Chinese Imagined Speech EEG Classification Method Based on Topological and Frequency-Band Priors

Ke Su[1][✉] and Haoran Guo[2] and Haoju Wang[3] and Liang Tian[4]
1 School of Art and Design, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China
2 Department of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China
3 Department of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China
4 Department of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China
coco_su0716@163.com
guo071128@163.com

Abstract. EEG-based imagined speech classification is an important topic in brain–computer interface research. However, Chinese imagined speech EEG signals are typically characterized by low signal-to-noise ratio, strong non-stationarity, and subtle inter-class differences, which make stable modeling challenging. Existing convolutional methods are limited in capturing long-range dependencies, while Transformer-based models often fail to fully exploit spatial topology and frequency-band information. To address these issues, this paper proposes a topology- and frequency-band-prior-guided convolution–Transformer hybrid model, named TFP-CTNet. The model incorporates a channel topology-aware embedding at the input stage to preserve electrode spatial relationships. A multi-scale convolutional module is then used to enhance local discriminative representations, and a frequency-band-aware gated residual MLP is introduced at the classification stage to selectively refine high-level features. Experiments on the Chisco dataset show that the proposed method consistently outperforms several competitive baselines in terms of accuracy and F1-score, while also achieving improved cross-subject consistency. These results demonstrate that incorporating structural and frequency-domain priors is beneficial for robust Chinese imagined speech EEG classification.

Keywords: brain–computer interface, imagined speech, EEG signals, topological and frequency-band priors

1 Introduction

Brain–computer interface (BCI) technology based on electroencephalography (EEG) is gradually moving toward more natural human–computer interaction. Among these, imagined speech EEG has attracted continuous attention due to its potential for silent communication without actual vocalization. Related studies have shown that imagined

speech EEG decoding still faces challenges such as weak signals, blurred class boundaries, and significant inter-subject variability, making the task inherently difficult [19].

Existing research mainly follows two directions. One line attempts to use convolutional networks to extract local temporal patterns; for example, a convolutional–recurrent network has been proposed for 9-class silent EEG speech recognition [20]. The other line focuses on stronger global modeling architectures. Studies have indicated that Transformer-based models have gradually become an important direction in EEG-BCI, yet limitations remain in terms of data requirements, interpretability, and utilization of structural priors [21].

Regarding the current state of imagined speech EEG research, prior studies have summarized the main methods and challenges from both imagined speech decoding and covert speech EEG decoding perspectives [18, 19]. Further research suggests that there is still considerable room for improvement in preprocessing, feature extraction, and classification modeling for EEG-based speech imagination decoding [23]. These findings collectively indicate that integrating local features, multi-scale information, and global dependency modeling is a crucial direction for improving task performance [22].

To address these challenges, this paper proposes a convolutional–Transformer hybrid model (TFP-CTNet) that incorporates topological and frequency-band priors. The model introduces channel topology embedding at the input stage to preserve spatial structural information, enhances local discriminative representations through multi-scale convolution in the feature extraction stage, and employs a frequency-aware gating mechanism at the classification stage for selective reconstruction of high-level features, thereby improving overall performance. Based on this approach, the main contributions of this work can be summarized as follows:

1. A convolution–Transformer framework is constructed for Chinese imagined speech EEG classification, enabling collaborative local feature extraction and global dependency modeling.
2. Channel topology information is incorporated at the input stage to better exploit the spatial structure of EEG signals.
3. A multi-scale convolutional module is designed to improve discriminative feature representation at different temporal scales.
4. A frequency-band-aware gated residual mapping strategy is proposed to enhance classification robustness.

2 Related Work

2.1 Local Feature Extraction and Deep Learning Methods

Deep learning has shown clear advantages in complex semantic EEG classification. Existing studies have demonstrated that deep models are effective in multi-character sequence tasks, and data-driven approaches can improve the representation of imagined speech features [1,2]. However, review studies have shown that improving feature robustness under low-SNR conditions remains a core challenge, and deep learning methods still suffer from data dependence and limited cross-subject generalization [3,8,13].

To strengthen local feature modeling, researchers have explored multi-scale convolutional networks, frequency-domain analysis, and convolutional-recurrent hybrid models, which are able to capture local patterns in both the temporal and spectral domains [15,18]. Incorporating band-power features and channel-space structure can improve the model's adaptability to low-SNR EEG signals and enhance its ability to distinguish subtle class differences. In addition, sparse Transformer structures and local–global hybrid networks have shown promising performance in cross-subject and complex-task scenarios [5,6,15,17], indicating their advantages in modeling complex semantic sequences and long-term dependencies. Further evidence suggests that combining local features with multi-scale information can help reduce inter-subject variability and improve classification stability and generalization [14,18]. Overall, these methods provide an important basis for Chinese imagined speech EEG decoding.

2.2 Global Modeling and Structural Optimization

As research has advanced, Transformer architectures with strong global modeling capability have been widely introduced into EEG decoding tasks and have achieved substantial improvements in mental imagery classification [4,16]. Based on a sliding-window mechanism, Swin Transformer performs hierarchical modeling over time and channel dimensions, thereby improving cross-subject generalization [11]. EEG Conformer integrates convolution and Transformer modules in a compact design, achieving efficient joint modeling of local features and global dependencies [9]. EEGNet, with its lightweight convolutional design, provides efficient spatiotemporal feature extraction and has become an important baseline model [3].

