and Mean Absolute Error (MAE). The results are summarized in Table 5.

**Table 5: Cross-Session EEG forecasting Performance**

| Prediction Length | GatedConformer MSE | GatedConformer MAE | PatchTST MSE | PatchTST MAE |
|---|---|---|---|---|
| 46(204) | 0.27 | 0.32 | 1.01 | 0.65 |
| 70(180) | 0.33 | 0.36 | 1.05 | 0.66 |
| 94(156) | 0.45 | 0.42 | 1.11 | 0.68 |
| 118(132) | 0.50 | 0.44 | 1.21 | 0.71 |
| 136(114) | 0.54 | 0.46 | 1.23 | 0.70 |

Table 5 reports the quantitative results of cross-session EEG forecasting. The proposed GatedConformer consistently outperforms PatchTST across all prediction lengths. Specifically:

- For a prediction length of 46 with a context window of 204, GatedConformer achieves an MSE of 0.27 and an MAE of 0.32, while PatchTST obtains an MSE of 1.01 and an MAE of 0.65.
- As the prediction length increases to 70, GatedConformer maintains superior performance with an MSE of 0.33 and an MAE of 0.36, compared to PatchTST's MSE of 1.05 and MAE of 0.66.
- At longer prediction horizons, the performance gap becomes more pronounced. For 94-step forecasting, GatedConformer yields an MSE of 0.45 and an MAE of 0.42, while PatchTST records an MSE of 1.11 and an MAE of 0.68.
- For the longest forecasting scenario, with 136 predicted steps and a 114-step look-back window, GatedConformer attains an MSE of 0.54 and an MAE of 0.46. In contrast, PatchTST reaches an MSE of 1.23 and an MAE of 0.70.

Figure 6 and Figure 7 illustrate the predicted EEG time series from our backbone model. The results demonstrate that the model effectively captures both the overall trend and the fine-grained oscillatory patterns of the EEG signal during silent speech recording. This is evident in its ability to reconstruct both peak amplitudes and high-frequency fluctuations.

For future patch prediction (Figure 7), the model achieves higher precision when forecasting EEG waves within 150 ms of the look-back window. However, for time patches beyond this range ( > 150 ms), the model struggles to accurately capture both the trend and oscillatory components. This limitation is reasonable, as EEG signals exhibit significant variability due to external stimuli and internal cognitive states, especially in silent speech experiments.

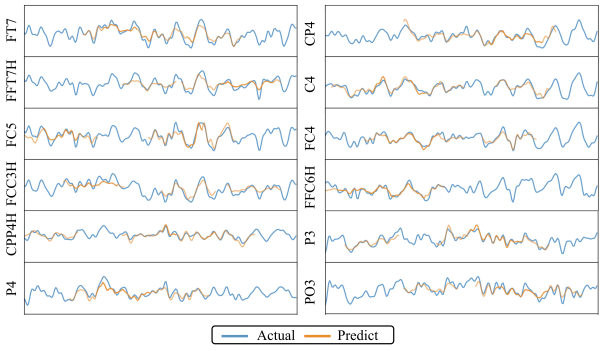



**Figure 6: EEG signal masked reconstruction visualization of a few channels. The blue curves are the actual data. The curves before the black dashed lines are the actual input to the model, and the orange curves are the predicted EEG signal.**

## 6 Conclusion

This paper introduces a novel self-supervised pretraining method and presents the Large Brain Language Model (LBLM), a 22-million-parameter model for silent speech decoding in active BCI. To support this, we collected a large-scale EEG dataset comprising over 120 hours of recordings. Our proposed Spectro-Temporal Predictive

