China Instrument and Control Society
Instrumentation
INSTRUMENTATION, Vol. 13, No. 2, June 2026 https://doi.org/10.15878/j.instr.202600334

Article

A Novel Deep Multi-Module Fusion Network for Speech Imagery EEG Decoding

Mengyao Yuan, Zhengdong Zhou*, Xiaoxi Yuan, Zeyi Yang, Zhi Cai

State Key Laboratory of Mechanics and Control for Aerospace Structures, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, China

* Corresponding author email: zzd_msc@nuaa.edu.cn

Abstract: To enhance the decoding performance of speech imagery EEG (SI-EEG) signals, a novel deep multi-module fusion network is proposed and evaluated on two public SI-EEG datasets. This method introduces a spatio-temporal-frequency convolution (STFCV) module to extract rich and fine-grained multi-domain features from the frequency-spatial domain of raw SI-EEG signals. A multi-head self-attention (MHSA) mechanism is further incorporated to emphasize critical EEG features within the SI-EEG signals. Additionally, a convolution-based sliding window data augmentation strategy is employed to enhance data diversity. A temporal convolutional network (TCN) is also integrated to effectively model long-range temporal dependencies, thereby boosting the decoding capability for SI-EEG. These modules interact in a collaborative and complementary manner, forming an organic and unified framework that enables coordinated feature extraction and fusion. Experimental results show that the proposed method achieves high decoding accuracy for both 5-class and 6-class SI tasks, outperforming existing state-of-the-art approaches on the BCI2020 and Coretto datasets. The model also achieves substantial gains in cross-subject validation. The proposed framework demonstrates that it may support practical use in speech imagery-based brain – computer interfaces, with an average inference time of approximately 0.85 s per trial, supporting its feasibility for online deployment in practical BCI scenarios.

Keywords: brain-computer interface(BCI); electroencephalogram(EEG); imagined speech; deep learning; temporal convolutional network(TCN)

Citation: Mengyao Yuan, Zhengdong Zhou, Xiaoxi Yuan, Zeyi Yang, Zhi Cai."A Novel Deep Multi-Module Fusion Network for Speech Imagery EEG Decoding." Instrumentation 13, no.2 (June 2026). https://doi.org/10.15878/j.instr.202600334

1 Introduction

Brain–computer interface (BCI) technology enables direct communication between the human brain and external devices[1]. This technology allows users to control computers, prosthetic limbs, or other electronic systems using neural activity, bypassing traditional neuromuscular pathways[2]. Among the various methods for acquiring brain signals, electroencephalography (EEG) is one of the most widely used noninvasive techniques. EEG works by placing electrodes on the scalp to capture the brain's electrical activity over time, providing an accurate reflection of the nervous system's dynamic processes[3]. Due to its high temporal resolution, low cost, portability, and ease of use, EEG technology has become a fundamental tool in cognitive neuroscience research and BCI development[4].

Most EEG-based BCI systems are currently built around classical paradigms, such as the P300[5], steady-state visual evoked potential (SSVEP) [6], and motor imagery (MI) [7]. These systems enable effective communication between users and external devices. However, these traditional paradigms face several limitations, including slow response times, heavy reliance

on user-specific training, and high operational burdens. These drawbacks limit their scalability and long-term adaptability in complex real-world applications. To improve the naturalness of interaction and enhance user experience, researchers have recently turned to a novel BCI paradigm based on speech imagery (SI). Speech imagery refers to the internal generation or rehearsal of linguistic content within the brain, without overt articulation or vocalization. The resulting EEG activity, known as speech imagery EEG (SI-EEG), provides a promising source of neural data for decoding intended speech[8-10]. Compared to conventional paradigms, SI-based BCIs are more aligned with natural human communication processes, offering greater intuitiveness and flexibility. Notably, SI-based systems show great potential for assisting individuals with speech impairments or locked-in syndrome, helping them express their intentions and restore communication capabilities[11]. In SI-EEG decoding tasks, traditional machine learning methods, such as support vector machines or random forests, typically rely on handcrafted features. However, due to the nonlinear, nonstationary, and low signal-to-noise ratio of EEG data, these methods often struggle with precision and robustness in feature extraction, limiting improvements in classification performance[12].

In recent years, deep learning techniques have become the dominant approach in speech imagery EEG decoding research. These techniques, particularly through end-to-end training frameworks, allow deep neural networks to automatically extract discriminative features, integrating feature learning with classification. This integration significantly enhances the model's representation and generalization capabilities[13]. Berg et al. proposed a two-dimensional convolutional neural network based on the EEGNet architecture, achieving an average accuracy of 29.7% on the four-class Think Out Loud dataset. This result demonstrated the feasibility of applying deep neural networks to raw EEG-based speech imagery decoding[14]. Li et al. developed a hybrid-scale spatio-temporal dilated convolutional neural network that achieved 54.31% accuracy on an eight-class word recognition task. Their work showcased the effectiveness of multi-scale spatio-temporal feature extraction in complex decoding scenarios[15]. Ahn et al. introduced a multi-scale convolutional Transformer model that integrates spatial, spectral, and temporal features. By incorporating a multi-head self-attention mechanism, they achieved 72% accuracy on a binary classification task using the ASU dataset, highlighting the model's strong capacity for salient EEG feature representation[16]. Pawar et al. proposed a feature fusion method based on discrete wavelet transform (DWT) and maximum linear cross-correlation (MaxLCor), utilizing a support vector machine (SVM) classifier for a five-class classification task on the BCI2020 dataset. Their approach achieved a classification accuracy of 40.64%[17]. Zheng et al. introduced a supervised classification method combining principal component analysis (PCA) and k-means clustering, presenting an improved CatPCA algorithm. This method significantly outperformed traditional PCA, achieving 58.51% accuracy on the same dataset[18]. Additionally, Bhalerao et al. combined multivariate swarm sparse decomposition (MSSDM) with a deep feature extraction network, reaching an average accuracy of 59.07% in the five-class classification task on the BCI2020 dataset[19].

Despite these advances, challenges persist due to the limited size of available SI-EEG datasets and significant inter-subject variability. Current approaches still face difficulties in fully capturing the complex spatio-temporal-frequency characteristics embedded in SI-EEG signals. As a result, improving multi-class decoding accuracy remains an open problem.

To address the challenges outlined above, a novel deep multi-module fusion network for SI-EEG decoding is proposed and evaluated on two public SI-EEG datasets. This method presents a synergistic integration of the spatio-temporal-frequency convolution (STFCV), multi-head self-attention (MHSA), and temporal convolutional network (TCN) modules, which are not simply connected in series, but instead form an organic whole that performs feature mining and fusion across three distinct dimensions: "frequency-spatial domain, " "feature importance, " and "long-range temporal domain. " The STFCV module extracts rich, fine-grained multi-domain features from raw SI-EEG signals, capturing both spatial and temporal dependencies within the frequency domain. The MHSA mechanism is incorporated to prioritize critical EEG features within these signals, while the TCN effectively models the long-range temporal dependencies, enhancing the capture of fine-grained temporal features. Additionally, a convolution-based sliding window (SW) data augmentation strategy is employed to further enhance the diversity of the data, improving the robustness and generalization of the model.

2 Materials and Methods

2.1 Model Framework

The proposed novel deep multi-module fusion (MMF) network comprises an STFCV module, an MHSA module, a TCN module, and a convolution-based SW module. The overall architecture of the model is illustrated in Fig. 1.