To further improve stability and discriminative power, researchers have introduced physical constraints, topological priors, channel-space structure, and frequency-band information into model design [7,14,18]. These priors not only improve training efficiency but also enhance robustness in cross-task, cross-subject, and real-time decoding scenarios, alleviating issues such as low SNR, blurred class boundaries, and pronounced individual differences. Recent reviews indicate that combining local features, multi-scale information, and global dependency modeling is a key direction for improving imagined speech EEG decoding [14,16,17,18].

Moreover, global modeling methods show strong potential in practical BCI applications. By capturing long-range dependencies and global semantic relationships, they can support silent communication, assistive communication, and intelligent interaction. When combined with multi-scale convolution and frequency-band priors, these methods are expected to further improve generalization and stability in multi-subject and multi-task environments, providing a reliable technical pathway for end-to-end EEG-BCI systems targeting Chinese imagined speech.

3 Method

3.1 Overview

This section presents a topology- and frequency-aware convolutional Transformer hybrid model, termed TFP-CTNet. The model adopts a Transformer encoder as the backbone for global dependency modeling, while introducing a channel topology-aware embedding mechanism at the input stage to preserve the spatial structure of EEG channels and incorporate electrode relationships as prior information. For feature extraction, a multi-scale convolutional front-end is employed to enhance local temporal pattern representation, which is particularly important for capturing weak and noise-sensitive discriminative cues in imagined speech EEG signals. At the classification stage, a frequency-aware gated residual MLP module is designed to selectively refine high-level features, thereby improving both discriminability and robustness. Through these components, the model forms a unified framework that integrates topology-aware representation, local feature enhancement, global dependency modeling, and discriminative feature refinement. The overall architecture of TFP-CTNet is illustrated in Fig. 1, and the details of each component are described in the following subsections.

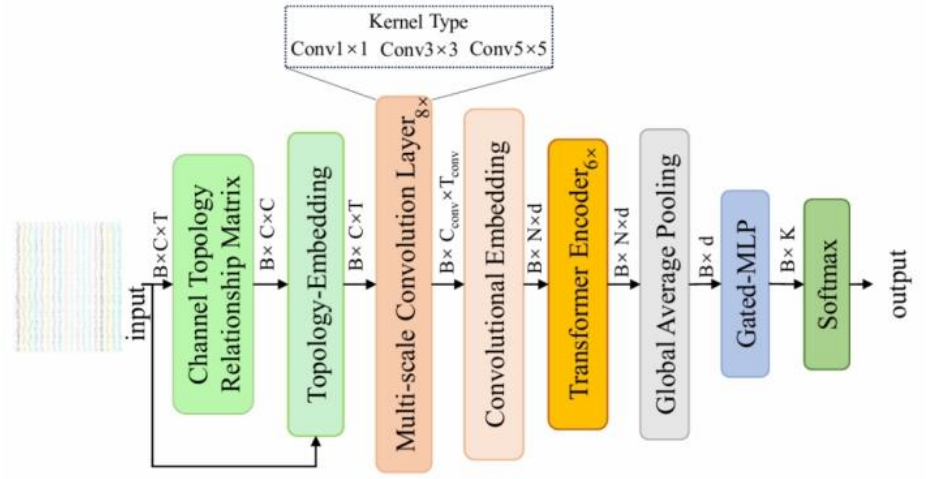

Fig. 1. The overall architecture of the proposed TFP-CTNet.

3.2 Input Representation and Channel Topology-Aware Embedding Module

In standard Transformer models, inputs are typically organized as homogeneous token sequences, where no explicit spatial structure constraints are assumed between tokens. However, for EEG signals, different channels correspond to electrodes located at different positions on the scalp, and there exist clear physical adjacency relationships and functional associations between channels. If the channel dimension is treated simply as an unordered feature dimension at the input stage, the model may fail to fully exploit inter-regional cooperative activity patterns during subsequent feature learning, which

can negatively affect classification performance. To address this, a channel topology-aware embedding mechanism is introduced at the input representation stage to encode the spatial structural priors of the electrodes into the input features. Specifically, a single EEG trial segment is represented as $X \in \mathbb{R}^{C\times T}$, where C is the number of channels and T is the number of temporal samples. Based on the three-dimensional electrode coordinates of the international 10–20 system, the topological correlation weights between channels are quantitatively computed by combining Euclidean distance and a Gaussian kernel function, thereby constructing a standardized channel adjacency matrix. Let the three-dimensional spatial coordinates of the *i*-th and *j*-th electrodes be p_i and p_j, respectively. The Euclidean distance between them is defined as:

$$D_{ij} = \left\| P_i - P_j \right\|_2 \tag{1}$$

The channel topological weight matrix $A \in \mathbb{R}^{C\times C}$ is thus defined, and its computation is given by:

$$A_{ij} = \begin{cases} exp\left(-\frac{D_{ij}^2}{2\sigma^2}\right), & D_{ij} < \tau \\ 0 \quad , D_{ij} \geq \tau \end{cases} \tag{2}$$

where σ denotes the distance scaling factor and τ represents the spatial adjacency threshold. This formulation is constrained by the physical spatial distance between electrodes: the Gaussian decay attenuates long-range electrode correlations while emphasizing local coupling among neighboring electrodes. As a result, the constructed topology matrix exhibits clear physiological spatial interpretability and full computational tractability.