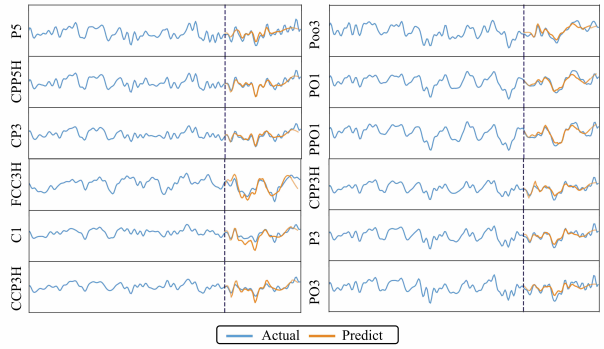



**Figure 7: EEG signal forecasting visualization of a few channels from the same EEG segment. The blue curves are the actual data. The curves before the black dashed lines are the actual input to the model, and the orange curves are the predicted EEG signal.**

(STP) pretraining method learns temporal and spectral EEG components in an autoregressive manner, enhancing the model's ability to capture the spatiotemporal dynamics of brain activity. The LBLM backbone is built by LayerGatedConformer blocks, paired with a spatiotemporal classifier to extract and integrate the most relevant features for the downstream classification tasks. Experimental results demonstrate that our approach consistently outperforms CNN- and transformer-based baselines in both word-level and semantic-group classification, particularly under cross-session protocols. Additionally, the model's ability to reconstruct missing signals and anticipate future EEG patterns enhances data reliability and real-time applicability. These findings highlight the effectiveness of the proposed next-patch prediction method for self-supervised pretraining. Our future work will focus on expanding the vocabulary size, improving cross-subject generalization, and further optimizing real-time integration to advance practical silent speech BCI systems.

## References

[1] He Pan, Peng Ding, Fan Wang, Tianwen Li, Lei Zhao, Wenya Nan, Yunfa Fu, and Anmin Gong. Comprehensive evaluation methods for translating bci into practical applications: usability, user satisfaction and usage of online bci systems. *Frontiers in Human Neuroscience*, 18:1429130, 2024.

[2] Jonathan R Wolpaw. Brain–computer interfaces. In *Handbook of clinical neurology*, volume 110, pages 67–74. Elsevier, 2013.

[3] Simanto Saha, Khondaker A Mamun, Khawza Ahmed, Raqibul Mostafa, Ganesh R Naik, Sam Darvishi, Ahsan H Khandoker, and Mathias Baumert. Progress in brain computer interface: Challenges and opportunities. *Frontiers in systems neuroscience*, 15:578875, 2021.

[4] Solène Le Bars, Sylvie Chokron, Rodrigo Balp, Khalida Douibi, and Florian Waszak. Theoretical perspective on an ideomotor brain-computer interface: toward a naturalistic and non-invasive brain-computer interface paradigm based on action-effect representation. *Frontiers in Human Neuroscience*, 15:732764, 2021.

[5] Brent J Lance, Scott E Kerick, Anthony J Ries, Kelvin S Oie, and Kaleb McDowell. Brain–computer interface technologies in the coming decades. *Proceedings of the IEEE*, 100(Special Centennial Issue):1585–1599, 2012.

[6] Jan Van Erp, Fabien Lotte, and Michael Tangermann. Brain-computer interfaces: beyond medical applications. *Computer*, 45(4):26–34, 2012.

[7] Arrigo Palumbo, Vera Gramigna, Barbara Calabrese, and Nicola Ielpo. Motor-imagery eeg-based bcis in wheelchair movement and control: A systematic literature review. *Sensors*, 21(18):6285, 2021.

[8] Gert Pfurtscheller and Christa Neuper. Motor imagery and direct brain-computer communication. *Proceedings of the IEEE*, 89(7):1123–1134, 2001.

[9] 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. Imagined speech can be decoded from low-and cross-frequency intracranial eeg features. *Nature communications*, 13(1):48, 2022.

[10] Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012.

[11] Sarah A Graham and Simon E Fisher. Decoding the genetics of speech and language. *Current opinion in neurobiology*, 23(1):43–51, 2013.

[12] 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.

[13] Yiqun Duan, Jinzhao 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.

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

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

[16] 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.

[17] 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.

[18] Aurélie de Borman, Benjamin Wittevrongel, Ine Dauwe, Evelien Carrette, Alfred Meurs, Dirk Van Roost, Paul Boon, and Marc M Van Hulle. Imagined speech event detection from electrocorticography and its transfer between speech modes and subjects. *communications biology*, 7(1):818, 2024.