The STFCV module extracts multi-scale and multi-dimensional features from SI-EEG signals by applying frequency convolution, channel-wise depth convolution, and spatial convolution. This enables it to simultaneously capture spectral, spatial, and temporal information, providing a rich representation of the raw EEG signals. The output features from the STFCV module are then augmented using a sliding window technique, which

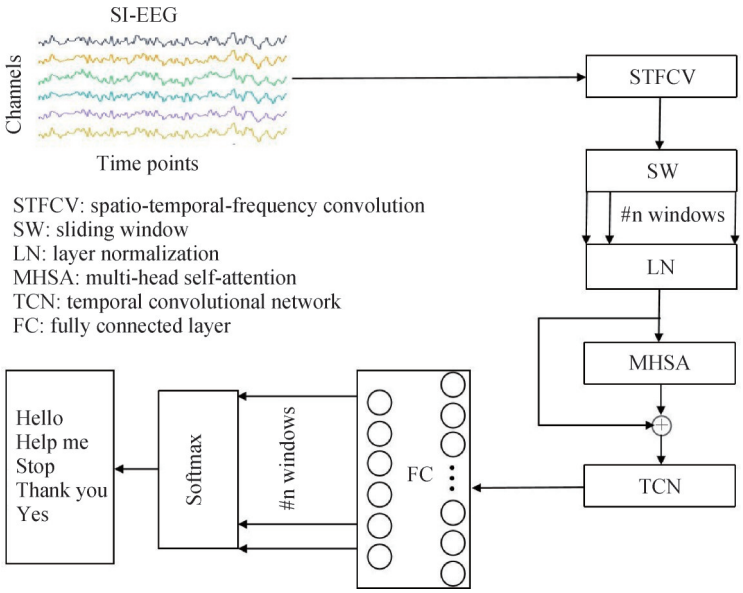


Fig.1 Overall model architecture

segments the features into smaller windows for further processing. Within each window, the features undergo layer normalization (LN) before being passed to the MHSA module.

The MHSA module is designed to prioritize the most informative features, enhancing their expressiveness and discriminative power, thereby enabling the model to focus on the most critical aspects of the signal. In addition, the TCN module extracts deep, fine-grained temporal dependencies from the SI-EEG signals, enabling the model to capture long-range temporal relationships.

The outputs from the MHSA and TCN modules are fused through a residual connection, which ensures feature stability and enhances the discriminative robustness of the model. Finally, the fused representations from all sliding windows are aggregated and passed through a fully connected (FC) layer, where a softmax classifier performs the final classification.

2.2 STFCV Module

The STFCV module consists of three types of convolution operations and two average-pooling layers, as shown in Fig. 2. The first two layers follow a three-branch parallel structure. In each branch, frequency convolution is applied along the temporal axis to extract spectral features from the SI-EEG signals. This step effectively learns multiple frequency filters, enabling the capture of key spectral information. After performing multi-scale frequency convolution and channel-wise depth separable convolution in each branch, the resulting feature maps are concatenated along the feature dimension to form a unified representation. This concatenation strategy preserves the complementary information learned at different temporal scales, enhancing the frequency-spatial representation of SI-EEG signals.

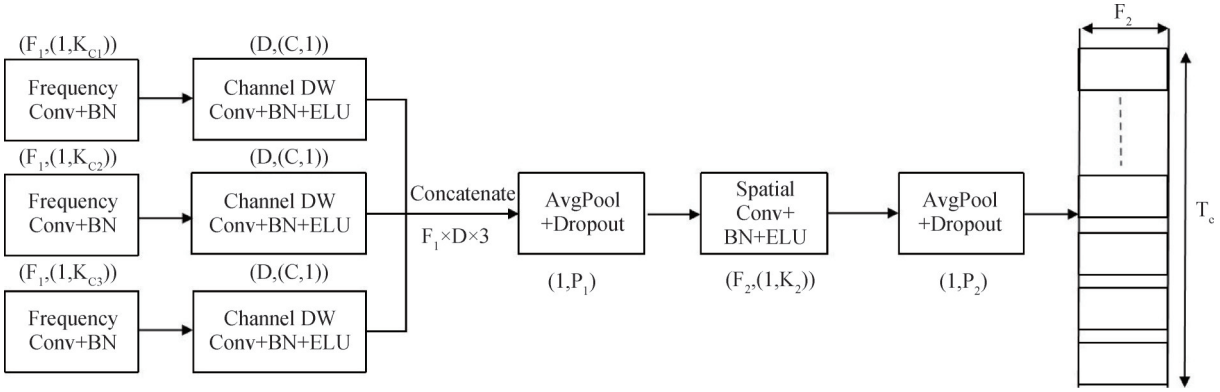


Fig.2 Spatio-temporal-frequency convolution module

The kernel sizes K_C are set to one-quarter, one-eighth, and one-sixteenth of the EEG signal's sampling frequency (e.g., for a sampling rate of 256 Hz, K_C = [64, 32, 16]). This design follows the multiscale temporal convolution strategy adopted in TSception[20] and the multiscale convolutional block of Ahn et al.[16] In these models, kernel lengths are chosen as fixed fractions of the sampling rate to align the receptive field with

physiologically meaningful oscillatory rhythms. At 256 Hz, kernel lengths of 64, 32, and 16 samples correspond to temporal windows of 250 ms, 125 ms, and 62.5 ms, respectively, making the three branches preferentially sensitive to activity above approximately 4 Hz, 8 Hz, and 16 Hz. These scales effectively span the 4-30 Hz frequency band, where speech imagery EEG activity is predominantly distributed, enabling the STFCV module to capture complementary low-, mid-, and relatively high-frequency components. Prior multiscale convolutional designs have empirically demonstrated improved performance in speech imagery and emotion-related EEG classification tasks. After frequency convolution, each branch applies depthwise separable convolution with a kernel size of $(C,1)$, where C denotes the number of EEG electrode channels. Since the kernel length matches the channel dimension of the input EEG segment, it can capture global spatial relationships. This operation extracts spatial features across different electrode channels of the SI-EEG signals.

To overcome the limitation of fixed kernel lengths, a parallel multi-branch structure with multiple convolutional kernel sizes is introduced. This enhances the model's ability to capture frequency-domain variations and better adapt to the wide range of spectral features present in the SI-EEG signals.

After the frequency and depthwise convolutions, the feature maps from all branches are fused through concatenation. To reduce feature dimensionality and mitigate overfitting, an average-pooling (AvgPool) layer with a kernel size of $(1, P_1)$, where $P_1 = 8$, is applied to compress the temporal dimension of each sample.

The third convolutional layer then applies F_2 filters with a kernel size of $(1, K_2)$, where $K_2 = 16$, to further integrate spatio-temporal-frequency information from the downsampled signal.

An additional average-pooling layer with a kernel size of $(1, P_2)$, where $P_2 = 6$, is then applied to further reduce feature dimensionality. P_2 controls the final sequence length of the output features. To accelerate convergence, batch normalization (BN) is applied after the second and third convolutional layers. The exponential linear unit (ELU) activation function is used to introduce nonlinearity, enhancing model stability.

Finally, the STFCV module outputs a feature representation sequence $z \in R^{Tc \times d}$, where T_c denotes the number of temporal feature vectors and $d = F_2$ represents the feature dimension of each vector. The value of T_c is determined by the following equation (1):

$$T_c = \frac{T}{P_1 P_2}. \tag{1}$$

where T is the original sample length of the SI-EEG signal, P_1 and P_2 correspond to the pooling parameters of the two average-pooling layers, respectively.