Considering that different channels correspond to distinct functional properties of brain regions, the topology matrix is employed to encode channel features in a structured manner, thereby obtaining input representations enriched with spatial prior information. To further enhance the model's sensitivity to channel structural characteristics, the topological prior is incorporated into the original features in the form of an embedding term. The channel topology embedding is defined as:

$$E_{topo} \in \mathbb{R}^{\mathbb{C}\times \mathbb{d}} \tag{3}$$

where d denotes the embedding dimension. Through a linear mapping or a learnable embedding matrix, the channel indices and their topological information are projected into a unified feature space and then fused with the projected original input features, yielding the topology-aware input representation:

$$X_{topo} = \phi(X) + E_{topo} \tag{4}$$

where $\phi(\cdot)$ denotes a one-dimensional depthwise convolutional projection rather than a simple linear transformation. Specifically, this operation adopts grouped 1D convolutions to independently map the temporal features of each channel, thereby preserving the spatial structural information across channels while achieving feature dimensional alignment and shallow temporal encoding. This design enables the model to

jointly capture temporal dynamics and channel-wise spatial structure in the subsequent local convolutional extraction and Transformer encoding stages.

3.3 Multi-Scale Convolutional Enhancement Module

Transformer encoders have clear advantages in modeling long-range dependencies; however, their self-attention mechanism is relatively less sensitive to local weak discriminative patterns in low signal-to-noise ratio scenarios. For Chinese SI-EEG tasks, task-related information often appears as fine-grained waveform variations within local temporal windows, which are easily affected by background EEG activity and artifacts. If the input is directly fed into the Transformer encoder, the model may be more susceptible to global noise components, thereby weakening local discriminative information. Therefore, as shown in Fig. 2, a multi-scale convolutional enhancement module is designed before the Transformer encoder to strengthen local pattern extraction in a prior manner. The core idea of the multi-scale convolutional enhancement frontend is to employ convolution kernels with different receptive fields to extract input features in a parallel manner, thereby capturing discriminative patterns across different temporal scales. Let the topology-aware input representation be denoted as X_{topo}; the feature extracted by the m-th scale convolution branch is given by:

$$F^{(m)} = \sigma\left(\text{Conv}^{(m)}\left(X_{\text{topo}}\right)\right), \quad m = 1,2,\dots,M \tag{5}$$

where $\text{Conv}^{(m)}(\cdot)$ denotes the convolution mapping of the m-th scale ($m \leq 3$), $\sigma(\cdot)$ denotes the nonlinear activation function, and M is the number of multi-scale branches. Convolutions at different scales correspond to receptive fields over different temporal ranges: smaller kernels capture short-term transient variations and fine-grained local waveform details, while larger kernels cover longer temporal spans and capture relatively slower dynamic trends. To integrate information across multiple scales, the features from all branches are fused to obtain a locally enhanced representation:

$$F_{ms} = \psi\left(F^{(1)}, F^{(2)}, \dots, F^{(M)}\right) \tag{6}$$

where $\psi(\cdot)$ denotes an element-wise summation-based feature fusion operation. The fused features not only preserve discriminative patterns across multiple temporal scales but also help suppress irrelevant noise components to a certain extent while aggregating locally stable representations. This multi-scale structure is capable of capturing both short-term transient waveform variations and long-term dynamic trends, resembling a multi-resolution analysis mechanism across temporal scales. It thereby enhances the stability of discriminative features in low signal-to-noise ratio EEG signals.

In addition, to ensure that the output of the convolutional frontend is compatible with the input format required by the subsequent Transformer encoder, a convolutional embedding layer is introduced after the multi-scale convolution module. This layer trans-

forms the locally enhanced features into token representations with a unified dimensionality, effectively mapping the EEG representations from the local convolutional feature space into a sequential representation space suitable for self-attention modeling.