[19] Kathrin Müsch, Kevin Himberger, Kean Ming Tan, Taufik A Valiante, and Christopher J Honey. Transformation of speech sequences in human sensorimotor circuits. *Proceedings of the National Academy of Sciences*, 117(6):3203–3213, 2020.

[20] Connie Cheung, Liberty S Hamilton, Keith Johnson, and Edward F Chang. The auditory representation of speech sounds in human motor cortex. *elife*, 5:e12577, 2016.

[21] Michael P Broderick, Andrew J Anderson, Giovanni M Di Liberto, Michael J Crosse, and Edmund C Lalor. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, 28(5):803–809, 2018.

[22] Jonathan R Brennan and John T Hale. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PloS one*, 14(1):e0207741, 2019.

[23] 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.

[24] Charles S DaSalla, Hiroyuki Kambara, Makoto Sato, and Yasuharu Koike. Single-trial classification of vowel speech imagery using common spatial patterns. *Neural networks*, 22(9):1334–1339, 2009.

[25] Nicolás Nieto, Victoria Peterson, Hugo Leonardo Rufiner, Juan Esteban Kamienkowski, and Ruben Spies. Thinking out loud, an open-access eeg-based bci dataset for inner speech recognition. *Scientific Data*, 9(1):52, 2022.

[26] Shunan Zhao and Frank Rudzicz. Classifying phonological categories in imagined and articulated speech. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 992–996. IEEE, 2015.

[27] Jinzhao Zhou, Yiqun Duan, Fred Chang, Thomas Do, Yu-Kai Wang, and Chin-Teng Lin. Belt-2: Bootstrapping eeg-to-language representation alignment for multi-task brain decoding. *arXiv preprint arXiv:2409.00121*, 2024.

[28] Kostas et al. Bendr: using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

[29] Qiushi Zhu, Xiaoying Zhao, Jie Zhang, Yu Gu, Chao Weng, and Yuchen Hu. Eeg2vec: Self-supervised electroencephalographic representation learning. *arXiv preprint arXiv:2305.13957*, 2023.

[30] Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. Uncovering the structure of clinical eeg signals with self-supervised learning. *Journal of Neural Engineering*, 18(4):046020, 2021.

[31] Joseph Y Cheng, Hanlin Goh, Kaan Dogrusoz, Oncel Tuzel, and Erdrin Azemi. Subject-aware contrastive learning for biosignals. *arXiv preprint arXiv:2007.04871*, 2020.

[32] Navid Mohammadi Foumani, Geoffrey Mackellar, Soheila Ghane, Saad Irtza, Nam Nguyen, and Mahsa Salehi. Eeg2rep: enhancing self-supervised eeg representation through informative masked inputs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5544–5555, 2024.

[33] Dustin E Stansbury, Thomas Naselaris, and Jack L Gallant. Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron*, 79(5):1025–1034, 2013.

[34] Diego Lopez-Bernal, David Balderas, Pedro Ponce, and Arturo Molina. A state-of-the-art review of eeg-based imagined speech decoding. *Frontiers in human neuroscience*, 16:867281, 2022.

[35] Lea Tøttrup, Kasper Leerskov, Johannes Thorling Hadsund, Ernest Nlandu Kamavuako, Rasmus Leck Kæseler, and Mads Jochumsen. Decoding covert speech for intuitive control of brain-computer interfaces based on single-trial eeg: a feasibility study. In *2019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR)*, pages 689–693. IEEE, 2019.

[36] Siyi Deng, Ramesh Srinivasan, Tom Lappas, and Michael D'Zmura. Eeg classification of imagined syllable rhythm using hilbert spectrum methods. *Journal of neural engineering*, 7(4):046006, 2010.

[37] Bram van den Berg, Sander van Donkelaar, and Maryam Alimardani. Inner speech classification using eeg signals: A deep learning approach. In *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*, pages 1–4. IEEE, 2021.

[38] Koji Koizumi, Kazutaka Ueda, and Masayuki 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.