2.3 SW Module

To enhance data diversity and improve decoding accuracy, a convolution-based sliding window mechanism is employed for feature segmentation and data augmentation, without introducing additional learnable parameters. This operation allows for both feature partitioning and identity transformation of the SI-EEG representations[21]. Specifically, a convolutional sliding window with a length of T_w and a stride of 1 is applied to the feature sequence $z \in R^{Tc \times d}$, dividing it into n overlapping windows $z^w \in R^{Tw \times d}$, where $w = 1, ..., n$ denotes the window index. Each segmented feature sequence z^w is subsequently fed into the MHSA module, followed by the TCN module for further temporal dependency learning. The window length T_w is determined by the following equation (2):

$$T_w = T_c - n + 1, T_c > n \geq 1. \tag{2}$$

If the STFCV module applies two temporal pooling operations with sizes $P_1 = 8$ and $P_2 = 6$, the resulting temporal sequence z will consist of $T_c = 16$ vectors, as shown in Eq. (1), where $T = 795$. If the number of windows $n = 5$, the window length $T_w = 12$, as shown in Eq. (2). Each vector in z corresponds to 48 (8 × 6) time-points from the original SI-EEG signal. Consequently, performing a single sliding step on the sequence z is equivalent to sliding by 48 time-steps in the original signal. This design reflects a balance between generating sufficiently diverse augmented segments and avoiding excessive overlap or redundancy, thereby improving generalization.

2.4 MHSA Module

In deep learning models, the attention mechanism is designed to mimic the human brain's ability to selectively focus on important elements while ignoring less relevant information. The multi-head self-attention (MHSA) mechanism has demonstrated significant success in fields such as natural language processing and computer vision[22].

In this study, the MHSA module is used to capture multidimensional salient EEG features by performing multiple self-attention operations in parallel. As shown in Fig. 3, the MHSA module consists of several attention heads arranged in parallel, with each head comprising

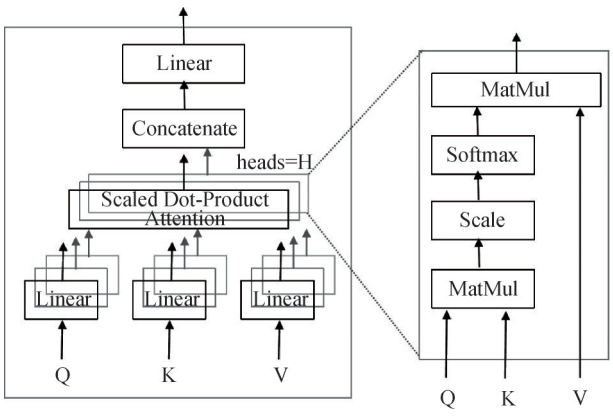


Fig.3 Multi-head self-attention mechanism

three learnable vector sets: queries (***Q***), keys (***K***), and values (***V***)[23].

For each given window feature sequence z^w, the MHSA module first generates the corresponding query (***Q***), key (***K***), and value (***V***) matrices through trainable linear projection weights. Using the scaled dot-product attention mechanism, the similarity score between ***Q*** and ***K*** is computed and normalized with a Softmax function to obtain the attention weight distribution. The weighted sum of the ***V*** vectors, based on the attention weights, results in the local contextual feature representation.

Each attention head uses a key dimension of 8, meaning the ***Q***, ***K***, and ***V*** vectors are all projected into 8-dimensional feature spaces per head. Based on experimental validation, the study used the optimal configuration of 3 attention heads, so the concatenated output of the multi-head attention layer has a dimension of 24. A final linear projection is then applied to map this 24-dimensional concatenated vector back into the feature space used by the subsequent layers.

The MHSA module concatenates the contextual representations generated by all attention heads and performs a linear projection to combine them into an integrated multi-dimensional representation. Finally, a residual connection is applied to merge the fused output with the original input feature sequence z^w, thereby enhancing feature stability and improving gradient flow during training.

2.5 TCN module

The temporal convolutional network is a convolution-based neural architecture that enhances the conventional convolutional neural network (CNN) by incorporating three key features: causal convolution, dilated convolution, and residual blocks. These features enable the TCN to model temporal dependencies effectively in sequential data and enhance its ability to capture long-range temporal patterns[24]. As shown in Fig. 4, the TCN module is constructed by stacking multiple residual blocks. Each residual block consists of two dilated causal convolutional layers. After each convolutional layer, batch normalization, an exponential linear unit (ELU) activation function, and a dropout layer are applied to reduce the risk of overfitting[25].

Causal convolution ensures that the output at each time step depends only on the current and past inputs, effectively preventing the leakage of future information. The introduction of dilated causal convolution expands the receptive field size (RFS), allowing the network to capture long-range dependencies within the temporal sequence. The RFS grows exponentially with the number of residual blocks L, and it is defined as follows:

$$\mathrm{RFS} = 1 + 2(K_\mathrm{T} - 1)(2^L - 1). \tag{3}$$

Where K_T represents the kernel size of the dilated causal convolution, and L denotes the number of residual blocks. To ensure that the network can fully utilize the entire feature sequence, the parameters K_T and L should satisfy the condition $\mathrm{RFS} \geqslant T_w$.

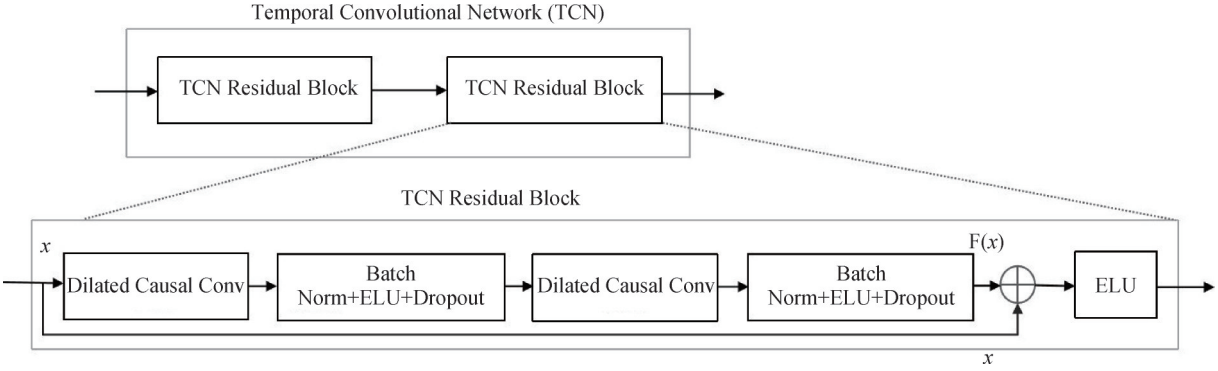


Fig.4 Temporal convolutional network

In the TCN module, the number of residual blocks L determines the depth of modeling in the temporal dimension, with typically two or more residual blocks being used. In this study, L is set to 2. The dilation coefficients are used to expand the receptive field in convolution operations, allowing the model to capture temporal dependencies at different time scales. The dilation rates are usually chosen in an exponentially increasing sequence of 1, 2, 4, $\cdots$, 2^{L-1}. In this implementation, with L = 2, the corresponding dilation rates are 1 and 2. The kernel size for temporal convolutions K_T is set to 4, which focuses on local temporal patterns while keeping the receptive field compact. Each TCN convolutional layer uses 32 filters to balance representational capability and computational complexity. The dropout rate is set to 0.3 to mitigate overfitting. Table 1 summarizes the key hyperparameters for the TCN module.

3 Results and Discussion

3.1 Experimental Environment

All experiments were conducted on a Windows 10 operating system. The model was trained on a workstation equipped with an Intel Core i7-8700 CPU

Table 1 Key hyperparameters for the TCN module

Hyperparameters	values
The number of residual blocks (L)	2
kernel size (K_T)	4
Filters	32
Dropout rate	0.3
The dilation coefficients	1, 2

(3.20 GHz, 6 cores) and an NVIDIA GeForce GTX 1060 GPU with 6 GB of memory. The software environment included the PyCharm 2024.1.4 integrated development environment (IDE), Python 3.7 interpreter, and TensorFlow 2.7 deep learning framework.