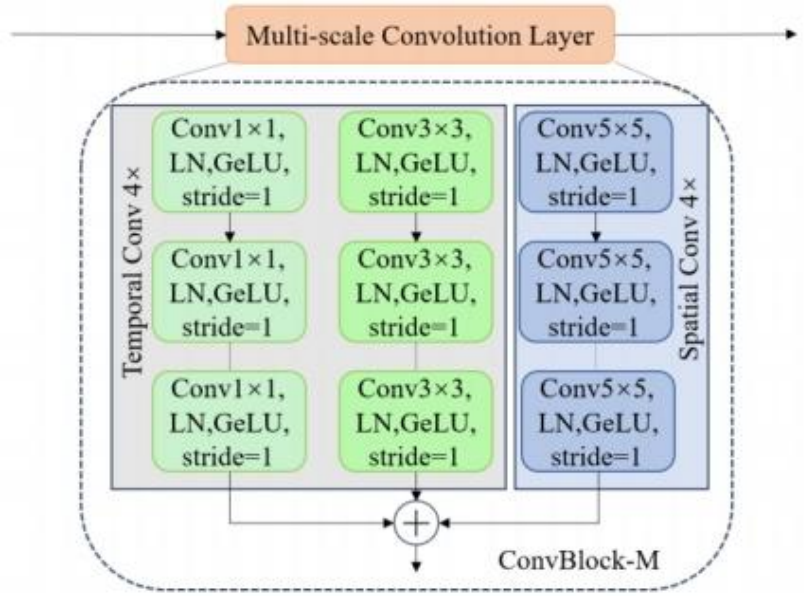

Fig. 2. Multi-Scale Convolutional Enhancement Component

3.4 Transformer Encoder for Global Modeling Module

After completing topology-aware input encoding and multi-scale local feature enhancement, the model proceeds to the global dependency modeling stage. A Transformer encoder is adopted as the backbone module, where a multi-head self-attention mechanism is employed to model the relationships among different temporal segments and cross-channel features, thereby addressing the limitations of convolutional structures in capturing long-range dependencies. Let the features after multi-scale convolutional enhancement, following reorganization and linear projection, be represented as a token sequence:

$$Z_0 = [z_1, z_2, \dots, z_N] \in \mathbb{R}^{\mathbb{N}\times\mathbb{d}} \tag{7}$$

where (N) denotes the number of tokens and (d) represents the embedding dimension. To preserve temporal positional information, positional encoding is added to the token sequence before it is fed into the encoder:

$$\tilde{Z}_0 = Z_0 + E_{pos} \tag{8}$$

where $E_{\text{pos}} \in \mathbb{R}^{N\times d}$ is the positional encoding matrix, which adopts a fixed positional encoding scheme.

As illustrated in Fig.3, the Transformer encoder consists of six stacked encoder blocks. Each encoder block mainly includes a multi-head self-attention (MHSA) sub-layer with 8 attention heads and a feed-forward network (FFN) sub-layer, along with residual connections and layer normalization. For the l-th encoder block, given the input Z_{l-1}, the output of the self-attention mechanism can be expressed as:

$$\widehat{Z}_l = \mathrm{MHSA}\big(\mathrm{LN}(Z_{l-1})\big) + Z_{l-1} \tag{9}$$

where $LN(.)$ denotes the layer normalization operation. The output is then passed through a feed-forward transformation:

$$Z_l = FFN\left(LN\left(\widehat{Z}_l\right)\right) + \widehat{Z}_l \tag{10}$$

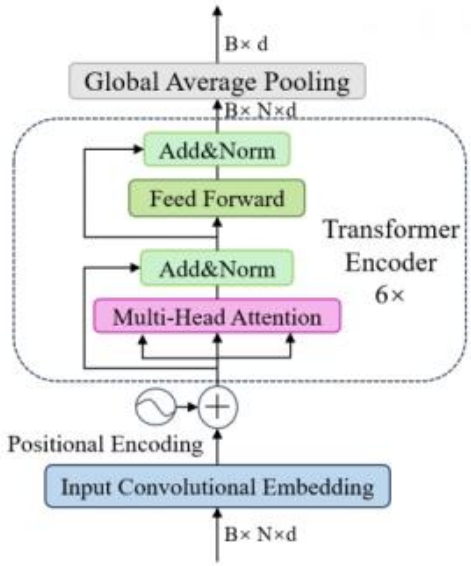

Fig. 3. Transformer Encoder Module

The multi-head self-attention mechanism learns relationships among tokens in parallel across different subspaces, enabling the model to simultaneously focus on various temporal segments, channel combinations, and their interaction patterns. For SI-EEG, this mechanism facilitates the capture of discriminative information distributed across different time positions and establishes high-level dependencies across time and channels, thereby forming intermediate feature representations with stronger global semantic consistency. After stacking (L) encoder layers, the high-level encoded feature $Z_L \in \mathbb{R}^{N\times d}$ is obtained. Subsequently, a feature aggregation strategy is applied to pool the token features, which can be expressed as:

$$h = \mathrm{Pool}(Z_L) \tag{11}$$

where Pool(·)denotes the global average pooling operation, and $h \in \mathbb{R}^d$represents the global semantic representation vector. The primary computational cost at this stage arises from the multi-head self-attention mechanism, with a time complexity of approximately O($N^2 d$). Since the preceding convolutional front-end has already performed local compression and feature reorganization, the input sequence length to the subsequent encoder can be reduced to some extent, thereby alleviating the overall computational burden.