[39] Ciaran Cooney, Attila Korik, Raffaella Folli, and Damien Coyle. Evaluation of hyperparameter optimization in machine and deep learning methods for decoding imagined speech eeg. *Sensors*, 20(16):4629, 2020.

[40] 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, 2023.

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

[42] Jinzhao Zhou, Yiqun Duan, Ziyi Zhao, Yu-Cheng Chang, Yu-Kai Wang, Thomas Do, and Chin-Teng Lin. Towards linguistic neural representation learning and sentence retrieval from electroencephalogram recordings. In *Proceedings of the 1st International Workshop on Brain-Computer Interfaces (BCI) for Multimedia Understanding*, pages 19–28, 2024.

[43] Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*, 2024.

[44] Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[45] Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. Neurolm: A universal multi-task foundation model for bridging the gap between language and eeg signals. *arXiv preprint arXiv:2409.00101*, 2024.

[46] Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in neuroscience*, 10:196, 2016.

[47] Robert H Shumway, David S Stoffer, Robert H Shumway, and David S Stoffer. Arima models. *Time series analysis and its applications: with R examples*, pages 75–163, 2017.

[48] Chris Chatfield. The holt-winters forecasting procedure. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 27(3):264–279, 1978.

[49] Pradeep Hewage, Ardhendu Behera, Marcello Trovati, Ella Pereira, Morteza Ghahremani, Francesco Palmieri, and Yonghuai Liu. Temporal convolutional neural (tcn) network for an effective weather forecasting using time-series data from the local weather station. *Soft Computing*, 24:16453–16482, 2020.

[50] Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv:1905.10437*, 2019.

[51] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International journal of forecasting*, 36(3):1181–1191, 2020.

[52] Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021.

[53] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.

[54] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

[55] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.

[56] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. *arXiv preprint arXiv:2402.02592*, 2024.

[57] Xu Liu, Juncheng Liu, Gerald Woo, Taha Aksu, Yuxuan Liang, Roger Zimmermann, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Moirai-moe: Empowering time series foundation models with sparse mixture of experts.

*arXiv preprint arXiv:2410.10469*, 2024.

[58] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

[59] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

[60] Kaan Gokcesu and Hakan Gokcesu. Generalized huber loss for robust learning and its efficient minimization for a robust statistics. *arXiv preprint arXiv:2108.12627*, 2021.

[61] Ming-ai Li, Jian-fu Han, and Jin-fu Yang. Automatic feature extraction and fusion recognition of motor imagery eeg using multilevel multiscale cnn. *Medical & Biological Engineering & Computing*, 59(10):2037–2050, 2021.

[62] Thorir Mar Ingolfsson, Michael Hersche, Xiaying Wang, Nobuaki Kobayashi, Lukas Cavigelli, and Luca Benini. Eeg-tcnet: An accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2958–2965. IEEE, 2020.

[63] Michael Hersche, Tino Rellstab, Pasquale Davide Schiavone, Lukas Cavigelli, Luca Benini, and Abbas Rahimi. Fast and accurate multiclass inference for mi-bcis using large multiscale temporal and spectral features. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1690–1694. IEEE, 2018.

[64] Haixian Wang and Wenming Zheng. Local temporal common spatial patterns for robust single-trial eeg classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 16(2):131–139, 2008.

[65] 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.

[66] Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022.

[67] Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie. Transformer-based spatial-temporal feature learning for eeg decoding. *arXiv preprint arXiv:2106.11170*, 2021.

[68] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.

[69] Joanna Górecka and Przemysław Makiewicz. The dependence of electrode impedance on the number of performed eeg examinations. *Sensors*, 19(11):2608, 2019.

[70] Carlos Valle, Carolina Mendez-Orellana, Christian Herff, and Maria Rodriguez-Fernandez. Identification of perceived sentences using deep neural networks in eeg. *Journal of neural engineering*, 21(5):056044, 2024.

[71] Arnaud Delorme and Scott Makeig. Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods*, 134(1):9–21, 2004.