The model parameters were initialized using the Glorot uniform initializer provided by TensorFlow. During training, the Adam optimizer and categorical cross-entropy loss function were used. The initial learning rate was set to 0.001 and was dynamically adjusted using an exponential decay strategy. For the BCI2020 dataset, which contains a relatively large number of samples, the batch size was set to 64. In contrast, for the Coretto dataset with fewer samples, the batch size was reduced to 16. A stratified five-fold cross-validation was applied to the SI-EEG data. Each fold was trained for 500 epochs, and model performance was evaluated on the validation set after each epoch. An early stopping strategy with a patience value of 20 was employed to prevent overfitting.

3.2 Experimental Datasets

The BCI2020 public dataset used in this study was derived from Track 3 of the 2020 International BCI Competition, which was focused on speech imagery classification. EEG signals were recorded from 64 channels, arranged following the international 10-20 system, with a sampling rate of 256 Hz. Each task lasted for 2 seconds. The dataset includes EEG recordings from 15 healthy participants, aged between 20 and 30 years, who performed five speech imagery tasks. These tasks correspond to the phrases "Hello," "Help me," "Stop," "Thank you," and "Yes." Each participant completed 70 trials per phrase, resulting in 350 trials per participant, and a total of 5,250 samples across all subjects[26].

Expanding validation to multiple datasets can enhance the generalizability of the proposed method. Another public dataset used in this study was recorded by Coretto et al. at the National University of Entre Ríos[27]. EEG signals were collected from 6 channels, following the international 10-20 system, with the electrodes placed near the speech center at F3, F4, C3, C4, P3, and P4. The sampling rate was 1024 Hz, and each task lasted for 4 seconds. The dataset includes EEG recordings from 15 healthy participants, each performing speech imagery tasks involving six Spanish words: "arriba" (up), "abajo" (down), "derecha" (right), "izquierda" (left), "adelante" (forward), and "atrás" (backward). A total of 2, 852 samples were recorded, corresponding to the six word tasks performed by the 15 participants.

The EEG acquisition system used in this study included a built-in notch filter, which effectively suppressed power line interference, eliminating the need for additional baseline correction[26,27]. To evaluate the model's performance, the dataset for each subject in intra-subject (individual decoding) tasks was divided into five mutually exclusive and approximately equal subsets using stratified five-fold cross-validation. For cross-subject (group decoding) tasks, the data from all subjects in each dataset were integrated and then divided into five mutually exclusive and approximately equal subsets using the same stratified five-fold cross-validation method. The hyperparameters for this study were chosen based on cross-validation with grid search. In each iteration, four subsets were used for training, and the remaining one was used for testing to ensure class balance. Preprocessing and training were performed independently within each fold. The final result was determined by reporting the average classification accuracy across the five iterations.

During data partitioning, the full-channel data from each trial were assigned exclusively to a single subset to prevent cross-channel overlap. Additionally, normalization parameters were calculated only from the training data and then applied to the test data. This approach prevented data leakage, ensuring that the evaluation of the model remained valid and reliable.

3.3 Performance Metrics

The model's performance was evaluated using two metrics: accuracy (ACC) and kappa score (K_{score}).

Accuracy (ACC) is calculated as:

$$ACC = \frac{\sum_{i=1}^{s} TP_i / I_i}{s}. \tag{4}$$

where, TP_i denotes the number of true positive samples correctly predicted in class i, I_i represents the total number of samples in class i, and s is the total number of classes.

The kappa score (K_{score}) is computed as:

$$K_{\text{score}} = \frac{1}{s}\sum_{a=1}^{s} \frac{P_a - P_e}{1 - P_e}. \tag{5}$$

where, s represents the number of classes, P_a is the actual observed agreement, which is the percentage of consistent classifications, and P_e is the expected agreement percentage by chance.

3.4 Ablation Study

The model architecture plays a significant role in determining the upper limit of decoding performance. Therefore, selecting the appropriate hyperparameters and components is essential for improving practical performance. To better understand the key factors that

influence performance, it is important to conduct systematic ablation experiments. These experiments can help identify which components and configurations have the most significant impact on the model's effectiveness.

3.4.1 Effect of the Number of Sliding Windows on Decoding Performance

The feature representation sequence $z \in R^{Tc \times d}$ generated by the STFCV module contains feature vectors that correspond to the sampled information of the original SI-EEG signal in the $C \times P_1 \times P_2$ space, thus capturing multi-domain features. The way this feature representation sequence is divided using a sliding window (with a fixed frame shift of 1) directly affects the final decoding performance. Different window lengths lead to different multi-dimensional feature aggregation strategies.

To assess the impact of the number of sliding windows n on SI-EEG decoding performance, n was varied within the range [1, T_c - 1]. For the BCI2020 dataset, experiments were conducted using three feature representation sequence lengths: T_c =14, T_c =16, and T_c = 19, as shown in Fig. 5. When n=1 , no sliding window was applied.

As shown in Fig. 5, the decoding accuracy increases progressively with the number of sliding windows. However, once the number of sliding windows exceeds a certain threshold, the accuracy begins to stabilize or gradually decrease. This occurs because increasing the number of sliding windows helps with data augmentation and allows the model to capture SI-EEG information at different time points. Yet, as the number of sliding windows continues to rise, the length of each window becomes shorter, which may lead to insufficient information within each window. This can negatively impact the decoding performance.

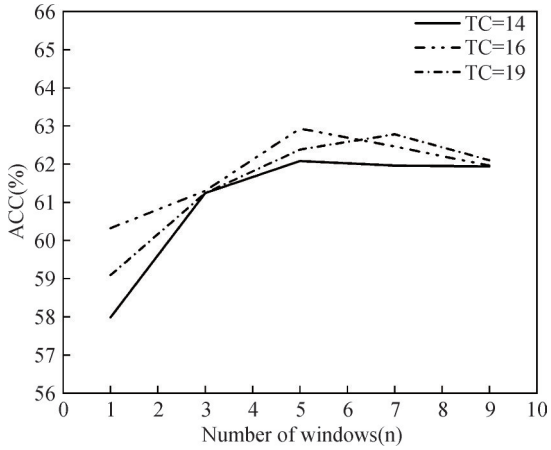


Fig.5 Effect of the number of sliding windows on decoding accuracy

On the BCI2020 dataset, when the sequence length is 16 and the number of sliding windows n=5 (corresponding to a window length of 12), the model achieves the highest decoding accuracy. This finding further highlights the importance of reasonable temporal aggregation in effectively extracting multi-domain fused EEG features from SI-EEG.

3.4.2 Effect of the Attention Mechanism on Decoding Performance

Table 2 presents a comparison of the impact of several commonly used attention mechanisms on the decoding performance of the model proposed in this paper. The attention mechanisms evaluated include the Convolutional Block Attention Module (CBAM) [28], SE Attention[29], and Multi-Head Self-Attention (MHSA). As shown in Table 2, using MHSA with 3 attention heads yields the best performance on the BCI2020 competition dataset, achieving an accuracy of 62.93% and a Kappa value of 0.537. This performance significantly outperforms the other attention mechanisms, indicating that an optimal number of attention heads enhances the model's ability to represent multi-domain features and improves its classification performance.