3.5 Frequency-Aware Gated Residual MLP Classification Module

Chinese SI-EEG signals are characterized by low signal-to-noise ratio and uneven distribution of discriminative information across frequency bands. If the classification stage lacks selective reconstruction of high-level features and guidance from frequency-band information, discriminative components may be mixed with redundant noise responses during the mapping process, thereby limiting the final classification performance. To address this issue, a frequency-aware gated residual MLP module is designed at the classification stage to enhance the discriminability and robustness of high-level feature representations. As illustrated in Fig.4, to improve the model's ability to utilize discriminative information from different frequency bands, a frequency-aware guidance vector (g_f) is first constructed, which is generated by a frequency-band mapping function based on intermediate features:

$$g_f = \rho(h) \in \mathbb{R}^{\mathbb{d}} \tag{12}$$

where $\rho(.)$ denotes the frequency-aware mapping function, which is used to reconstruct frequency-band-related weighting information. Subsequently, the activation function σ_f (.) is applied to normalize it into the range ((0,1)), yielding the frequency-guided weights:

$$\alpha_f = \sigma_f(g_f) \tag{13}$$

Based on this, a gated residual MLP structure is designed to perform discriminative mapping on the global features. Let the basic MLP transformation be defined as:

$$u = \varphi_2\left(\delta(\varphi_1(h))\right) \tag{14}$$

where φ1 (.) and φ2 (.) denote two linear transformations, respectively, and δ(.) represents a nonlinear activation function. To suppress noise-driven redundant activations, a gating mechanism is introduced to adaptively reweight the output features, and it is further combined with frequency-guided weights to obtain the gated output:

$$\tilde{u} = (\alpha_g \odot u) \odot \alpha_f \tag{15}$$

α_f represents the frequency-band weighting, which enhances EEG rhythms relevant to imagined speech while suppressing noise-dominated frequency components. α_g denotes the feature gating weights, which adaptively filter out redundant and interference-prone features. Through element-wise combination, the two mechanisms jointly optimize feature representations from both the frequency domain and the feature dimension, thereby improving the discriminability of EEG features under low signal-to-noise ratio conditions.

Considering that strong gating may lead to the attenuation of useful information, a residual pathway is further introduced to preserve the original global representation, thereby constructing a residual gated discriminative mapping:

$$h_{,} = \tilde{u} + h \tag{16}$$

This residual design helps enhance feature selectivity while maintaining stable information flow, preventing the model from over-relying on the gating branch and thus avoiding the loss of discriminative information. Finally, the feature $h_{,}$ after residual gated mapping is fed into the classification head to produce the class logits:

$$o = W_c h_{,} + b_c \tag{17}$$

where W_c and b_c denote the weights and bias of the classification layer, respectively. The predicted probability distribution over classes is obtained after applying the Softmax normalization. In summary, TFP-CTNet achieves efficient classification of Chinese imagined speech EEG signals through a synergistic four-stage framework, comprising topology prior constraints, local convolutional enhancement, global dependency modeling, and frequency-aware gated discriminative mapping.

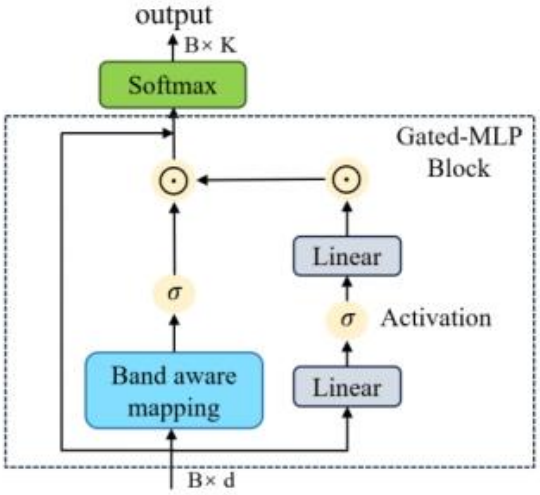

Fig. 4. Frequency-Aware Gated Residual MLP Discriminative Module

4 Experimental Setup and Results Analysis

4.1 Dataset and Data Preprocessing

In Chinese imagined speech EEG classification tasks, model performance is influenced not only by network architecture but also by experimental paradigms, stimulus types, and signal acquisition conditions. Compared with phoneme-level tasks, word-level imagination involves more complex semantic organization and articulation structures, exhibiting more pronounced temporal coordination patterns. In this study, the Chisco Dataset[14] is used as the experimental dataset, which contains 39 classes with a total of 6,681 sentences of daily Chinese expressions. To ensure a balanced and effective data distribution, the top 7 categories with the largest sample proportions are selected as the core subset (as shown in Table 1). The data were collected from three healthy right-handed subjects (S1–S3, 2 males and 1 female, aged 22–30) using the international 10–20 electrode placement system, with a sampling rate of 1 kHz to ensure high temporal resolution. The experimental paradigm is divided into three stages, including cue presentation (5 s), imagined speech (3.3 s), and rest interval (1.8 s). To maintain the purity of supervised samples, each trial is treated as the basic analysis unit, and only EEG signals from the imagined speech phase are extracted, thereby avoiding potential information contamination caused by arbitrary sliding window segmentation of continuous signals.