# Appendices

## A Implementation Details

### A.1 Hyperparameters

Our experiments are implemented using the PyTorch framework and executed on NVIDIA H100 GPUs. During the pretraining of the backbone model, we use a batch size of 1024 and an initial learning rate of 1e-3. Training is optimized using the LAMB optimizer [68], combined with a cosine annealing learning rate scheduler. We set the weight decay to 0.01 and apply gradient clipping to cap the gradient norm at 1.0 for stable training. For the Huber loss in pretraining, we $delta = 1$ to provide a balanced trade-off between MSE and MAE. We also set $\lambda_1 = \lambda_2 = 0.1$ for balancing the learning of the amplitude and phase components in the EEG signals. The BLBM backbone model is pretrain for 80 epochs on our dataset.

### A.2 LayerGatedConformer Implementation Details

The convolution module used in our LayerGatedConformer is displayed in Figure 8 which align with the conformer paper [58]. The

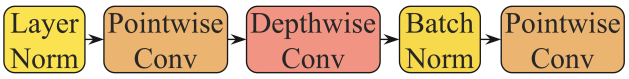



**Figure 8: The detailed structure of the convolution module in the LayerGatedConformer blocks.**

detailed setting of the BLBM backbone model is displayed in Table ??.

### A.3 Spatio-Temporal Classifier Implementation Details

## B Training Details

### B.1 Reversible Normalization

We consider using reversible instance normalization (RevIN) to stabilize the training process and avoid distortions to the original data distribution. Each EEG channel will be standardized independently across the time axis:

$$\hat{\mathbf{x}}_p^{(i)} = \frac{\mathbf{x}_p^{(i)} - \mu_p^{(i)}}{\sigma_p^{(i)}} \tag{5}$$

where $\mu_p^{(i)}$ and $\sigma_p^{(i)}$ denote the mean and variance of a specific EEG channel $i \in \{1, \cdots, M\}$. After model processing, the patch prediction output $\mathbf{y}$ will be de-normalized in a reverse process:

$$\hat{\mathbf{y}}_p^{(i)} = \mathbf{y} \cdot \sigma_p^{(i)} + \mu_p^{(i)} \tag{6}$$

which ensures the prediction output remains the same distribution as the input.

### B.2 Multi-band Data Mixing

Following recent trends in time-series modeling [55], we adopt a multi-band mixing strategy (Figure 9) to address the limited availability of EEG data. Specifically, each raw EEG waveform is decomposed via bandpass filters into alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–50 Hz) components, identified as key frequency bands by our rm-ANOVA analysis (Section C.3). We then generate augmented training samples by mixing the raw waveforms with these band-isolated signals, thereby increasing data diversity and reinforcing frequency-specific patterns. Empirically, this approach mitigates the data shortage common in EEG research and improves the time-series model's ability to capture both broadband and band-specific oscillatory features, ultimately enhancing overall model performance.

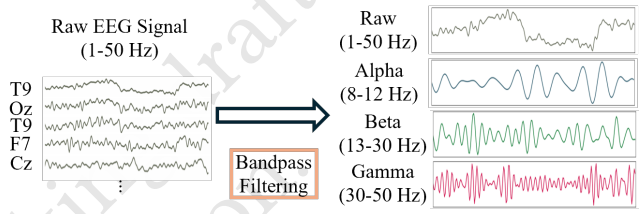



**Figure 9: Dataset mixing strategy used in this paper. We bandpass filtered the preprocessed raw input EEG data into different frequency bands. Allowing for more simple learning task to warm up the model training and increasing the diversity of the training dataset.**

### B.3 Supervised Finetuning for Classification

For the supervised finetuning, we train our classifier with the pretrained backbone using the cross-entropy loss as follow.

$$\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) = -\sum y_c \log(\hat{y}_c), \tag{7}$$

where $\hat{y}_c$ and $y_c$ denotes the groundtruth and the predicted word label from the EEG segment.