Table 2 Classification results of different attention mechanisms on the BCI2020 dataset

Attention	*ACC*/%	K_{score}
None	61.39	0.517
SE	61.62	0.520
CBAM	61.52	0.519
MHSA-1	62.02	0.525
MHSA-2	62.63	0.533
MHSA-3	**62.93**	**0.537**
MHSA-4	62.15	0.527
MHSA-5	62.13	0.527
MHSA-6	61.83	0.523

However, as the number of attention heads increases further, both the accuracy and Kappa value of the model decrease. This suggests that an excessive number of attention heads may lead to increased model complexity and redundant representations. It can also introduce irrelevant or noisy information, which weakens the model's discriminative ability. Therefore, when designing a multi-head attention mechanism, it is crucial to select the optimal number of heads based on the task complexity and feature space size to achieve the best performance.

3.4.3 Effect of Different Modules on Decoding Performance

To validate the effectiveness of the modules in the model, ablation experiments were conducted on the BCI2020 dataset.

In each experiment, one or more modules were removed prior to training and testing, and the impact on

speech imagery decoding performance was evaluated. The experimental results are presented in Table 3.

Table 3 Classification results of different ablation experiment settings on the BCI2020 dataset

STFCV	MHSA	SW	TCN	*ACC*/%	K_{score}
√	×	×	×	52.76	0.410
√	√	×	×	60.04	0.501
√	×	√	×	54.00	0.425
√	×	×	√	56.40	0.455
√	√	√	×	61.87	0.523
√	√	×	√	58.97	0.487
√	×	√	√	61.39	0.517
×	√	√	√	52.72	0.409
√	√	√	√	**62.93**	**0.537**

As shown in Table 3, the complete model in this study demonstrates the best classification performance, which fully validates the rationality and effectiveness of the proposed model structure. Specifically, introducing different sub-modules into the basic STFCV module leads to varying degrees of performance improvement. The inclusion of the MHSA module, SW module, and TCN module resulted in increases in decoding accuracy of 7.28%, 1.24%, and 3.64%, respectively.

When multiple modules were combined, the decoding performance further improved. For instance, combining MHSA and SW modules led to a 9.11% accuracy increase; adding MHSA and TCN modules improved accuracy by 6.21%; and combining SW and TCN modules resulted in an 8.63% improvement. The best performance was achieved when all three modules—MHSA, SW, and TCN—were integrated, boosting accuracy by 10.17%. These results highlight the complementarity between different modules, and their reasonable combination significantly enhances the model's decoding ability.

On the other hand, when the three-branch parallel convolution structure in the STFCV module was replaced with a single convolution layer using a fixed-length kernels and depthwise separable convolutions (i. e., by ablating the STFCV module), the model's average decoding accuracy dropped by 10.21%. This degradation indicates that the multi-branch design of the STFCV module is crucial for effective frequency-domain feature extraction. In the original STFCV module, the three parallel branches employ different kernel lengths, which effectively act as a bank of temporal – frequency filters that capture EEG oscillations at multiple frequency bands and time scales. The features from these branches are then concatenated, providing complementary spectral information and improving the separability of SI-EEG patterns across classes. In contrast, a single-branch convolution with a fixed kernel length provides a much narrower receptive field in the frequency domain and cannot adequately model the diverse and subject-dependent frequency components present in SI-EEG signals. As a result, some discriminative spectral patterns are either suppressed or mixed, leading to weaker feature representations and a noticeable decline in decoding performance. These results further confirm that the multi-branch STFCV structure is a key component for robust and accurate SI-EEG decoding.

In conclusion, each module plays a crucial role in enhancing the model's performance. The STFCV, MHSA, and TCN modules are not simply connected in series, but rather form an organic whole that performs feature mining and fusion across three distinct dimensions: the "frequency-spatial domain, " "feature importance, " and the "long-range temporal domain," thereby improving the performance of speech imagery EEG decoding. In practical applications, the combinations of these modules can be flexibly adjusted to meet specific decoding accuracy and real-time requirements, enabling an optimal balance between performance and efficiency.

3.5 Comparative Experiments

To validate the effectiveness and advancement of the proposed method, comparative experiments and performance analysis were conducted using relevant methods that have been studied on the same public dataset. Table 4 summarizes the intra-subject (individual-level) and cross-subject (group-level) classification performance of all subjects on the BCI2020 dataset for different models. To further assess the model's generalizability, the evaluation is extended to a multi-dataset setting: Table 5 reports the corresponding intra-subject and cross-subject results on the Coretto dataset, providing an additional benchmark for cross-dataset performance.

Table 4 Comparison of classification results of different models on the BCI2020 dataset

Authors	Model	Intra-subject(*ACC*/%)	Cross-subject(*ACC*/%)	*P*-value
Pawar et al. [2022] [17]	DWT+MaxLCor	40.64±2.45	NA	< 0.05
Zheng et al. [2023] [18]	CatPCA	58.51±4.70	NA	NA
Bhalerao et al. [2025] [19]	MSSDM-SqueezeNet-JTFDF-HDC	59.07±8.26	NA	< 0.01
Ours	MMF	62.93±4.51	35.83±0.79	< 0.001

Table 5 Comparison of classification results of different models on the Coretto dataset

Authors	Model	Intra-subject(*ACC*/%)	Cross-subject(*ACC*/%)	*P*-value
Coretto et al. [2017] [27]	DWT+RF	NA	18.58±1.47	NA
Cooney et al.[2020] [30]	CNN(EEGNet)	24.46±1.75	24.90±0.93	$< 1\times10^{-7}$
Lee et al. [2020] [31]	Siamese NN	31.40±2.73	NA	< 0.001
Carvalho et al. [2024] [32]	DE-DDA+SVM	33.60	16.66	NA
Bhalerao et al. [2025] [19]	MSSDM-ResNet-JTFDF-CCA	60.80±2.17	NA	< 0.01
Ours	MMF	64.03±5.45	34.33±0.72	< 0.001

Note:The *p*-value is an important statistical measure used to assess the significance of results (where NA indicates missing values).

On the BCI2020 dataset, the proposed MMF model achieves an intra-subject decoding accuracy of 62.93% ± 4.51%, outperforming the compared models, such as Pawar et al.'s DWT + MaxLCor [17] (40.64% ± 2.45%) and Zheng et al. 's CatPCA[18] (58.51% ± 4.70%). Moreover, when compared to the best existing method, MSSDM-SqueezeNet-JTFDF-HDC [19], which achieves 59.07% ± 8.26%, the MMF model surpasses it by 3.86%.

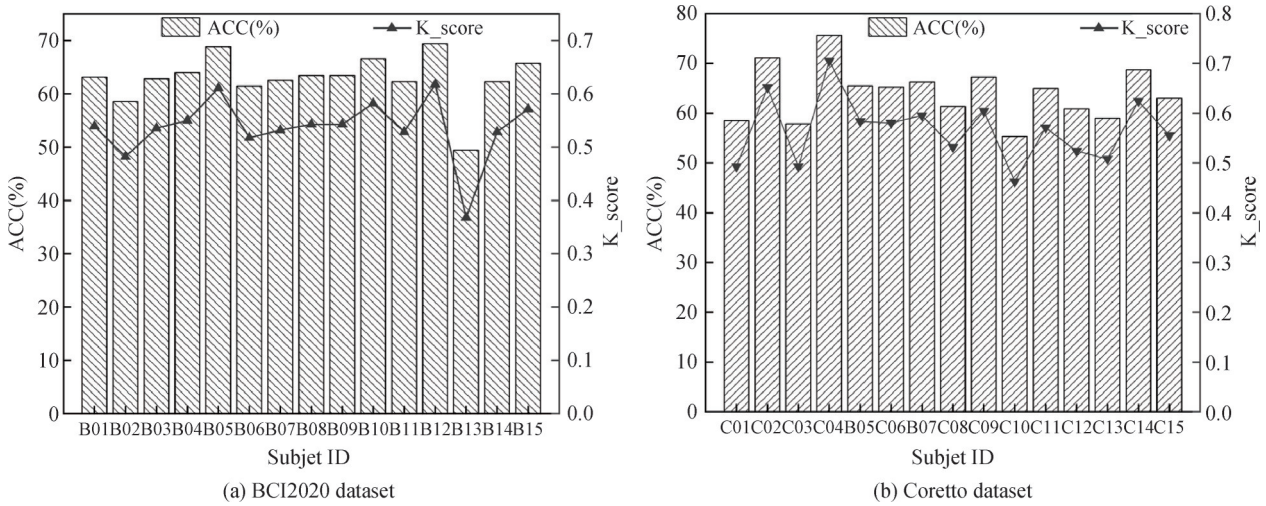
On the Coretto dataset, the MMF model achieves an intra-subject decoding accuracy of 64.03% ± 5.45%, outperforming models like Cooney et al. 's CNN (EEGNet) [30] (24.46% ± 1.75%) and Lee et al.'s Siamese NN model[31] (31.40% ± 2.73%), and surpassing MSSDM-ResNet-JTFDF-CCA[19] (60.80%±2.17%) by 3.23%.