In the preprocessing stage, this study follows the principle of "minimal intervention and preservation of valid information," and performs standardized processing based on the MNE toolbox. First, the original 1 kHz signals are resampled to 500 Hz, reducing computational complexity while preserving frequency components below 250 Hz. Then, the PREP pipeline is applied for robust re-referencing, abnormal channel detection, and interpolation-based repair to improve spatial consistency. For frequency-domain processing, only a 50 Hz notch filter is applied to remove power-line interference, and a 1 Hz zero-phase high-pass filter is used to suppress low-frequency drifts, while no low-pass filtering is applied to avoid restricting high-frequency components. Based on event markers, the 5–8.3 s imagined speech segments are extracted as model inputs. In addition, the Autoreject algorithm is used to automatically remove abnormal data segments, and finally extended Infomax ICA is applied to eliminate artifacts, resulting in trial-level sample tensors with stable structure and controlled noise, thereby providing reliable inputs for subsequent model training.

Table 1. Core subset of 7 classes from the Chisco dataset

Classes	Number of Samples
Job Hunting and Career Development	321
Food and Dining	307
Personal Behavior and Daily Activities	274
Emotions and Interpersonal Relationships	274
Inquiries and Personal Matters	267
Health and Safety	267
Work	261

4.2 Experimental Setup

To ensure reproducibility and fair comparison of experimental results, all models are trained and evaluated under a unified hardware and software environment. The experiments are conducted on an NVIDIA GeForce RTX 4090 GPU with 24 GB of memory, an Intel Xeon(R) Platinum 8352V CPU (16 cores), and 120 GB of RAM. The software platform is implemented based on Python 3.8 and PyTorch 2.0.0.

In terms of training strategy, the model parameters are optimized using the AdamW optimizer to balance adaptive learning rates and weight decay. The initial learning rate is set to $1e$-3, the batch size is 64, and the maximum number of training epochs is 100. To mitigate overfitting, Dropout is applied in both the multi-scale convolution module and the classification MLP module, with rates of 0.2 in the convolutional frontend, 0.1 within the Transformer encoder, and 0.2 in the MLP layers. The loss function is defined as cross-entropy loss, combined with an early stopping strategy to monitor validation performance.

In the evaluation protocol, considering the significant inter-subject variability and class distribution fluctuations in imagined speech EEG data, a subject-specific stratified 5-fold cross-validation scheme is adopted. In each fold, the class distribution is kept consistent, with approximately 80% of the data used for training and 20% for testing. The final results are reported as the mean accuracy and standard deviation across all folds. To ensure that performance differences are attributed to model architecture improvements, all comparison methods are trained and evaluated under the same training strategy and data partitioning scheme.

4.3 Evaluation Metrics

This study adopts classification accuracy (ACC) and F1-score as the primary evaluation metrics. Accuracy reflects the overall proportion of correctly classified samples and serves as the most intuitive measure of overall performance. However, when class distributions are imbalanced, accuracy alone may obscure the model's ability to recognize minority classes; therefore, the F1-score is introduced to provide a balanced assessment

by jointly considering precision and recall. Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives in the confusion matrix, respectively. The accuracy is defined as:

$$\mathrm{ACC} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \tag{18}$$

The F1-score is defined as the harmonic mean of precision $PR = \frac{TP}{TP+FP}$ and recall $RE = \frac{TP}{TP+FN}$:

$$F1 = \frac{2 \cdot PR \cdot RE}{PR+RE} \tag{19}$$

4.4 Comparison Results and Analysis

To comprehensively evaluate the effectiveness of TFP-CTNet in Chinese imagined speech EEG classification, it is compared with several representative methods in the SI-BCI domain, including CSP+SVM [22], MBBN-PCNN [23], EEGNet [24], EEG-Transformer [25], and TH-LSTM-Transformer [26], covering different paradigms such as traditional machine learning, classical convolutional neural networks, attention-enhanced architectures, and temporal attention-based hybrid models. The experimental results are shown in Table 2.

Table 2. Comparison Results on the Chisco Dataset

Method	Subjects							
	S1		S2		S3		Overall	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1
CSP+SVM	52.16	48.32	50.71	49.16	54.15	55.15	52.34	50.88
MBBN-PCNN	58.43	56.55	60.53	62.55	61.08	60.11	60.01	59.74
EEGNet	64.29	65.84	62.82	60.42	63.25	63.92	63.45	63.39
EEG-Transformer	68.52	67.06	69.17	66.84	69.46	66.43	69.05	66.78
TH-LSTM-Transformer	69.31	67.11	68.44	65.36	70.39	68.75	69.38	67.07
Ours	**71.74**	**70.27**	**70.35**	**68.78**	**72.58**	**71.26**	**71.56**	**70.10**

Overall, TFP-CTNet achieves the best performance across all three subjects, with an average accuracy of 71.56% and an average F1-score of 70.10%. Compared with the current strongest baseline, TH-LSTM-Transformer (Overall ACC 69.38%, F1 67.07%), TFP-CTNet improves by 2.18 and 3.03 percentage points, respectively. Compared with EEG-Transformer (69.05% ACC, 66.78% F1), it achieves improvements of 2.51 and 3.32 percentage points. Against the classical EEGNet (63.45% ACC, 63.39% F1), the

gains reach 8.11 and 6.71 percentage points, respectively. Compared with the traditional CSP+SVM (52.34% ACC, 50.88% F1), the improvements are as high as 19.22 percentage points in both metrics.