## C EEG data collection

Twelve right-handed native English speakers from Australia (6 males, 6 females, aged 28 ± 0.85 years old) with no visual or speech impairment or neurological disorders were participated in this study. They performed silent speech task during the experiment. The subjects were seated on a comfortable chair and were instructed to keep their eyes open while observing a computer monitor placed at a comfortable distance. Subjects were also asked to keep their body movement o minium during the experiment to minimise muscle. Informed consent was obtained in writing from each participant prior to the experiment. The EEG signals were recorded at a sampling rate of 1000Hz using a 128-channel neuroscan quik-cap[1] connected to a SynAmps RT 128-channel Amplifier[2]. The CURRY 9 software[3]

[1]https://au.neuromedicalsupplies.com/product/compumedics-neuroscan-quik-cap-128-ch-large-62-68cm-synamps2-agagcl-sintered-electrodes-2/
[2]https://compumedicsneuroscan.com/product/synamps-rt-128-channel-eeg-erp-ep-amplifier/
[3]https://compumedicsneuroscan.com/product/curry-9-x-data-acquisition-and-online-processing/

**Table 6: Summary of Data Collection Settings**

| Feature | Details |
|---|---|
| Subjects | 12 |
| Modality | EEG |
| Channel Number | 122 |
| Sessions Per Subject | 16 |
| Total Trials Per Subject | 6000 |
| Vocabulary Size | 24 |
| Total Hour Per Subject | 10 |

was used for data acuisition. The placement of electrodes follows the 10–10 international system[**?** ]. Impedance checks were conducted before each experiment to ensure that the impedance value of all electrodes remains below $10k\omega$ and balanced across electrodes [69, 70]. A ground electrode was referenced to maintain signal consistency.

## C.1 Experiment Design

Our silent speech experiment involves 24 commonly used words in English. The words are selected based on 6 semantic groups. Table 7 shows the 24 words and their corresponding semantic group. Since our goal is to facilitate silent speech decoding for active BCI research, we include a total of 6 semantic groups ranging from motion, emotion, location, person, object, and number to increase vocabulary diveristy. The participants were instructed to make articulation attempts without emitting sounds. Each word will be repeated 5 times and the sequence of the words is randomized for each session. The experiment was implemented in Python using Psychopy library. Statistical information of the collected dataset is shown in Table 6.

## C.2 EEG data preprocessing

An overview of the preprocessing pipeline is shown in Figure 10. Using our 128-channel neuroscan system, each EEG recording is saved in *.cdt* format and consists of 133 channels: 122 EEG channels capturing brain activity, 4 electrodes capturing eye movements, and 6 reference electrodes capturing muscle and environment noise, and 1 trigger channel for condition labeling. The EEG data is first band-pass filtered between 1 and 75Hz using a finite impulse response (FIR) filter to remove slow drifts and high-frequency noise, followed by a 50Hz notch filter to eliminate line interference. Then the EEG signals are re-referenced by calculating the average of the 122 channels. Next, ICA decomposition is performed to identify and remove artifact components such as eye movement, muscle activity, and cardiac artifacts, with a confidence threshold of 90% for ICA classification. Finally, the EEG data is segmented into epochs of 2 seconds with a 0.5-second overlap. When used for training and classification, the EEG data is downsampled to 250Hz to improve computational efficiency. Our preprocessing pipeline is implemented using the EEGLAB library [71].