In addition to intra-subject decoding, the MMF model also shows strong performance in cross-subject decoding. On the BCI2020 dataset, it achieves an accuracy of 35.83% ± 0.79%, while on the Coretto dataset, it achieves an accuracy of 34.33% ± 0.72%. These results demonstrate the model's generalizability across different subjects, showcasing its robustness in practical brain-computer interface applications. The improvement in both intra-subject and cross-subject accuracies indicates that the MMF model is not only effective for individual subjects but also adaptable to different subject groups.

To further validate the robustness and reliability of the proposed model, a one-sample Wilcoxon signed-rank test was conducted on the 5-fold average accuracy for each of the 15 subjects. The results (p<0.001) indicate that the MMF model's performance is statistically significantly higher than the random baseline. This reinforces the effectiveness of the MMF model, supporting its potential for practical BCI applications.

In addition to significant accuracy improvements, the proposed model also exhibits excellent computational efficiency. The average inference time per trial is approximately 0.85 s, which is relatively small compared to the speech imagery task durations in the datasets used (2 s for BCI2020 and 4 s for Coretto). This suggests that the model can operate within a feasible time frame for prospective online implementation. Moreover, the model's file size is approximately 2,400 KB, making it suitable for resource-constrained applications, such as portable brain-computer interface devices and real-time brain-machine interaction systems.

To provide a comprehensive evaluation of the decoding performance of the proposed model across different individuals, intra-subject classification results for each subject are shown in Fig. 6. The figure reports



(a) BCI2020 dataset (b) Coretto dataset

Fig.6 Intra-subject decoding performance of the proposed model on two EEG datasets: (a) BCI2020 and (b) Coretto. For each subject (x-axis), the figure reports the average accuracy (ACC, %) and average Cohen's Kappa (K_score), both computed across cross-validation folds.

the 5-fold average decoding accuracy and the corresponding 5-fold average Cohen's kappa coefficient.

As shown in Fig. 6, there are significant individual differences in EEG signals, leading to varying classification performance across subjects. On the BCI2020 dataset, the model achieved the highest decoding accuracy of 69.43% for subject B12, the second-highest accuracy of 68.86% for subject B05, and the lowest accuracy of 49.43% for subject B13. On the Coretto dataset, the highest, second-highest, and lowest decoding accuracies are 75.65% for subject C04, 71.11% for subject C02, and 55.35% for subject C10, respectively. For most subjects, the decoding accuracy lies between 58% and 70%.

Notably, subjects B05, B12, C02, and C04 performed significantly better than the others. This may be related to factors such as their level of focus, cooperation, and familiarity with the task during the experiment. These factors could influence the stability of the EEG signals, which in turn affects the model's decoding performance.

In conclusion, although some subjects exhibited lower classification performance, the small standard deviation indicates that the decoding accuracy and Kappa coefficient for most subjects remained consistent and high. This further validates the reliability and stability of the proposed model.

Fig. 7 presents the average confusion matrices for all subjects across both datasets. Overall, the decoding performance of the proposed model is superior to that of the baseline methods, with more balanced and higher classification rates observed for both the five English phrases and the six Spanish words.

For the BCI2020 dataset, clear improvements are observed for phrases such as "Hello" and "Thank you." For example, "Hello" is correctly classified 64.3% of the time (Fig. 7(d)), whereas Pawar et al.'s and Zheng et al.'s models achieve only 41% (Fig. 7(a)) and 59.0% (Fig. 7 (b)), respectively. However, the decoding accuracy for the phrases "Help me" and "Stop" remains lower than that reported by Bhalerao et al., indicating that these categories are still challenging for the proposed model.

For the Coretto dataset, a similar trend is observed. Spanish words such as "abajo" (down) and "adelante" (forward) exhibit fewer misclassifications. In particular, "adelante" is correctly identified 66.1% of the time (Fig. 7 (g)), compared with 30% in Lee et al.'s model (Fig. 7(e)) and 42.3% in Bhalerao et al. 's model (Fig. 7(f)). Nevertheless, the decoding accuracy for "atrás" and "derecha" is lower than that of Bhalerao et al.'s method, suggesting that certain word categories remain difficult to decode.

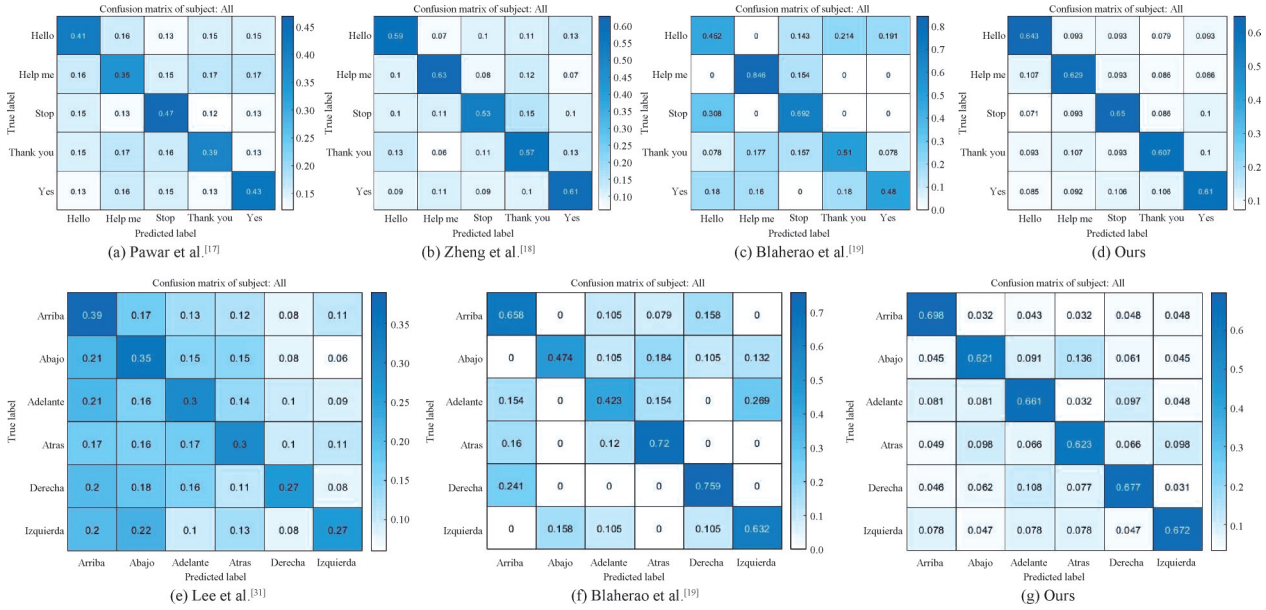


Fig.7 Average confusion matrices for all subjects across two datasets, comparing the proposed model with baseline methods. Panels (a)-(d) correspond to the BCI2020 dataset. Panels (e)-(g) correspond to the Coretto dataset. Each matrix entry is the mean classification rate averaged over subjects and cross-validation folds. Higher diagonal values indicate better per-class recognition accuracy. This layout enables direct visual comparison of classification balance and error patterns across different models and datasets.