From the perspective of cross-subject consistency, TFP-CTNet achieves ACC values of 71.74%, 70.35%, and 72.58% on subjects S1, S2, and S3, respectively, with relatively small performance fluctuations. This indicates that the proposed method maintains stable discriminative capability despite inter-individual variations in EEG distributions. In contrast, traditional methods and some shallow deep-learning baselines exhibit more pronounced performance differences across the three subjects, suggesting that relying solely on handcrafted features or single-stage local convolutional representations is insufficient to capture the weak and dispersed discriminative cues in word-level Chinese imagined speech tasks.

4.5 Ablation Study Results and Analysis

To evaluate the contribution of each key component, ablation experiments are conducted in this chapter. The Transformer is used as the baseline model, and under the condition that all other training settings remain unchanged, the channel topology-aware embedding (Topo-embedding), the multi-scale convolutional enhancement module (ConvBlock-M), and the frequency-aware gated residual MLP discriminative module (Gated-MLPBlock) are introduced separately. Furthermore, the synergistic effects of different module combinations are also investigated. The experimental results are shown in Table 3.

From the perspective of individual module contributions, starting from the baseline (ACC 64.36%, F1 63.33%), adding only the Topo-embedding improves performance to 66.53% and 65.14%, yielding gains of 2.17% in accuracy and 1.81% in F1-score, respectively. This indicates that explicitly preserving channel spatial topology priors helps stabilize the spatial consistency modeling of the Transformer. When only the ConvBlock-M is introduced, performance increases to 67.24% (ACC) and 66.71% (F1), corresponding to improvements of 2.88% and 3.38%, suggesting that multi-scale convolution provides significant gains in capturing local temporal discriminative patterns in word-level imagined speech EEG, and offers cleaner local representations for subsequent global modeling. With only the Gated-MLPBlock, performance reaches 66.75% (ACC) and 65.47% (F1), improving by 2.39% and 2.14%, respectively, demonstrating that introducing frequency-aware gating and residual reweighting at the classification stage helps suppress irrelevant dimensions while enhancing key discriminative features.

Table 3. Ablation Study of Model Components on the Chisco Dataset

Baseline	Topo-embedding	ConvBlock-M	Gated-MLPBlock	ACC	F1
√	×	×	×	64.36	63.33
√	√	×	×	66.53	65.14
√	×	√	×	67.24	66.71
√	×	×	√	66.75	65.47
√	√	√	×	68.17	66.92
√	√	×	√	68.42	67.39
√	×	√	√	70.63	69.58
√	**√**	**√**	**√**	**71.56**	**70.10**

From the perspective of dual-module combinations, the combination of Topo-embedding and ConvBlock-M achieves 68.17% (ACC) and 66.92% (F1), while the combination of Topo-embedding and Gated-MLPBlock reaches 68.42% (ACC) and 67.39% (F1). Notably, the combination of ConvBlock-M and Gated-MLPBlock, even without introducing the topology embedding, achieves 70.63% (ACC) and 69.58% (F1), representing improvements of 6.27% and 6.25% over the baseline, respectively. This highlights that the "local enhancement + discriminative reconstruction" paradigm exhibits stronger complementarity in this task: the convolutional module improves the quality of local representations, while the gated discriminative module further performs selective enhancement during the classification stage, resulting in gains that significantly exceed linear expectations from single-module stacking. When all three modules are integrated, TFP-CTNet achieves 71.56% (ACC) and 70.10% (F1), corresponding to improvements of 7.20% and 6.77% over the baseline, respectively, providing module-level evidence for the optimal results reported in Table 2.

5 Conclusion

This study addresses challenges in Chinese imagined speech EEG classification, including complex rhythmic structures, blurred class boundaries, and significant cross-subject variability. To this end, the proposed TFP-CTNet framework incorporates topology-aware priors and frequency-band enhancement mechanisms to guide the model in learning under constraints of spatial connectivity and spectrally sensitive regions, while combining convolutional and Transformer architectures to enable joint learning of local and global features. Experimental results demonstrate that the proposed model achieves an improvement of approximately 2.18% in accuracy over SOTA methods

and about 7% over baseline models, while also outperforming mainstream approaches in terms of both accuracy and stability. Ablation studies further confirm the effectiveness of the topology- and frequency-guided modeling strategy.

Despite these promising results, the model still has certain limitations, such as the relatively small number of subjects, insufficient modeling of EEG non-stationarity and semantic context, and the need for improved real-time performance. Future work may focus on expanding dataset scale, multi-center data collection, dynamic frequency-band adaptation, lightweight model design, and multimodal fusion to further enhance performance and applicability.

Overall, TFP-CTNet demonstrates the effectiveness of integrating topological and frequency-band priors in complex Chinese imagined speech EEG classification tasks, providing a reference pathway for end-to-end brain–computer interface modeling.