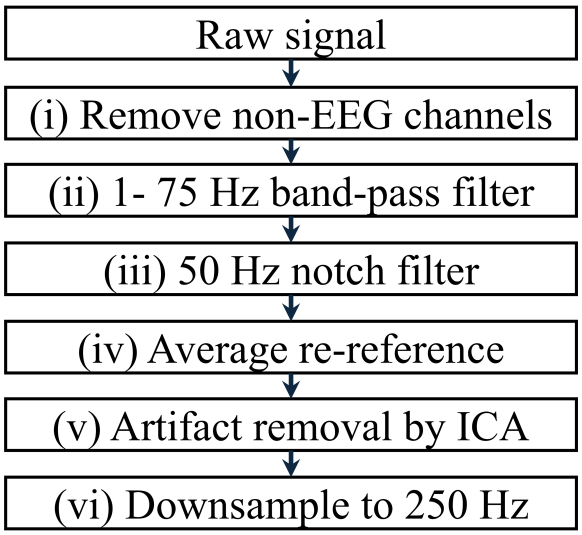



**Figure 10: Pre-processing steps for EEG data.**

## C.3 Correlation Analysis

To determine which frequency bands exhibited the greatest variation under different cognitive conditions, we conducted a repeated-measures analysis of variance (rm-ANOVA) on the power spectrum density (PSD) of delta (1 – 4Hz), theta (4 – 8Hz), alpha (8 – 13Hz), beta (13 – 30Hz), and gamma (30 – 50Hz). Three condition pairs were compared: rest versus reading, rest versus silent speech, and reading versus silent speech. We used F-scores to quantify differences among these conditions, with higher F-scores indicating larger variations in PSD. Figure 11(a) shows that in the delta band (1–4 Hz), strong frontal activation likely corresponds to eye-related artifacts, while the alpha band (8–13 Hz) reveals moderate differences centered around the occipital region, notably near the POOZ electrode. In contrast, the beta (13–30 Hz) and gamma (30–50 Hz) bands exhibit high F-scores in the left hemisphere, suggesting substantial neural variability that may be related to motor or cognitive processes. In Figure 11(b), similar frontal activations appear in the delta band, again pointing to eye movements or blinks. The alpha band, which overlaps with the mu rhythm, presents pronounced changes in both left and right temporal regions. Meanwhile, the beta and gamma bands show F-scores exceeding 10 in the occipital area, reflecting the visual engagement required for reading compared to resting states. Finally, Figure 11(c) highlights the beta and gamma bands in the left hemisphere as the main source of variation when comparing resting versus silent speech. The relative lack of occipital activation here contrasts with the reading condition, indicating reduced visual processing during silent speech and a stronger emphasis on cognitive or motor-planning regions. Based on these findings, alpha, beta, and gamma bands emerge as the most discriminative for silent speech decoding, justifying our focus on these frequency ranges in subsequent analyses. Altogether, these analysis results reveal that alpha, beta, and gamma bands contains particularly significant differences during silent speech

**Table 7: List of stimuli and their syllable number used in our experiment**

| Semantic Category | Word | Number of Syllables | Semantic Group | Word | Number of Syllables |
|---|---|---|---|---|---|
| Motion | Jumping | 2 | People | Mother | 2 |
| | Running | 2 | | Cowboy | 2 |
| | Swimming | 2 | | Professor | 3 |
| | Going | 2 | | Me | 1 |
| Emotion | Happy | 2 | Number | One | 1 |
| | Sad | 1 | | Three | 1 |
| | Fun | 1 | | Eleven | 3 |
| | Horrible | 3 | | Million | 2 |
| Location | College | 3 | Object | Spoon | 1 |
| | Home | 2 | | Alfa | 2 |
| | Battlefield | 3 | | Python | 2 |
| | Here | 2 | | Telephone | 3 |

activaties. This observation motivated our selection of these bands for multi-band data mixing in subsequent analyses.

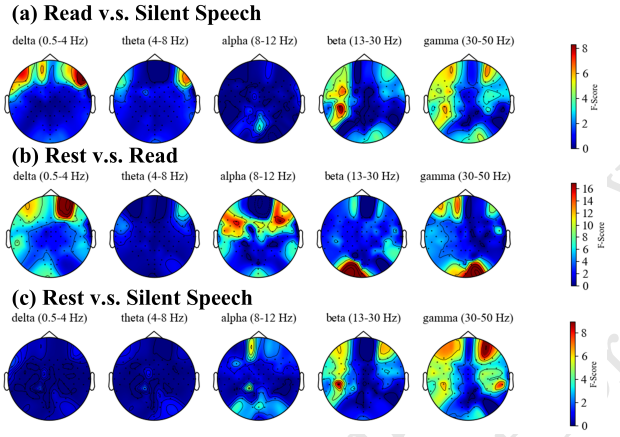



**Figure 11: rm-ANOVA analysis of paired conditions from all subject. (a) read vs. silent speech, (b) rest vs. read, (c) rest vs. silent speech. We visualize the F-score of the compared conditions.**