In summary, these results indicate that the proposed model provides more accurate and stable decoding performance across both datasets, although some limitations still persist for specific phrases and words.

In these models, certain phrases such as "Stop, " "Help me," and "Thank you" are frequently misclassified, exhibiting relatively high misclassification rates. For example, in Pawar et al. 's model, "Help me" is misclassified as "Thank you" at 17%, and vice versa, while "Stop" is misclassified as "Hello" at 15%. In Zheng et al.'s model, "Stop" is misclassified as "Thank you" at 15%. In Bhalerao et al. 's model, the misclassification rates for "Stop" as "Hello" (30.8%), "Hello" as "Thank you" (21.4%), "Help me" as "Stop" (15.4%), and "Thank

you" as "Help me" (17.7%) are notably higher. By contrast, our model significantly reduces these misclassifications: "Stop" is misclassified as "Yes" at 10%, "Help me" as "Hello" at 10.7%, and "Thank you" as "Help me" at 10.7%. These values are markedly lower than those of the baseline models, highlighting the improved discriminative capability of the proposed method.

Similarly, on the Coretto dataset, certain Spanish words such as "abajo", "adelante", "derecha" and "izquierda" are frequently misclassified in the baseline models. In Lee et al.'s model, "izquierda" is misclassified as "abajo" at 22%, "abajo" as "arriba" at 21%, and "adelante" as "arriba" at 21%. In Bhalerao et al.'s model, the misclassification rates for "adelante" as "izquierda" (26.9%), "derecha" as "arriba" (24.1%), "abajo" as "atras" (18.4%), and "izquierda" as "abajo" (15.8%) are also substantial. In contrast, our model achieves significantly lower error rates: "abajo" is misclassified as "atras" at 13.6%, "adelante" as "derecha" at 9.7%, and "derecha" as "adelante" at 10.8%. These results again demonstrate a clear performance advantage over competing models.

The frequent misclassification of certain phrases and words may be attributed to intrinsic properties of SI-EEG signals. Some expressions evoke highly similar neural activation patterns, leading to overlapping spatio-temporal EEG signatures that are difficult to distinguish. Variability in individual EEG characteristics, including differences in signal-to-noise ratio and response stability, may further obscure class-specific features. In addition, the cognitive processes underlying speech imagery often involve shared articulatory and phonological pathways, resulting in partially overlapping temporal dynamics among related phrases or words. Despite these challenges, the proposed multi-module fusion framework is able to capture more subtle spatio-temporal-frequency distinctions, thereby reducing misclassification rates compared with the baseline models.

4 Conclusion

To enhance the decoding performance of SI-EEG signals, a novel deep multi-module fusion (MMF) network for SI-EEG decoding was proposed and evaluated on two public SI-EEG datasets. In this framework, a synergistic integration of the STFCV, MHSA, and TCN modules is achieved. Rather than being simply connected in series, these modules form an organic whole that performs feature mining and fusion across three complementary dimensions: frequency-spatial domain, feature importance, and long-range temporal domain.

Within the MMF framework, rich and fine-grained multi-domain features are extracted from raw SI-EEG signals by the STFCV module, allowing frequency, spatial, and temporal information to be captured simultaneously. A convolution-based sliding window data augmentation strategy is employed to segment and expand the STFCV features, thereby increasing data diversity and enhancing robustness and generalization. Within each window, the MHSA module is used to emphasize the most informative EEG features so that their expressiveness and discriminative power are strengthened. In addition, temporal dependencies in the signals are effectively modeled by the TCN module, enabling long-range temporal relationships that are essential for accurate decoding to be captured and further improving the overall SI-EEG decoding capability.

Experimental results demonstrate that both the number of sliding windows and the selection of the attention mechanism significantly impact SI-EEG decoding performance. Appropriately increasing the number of sliding windows effectively improves decoding accuracy, while the multi-head self-attention mechanism also shows a clear advantage in classification performance. The study demonstrates the efficacy of the proposed MMF framework in accurately decoding SI-EEG for both 5-class and 6-class SI recognition tasks, outperforming existing state-of-the-art approaches on the BCI2020 and Coretto datasets. Moreover, the model achieves substantial gains in cross-subject validation. By extending the evaluation to cross-subject and multi-dataset scenarios, the generalizability of the method is further confirmed. Overall, the proposed framework significantly improves SI-EEG decoding accuracy and exhibits robust performance under varying data distributions and subject groups.

In future work, the MMF model will be extended to online SI-EEG decoding. Its real-time performance (e.g., response latency during continuous decoding) will also be systematically evaluated, thereby further enhancing the practical applicability of the proposed method in real-world BCI applications.

Author Contribution:

Mengyao Yuan: Model development, data curation, manuscript drafting, and revision. Xiaoxi Yuan: Research design, methodology formulation, and supervision. Zeyi Yang and Zhi Cai: Literature review and funding acquisition. Zhengdong Zhou: Manuscript review and editing, supervision, project administration, and guidance on research design and data analysis.

Acknowledgments:

The authors gratefully acknowledge the support of the Intelligent Imaging Laboratory, the Institute of Precision Driving, and the College of Aerospace Engineering at Nanjing University of Aeronautics and Astronautics for providing the computational resources used in this work.

Foundation Information:

This research was funded by the National Natural Science Foundation of China (General Program, Grant

No. 52375570), the First Batch of "Unveiling the List and Appointing the Commander" Project of the Chinese Academy of Aeronautics (Grant No. F2021109), and the Research and Practice Innovation Program of Nanjing University of Aeronautics and Astronautics (Grant Nos. xcxjh20240110 and xcxjh20240111).

Data Availability:

The authors declare that the main data supporting the findings of this study are available within the paper and its Supplementary Information files. This study utilized a publicly available dataset, the BCI2020 dataset (2020 International BCI Competition), which can be accessed through the respective official repositories (https://osf.io/pq7vb/overview).

Conflicts of Interest:

The authors declare no competing interests.

Dates:

Received 02 November 2025; Accepted 30 March 2026; Published online 10 July 2026

References

[1] Su K, Tian L. (**2025**). Systematic review: progress in EEG-based speech imagery brain-computer interface decoding and encoding research[J]. *PeerJ Computer Science*, 11, e2938. doi:10.7717/peerj-cs.2938.

[2] Zhang L Y, Zhou Y Y, Gong P L, et al. (**2025**). Speech imagery decoding using EEG signals and deep learning: a survey[J]. *IEEE Transactions on Cognitive and Developmental Systems*, 17(1), 22-39. doi: 10.1109/TCDS.**2024**.3431224.

[3] Edelman B J, Zhang S, Schalk G, et al. (**2025**). Non-invasive brain-computer interfaces: state of the art and trends [J]. *IEEE Reviews in Biomedical Engineering*, 18, 26-49. doi: 10.1109/RBME.**2024**.3449790.

[4] Musallam Y K, Alfassam N I, Ghulam M, et al. (**2021**). Electroencephalography-based motor imagery classification using temporal convolutional network fusion [J]. *Biomedical Signal Processing and Control*, 69, 102826. doi: 10.1016/j.bspc.**2021**.102826.