Acknowledgments. This study was funded by National Natural Science Foundation of China General Program: Research on Integration Design Method of Complex Product Based on Informed Tectonics（NO.52575294）

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

. References

1. Pan, H., Wang, Y., Li, Z., et al.: A complete scheme for multi-character classification using EEG signals from speech imagery. IEEE Transactions on Biomedical Engineering 71(8), 2454–2462 (2024)
2. Abdulghani, M.M., Walters, W.L., Abed, K.H.: Imagined speech classification using EEG and deep learning. Bioengineering 10(6), 649 (2023)
3. Alzahrani, S., Banjar, H., Mirza, R.: Systematic review of EEG-based imagined speech classification methods. Sensors 24(24), 8168 (2024)
4. Ahn, H.J., Lee, D.H., Jeong, J.H., et al.: Multiscale convolutional transformer for EEG classification of mental imagery in different modalities. IEEE Transactions on Neural Systems and Rehabilitation Engineering 31, 646–656 (2022)
5. Liu, M., Liu, Y., Shi, W., et al.: EMPT: a sparsity transformer for EEG-based motor imagery recognition. Frontiers in Neuroscience 18, 1366294 (2024)
6. Xue, S., Jin, B., Jiang, J., et al.: A hybrid local-global neural network for visual classification using raw EEG signals. Scientific Reports 14(1), 27170 (2024)
7. Altaheri, H., Muhammad, G., Alsulaiman, M.: Physics-informed attention temporal convolutional network for EEG-based motor imagery classification. IEEE Transactions on Industrial Informatics 19(2), 2249–2258 (2022)
8. Al-Saegh, A., Dawwd, S.A., Abdul-Jabbar, J.M.: Deep learning for motor imagery EEG-based classification: A review. Biomedical Signal Processing and Control 63, 102172 (2021)
9. Song, Y., Zheng, Q., Liu, B., et al.: EEG conformer: convolutional transformer for EEG decoding and visualization. IEEE Transactions on Neural Systems and Rehabilitation Engineering 31, 710–719 (2022)

10. Qamar, W.U.R., Abibullaev, B.: Multi-scale EEG feature decoding with Swin transformers for subject-independent motor imagery BCIs. Scientific Reports 16(1), 2503 (2026)
11. Lawhern, V.J., Solon, A.J., Waytowich, N.R., et al.: EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. Journal of Neural Engineering 15(5), 056013 (2018)
12. Zhang, Z., Ding, X., Bao, Y., et al.: Chisco: An EEG-based BCI dataset for decoding imagined speech. Scientific Data 11(1), 1265 (2024)
13. Hossain, A., Khan, P., Kader, M.F.: Imagined speech classification exploiting EEG power spectrum features. Medical & Biological Engineering & Computing 62(8), 2529–2544 (2024)
14. Haresh, M.V., Begum, B.S.: Towards imagined speech: identification of brain states from EEG signals for BCI-based communication systems. Behavioural Brain Research 477, 115295 (2025)
15. Nguyen, A.H.P., Oyefisayo, O., Pfeffer, M.A., et al.: EEG-TCNtransformer: a temporal convolutional transformer for motor imagery brain–computer interfaces. Signals 5(3), 605–632 (2024)
16. Panachakel, J.T., Ramakrishnan, A.G.: Decoding covert speech from EEG – a comprehensive review. Frontiers in Neuroscience 15, 642251 (2021)
17. Lopez-Bernal, D., Balderas, D., Ponce, P., et al.: A state-of-the-art review of EEG-based imagined speech decoding. Frontiers in Human Neuroscience 16, 867281 (2022)
18. Vorontsova, D., Menshikov, I., Zubov, A., et al.: Silent EEG-speech recognition using convolutional and recurrent neural network with 85% accuracy of 9 words classification. Sensors 21(20), 6744 (2021)
19. Abibullaev, B., Keutayeva, A., Zollanvari, A.: Deep learning in EEG-based BCIs: a comprehensive review of transformer models, advantages, challenges, and applications. IEEE Access 11, 127271–127301 (2023)
20. Keutayeva, A., Abibullaev, B.: Data constraints and performance optimization for transformer-based models in EEG-based brain-computer interfaces: a survey. IEEE Access 12, 62628–62647 (2024)
21. Rahman, N., Khan, D.M., Masroor, K., et al.: Advances in brain-computer interface for decoding speech imagery from EEG signals: a systematic review. Cognitive Neurodynamics 18(6), 3565–3583 (2024)
22. Wang, L., Huang, X., Ren, L., et al.: Signal analysis and classification of a novel active brain-computer interface based on four-category sequential coding. Biomedical Signal Processing and Control 78, 103857 (2022)
23. Wang, J., Wang, L.: Parallel convolutional neural network based on multi-band brain networks for EEG classification. In: Proc. 5th Int. Conf. Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), pp. 49–53. IEEE (2022)
24. Schirrmeister, R.T., Springenberg, J.T., Fiederer, L.D.J., et al.: Deep learning with convolutional neural networks for EEG decoding and visualization. Human Brain Mapping 38(11), 5391–5420 (2017)
25. Lee, Y.E., Lee, S.H.: EEG-transformer: self-attention from transformer architecture for decoding EEG of imagined speech. In: Proc. 10th Int. Winter Conf. Brain-Computer Interface (BCI), pp. 1–4. IEEE (2022)
26. Pan, H., Teng, B., Li, Z., et al.: A hybrid TH-LSTM-transformer model for text generation from EEG signals during imagined character speech. Biomedical Signal Processing and Control 113, 108871 (2026)