## D Visualizations

*Visualizations.* Visualization of the reconstruction performance from both tasks are shown in Figure 6 and 7. These results demonstrate that the proposed model can both fill in missing EEG information and anticipate upcoming neural activity with reasonable accuracy, thereby providing a richer signal representation that enhances decoding. In the masked reconstruction task (Figure 6), accurately inferring unseen segments yields more complete data for downstream analysis, reducing the detrimental impact of signal dropout or corruption on decoding performance. Meanwhile, the ability to forecast future EEG waveforms (Figure 7) enables systems to maintain responsiveness and continuity in scenarios where timely detection of neural events is critical. Overall, the model's demonstrated robustness in reconstruction and short-term prediction can strengthen decoding pipelines by offering a more stable and comprehensive view of ongoing brain activity.

Figures 6 and 7 illustrate how the proposed backbone model predicts EEG time series in both masked reconstruction and future patch prediction tasks. The results confirm that our model effectively captures overall trends and fine-grained oscillatory details, enabling it to reconstruct peak amplitudes as well as high-frequency fluctuations. In the masked reconstruction task (Figure 6), accurately inferring unseen segments helps counteract signal dropout or corruption, thereby providing higher-quality data for downstream analysis. Meanwhile, forecasting future EEG waveforms (Figure 7) helps maintain responsiveness by predicting upcoming neural activity in real time, an advantage for critical applications that depend on timely detection of brain states.

Although short-term predictions (within 150 ms of the look-back window) achieve higher precision, performance diminishes for forecasts beyond this range due to the inherently variable nature of EEG signals, particularly in silent speech contexts where slight cognitive shifts can substantially alter neural activity. Despite this limitation, the ability to fill in missing segments and anticipate near-future EEG changes yields a richer signal representation that strengthens decoding pipelines, offering a more stable and continuous view of ongoing brain activity.

Finally, Figure 12 shows the model's future predictions across alpha, beta, and gamma bands, offering a more granular view of its forecasting capabilities. In each sub-figure, the predicted waveforms (orange) closely track the actual signals (blue) within a short time horizon, successfully capturing both amplitude and finer oscillatory details. Notably, the alpha band exhibits smoother, lower-frequency fluctuations, while beta and gamma signals reveal higher-frequency patterns that the model still manages to approximate with reasonable accuracy. Although prediction quality gradually diminishes over longer forecast intervals—as inherent variability increases—these results highlight the model's ability to handle distinct spectral components, ultimately contributing to a more comprehensive and robust EEG decoding pipeline.

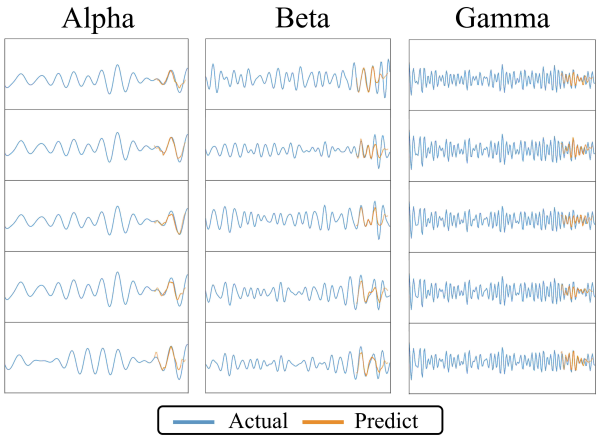



**Figure 12: EEG signal forecasting visualization of alpha, beta, and gamma band.**

**Table 8: List Of Variables in this Paper**

| | |
|---|---|
| Length of the input EEG data | $L$ |
| EEG data at a time step | $\mathbf{x}$ |
| Number of EEG channels | $M$ |
| patch length | $P$ |
| Number of EEG signal length to be predicted | $T$ |

## E Notions

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009