[5] Liu M, Lin C, Han J, et al. (**2023**). A review of brain-computer interface decoding algorithms based on P300 features [J]. *Journal of Signal Processing*, 39(8), 1367-1385. doi:10.16798/j.issn.1003-0530.**2023**.08.004.

[6] Li J X, Dai F Z, Yin D, et al. (**2023**). A method of SSVEP signal identification based on improved eCAA [J]. *Instrumentation*, 10(4), 1-11. doi: 10.15878/j. cnki. instrumentation.20231128.001.

[7] Brusini L, Stival F, Setti F, et al. (**2021**). A systematic review on motor-imagery brain-connectivity-based computer interfaces [J]. *IEEE Transactions on Human-Machine Systems*, 51(6), 725-733. doi:10.1109/THMS.**2021**.3115094.

[8] Mohan A, Anand R S. (**2024**). Exploring diverse augmentation strategies for imagined speech detection in brain-computer interface using machine learning models [C]. In IEEE (Eds.), 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2340-2345. IEEE, Kuching, Malaysia. doi:10.1109/SMC54092.**2024**.10831467.

[9] Zhou Y F, Zhang L W, Zhou Z D, et al. (**2024**). Classification of group speech imagined EEG signals based on attention mechanism and deep learning [J]. *Journal of Zhejiang University (Engineering Science)* , 58(12), 2540-2546. doi: 10.3785/j.issn.1008-973X.**2024**.12.013.

[10] Zhang L W, Zhou Z D, Xu Y F, et al. (**2022**). Classification of imagined speech EEG signals with DWT and SVM [J]. *Instrumentation* , 9(2), 56-63. doi: doi. org/10.15878/j. cnki. instrumentation.**2022**.02.004

[11] Liu Y P, Gong A M, Ding P, et al. (**2022**). Key technology of brain-computer interaction based on speech imagery [J]. *Journal of Biomedical Engineering*, 39(3), 596-611. doi: 10.7507/1001-5515.202107018.

[12] Panachakel J T, Ramakrishnan A G. (**2021**). Decoding covert speech from EEG—a comprehensive review [J]. *Frontiers in Neuroscience*, 15, 642251. doi:10.3389/fnins.**2021**.642251.

[13] Das A, Singh S, Kim J, et al. (**2025**). Enhanced EEG signal classification in brain computer interfaces using hybrid deep learning models[J]. *Scientific Reports*, 15(1), 27161. doi: 10.1038/s41598-025-07427-2.

[14] Berg B V D, Donkelaar S V, Alimardani M. (2021). Inner speech classification using EEG signals: a deep learning approach [C]. In IEEE (Eds.), 2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS), 1-4. IEEE, Magdeburg, Germany. doi: 10.1109/ICHMS53169.**2021**.9582457.

[15] Li F, Chao W B, Li Y, et al. (**2021**). Decoding imagined speech from EEG signals using hybrid-scale spatial-temporal dilated convolution network [J]. *Journal of Neural Engineering*, 18(4), 0460c4. doi:10.1088/1741-2552/ac13c0.

[16] Ahn H J, Lee D H, Jeong J H, et al. (**2023**). Multiscale convolutional transformer for EEG classification of mental imagery in different modalities [J]. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31, 646-656. doi:10.1109/TNSRE.**2022**.3229330.

[17] Pawar D, Dhage S. (**2022**). Imagined speech classification using EEG based brain-computer interface[C]. In IEEE (Eds.), 2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT), 662-666. IEEE, Indore, India. doi: 10.1109/CSNT54456.**2022**.9787644.

[18] Zheng X B, Ling B, Zheng S Y, et al. (**2023**). Supervised categorized principal component analysis for imagined speech classification via applying singular value decomposition on a symmetry matrix [J]. *Biomedical Signal Processing and Control*, 86(Pt C), 105324. doi:10.1016/j.bspc.**2023**.105324.

[19] Bhalerao S V, Pachori R B. (**2025**). Imagined speech – EEG detection using multivariate swarm sparse decomposition-

based joint time-frequency analysis for intuitive BCI[J]. *IEEE Transactions on Human-Machine Systems*, 55(3), 347-357. doi:10.1109/THMS.**2025**.3554449.

[20] Ding Y, Robinson N, Zeng Q H, et al. (**2020**). TSception: A deep learning framework for emotion detection using EEG [C]. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 1-7. doi:10.1109/IJCNN48605.**2020**.9206750.

[21] Schirrmeister R T, Springenberg J T, Fiederer L D J, et al. (**2017**). Deep learning with convolutional neural networks for EEG decoding and visualization [J]. *Human Brain Mapping* , 38(11), 5391-5420. doi:10.1002/hbm.23730.

[22] Jin X, Zhu F, Shen Y, et al. (**2025**). Data-driven dynamic graph convolution transformer network model for EEG emotion recognition under IoMT environment[J]. *Big Data Mining and Analytics*, 8(3), 712-725. doi: 10.26599/BDMA.**2024**.9020071.

[23] Brauwers G, Frasincar F. (**2023**). A general survey on attention mechanisms in deep learning[J]. *IEEE Transactions on Knowledge and Data Engineering*, 35(4), 3279-3298. doi: 10.1109/TKDE.**2021**.3126456.

[24] Altaheri H, Muhammad G, Alsulaiman M. (**2023**). Physics-informed attention temporal convolutional network for EEG-based motor imagery classification [J]. *IEEE Transactions on Industrial Informatics*, 19(2), 2249-2258. doi: 10.1109/TII.**2022**.3197419.

[25] Bi J Y, Wang F, Ping J Y, et al. (**2024**). FBN-TCN: Temporal convolutional neural network based on spatial domain fusion brain networks for affective brain-computer interfaces [J]. *Biomedical Signal Processing and Control*, 94, 106323. doi: 10.1016/j.bspc.**2024**.106323.

[26] BCI Competition Committee. (**2022**). 2020 International BCI Competition. OSF. doi:10.17605/OSF.IO/PQ7VB.

[27] Coretto G A P, Gareis I E, Rufiner H L. (**2017**). Open access database of EEG signals recorded during imagined speech[C]. In Proceedings of the SPIE, 12th International Symposium on Medical Information Processing and Analysis, 1016002. doi: 10.1117/12.2255697.

[28] Woo S, Park J, Lee J Y, et al. (**2018**). CBAM: convolutional block attention module[C]. In Ferrari, V., Hebert, M., Sminchisescu, C., & Weiss, Y. (Eds.), Computer Vision – ECCV 2018, 3-19. Springer, Cham. doi: 10.1007/978-3-030-01234-2_1.

[29] Hu J, Shen L, Sun G. (**2018**). Squeeze-and-excitation networks[C]. In IEEE/CVF (Eds.), 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7132-7141. IEEE, Salt Lake City, UT, USA. doi: 10.1109/CVPR.**2018**.00745.

[30] Cooney C, Korik A, Folli R, et al. (**2020**). Evaluation of hyperparameter optimization in machine and deep learning methods for decoding imagined speech EEG[J]. *Sensors (Basel)*, 20(16): 4629. doi:10.3390/s20164629.

[31] Lee D, Lee M, Lee S. (**2020**). Classification of imagined speech using Siamese neural network [C]. 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2979-2984. doi:10.1109/SMC42975.**2020**.9282982.

[32] Carvalho V R, Mendes E M A M, Fallah A, et al. (**2024**). Decoding imagined speech with delay differential analysis[J]. *Frontiers in Human Neuroscience*, 18. doi: 10.3389/fnhum.**2024**.1398065.