scientific reports

OPEN

SATrans-Net: Sparse Attention Transformer for EEG-based motor imagery decoding

Tianhua Miao[1,2,4], Liansen Sha[1,3,4], Kun Huang[1,3], Yongbin Li[1,3] & Bin Liu[1,3]✉

Brain-computer interface (BCI) technology decodes electroencephalography (EEG) signals to identify motor intentions associated with motor imagery (MI), offering assistive solutions for individuals with motor impairments. However, current deep learning methods often overlook the long-sequence nature of EEG-MI signals, leading to limited feature extraction and reduced decoding accuracy. To address this, we propose SATrans-Net, an end-to-end framework that models long-range dependencies in EEG-MI signals to enhance decoding performance. SATrans-Net uses two-dimensional depthwise separable convolution (2D-DSC) to extract spatiotemporal features and incorporates a Top-K Sparse Attention (TKSA) mechanism into the Transformer architecture, improving long-range modeling while reducing computational cost. By fusing local and global features, the model achieves accurate classification via a fully connected layer. For interpretability, Grad-CAM is applied to generate Class Activation Topography (CAT) maps, visualizing spatial attention over cortical regions. Cross-session evaluations show that SATrans-Net achieves average accuracies of 84.72%, 89.76%, and 96.79% on the BCI IV-2a, BCI IV-2b, and High-Gamma datasets, respectively, outperforming existing methods. Ablation studies further verify the critical role of the TKSA module. Overall, SATrans-Net demonstrates high decoding accuracy and strong interpretability, paving the way for the application of computational techniques in biomedical signal processing. Source Code:https://github.com/Jasmin-Tianhua/EEG-research_SATrans-Net

Keywords Brain-computer interface (BCI), EEG-MI decoding, Deep learning, Long-range dependency, Transformer

Brain-computer interface (BCI) technology enables human-computer interaction by decoding brain signals and has demonstrated significant potential in neural rehabilitation and assistive control. In particular, it enhances the quality of life for individuals with disabilities by facilitating rehabilitation training, prosthetic control, and smart home applications[1–4]. Motor imagery (MI) tasks play a crucial role in BCI systems, allowing individuals with physical limitations to control external devices using EEG signals, thereby offering new possibilities for rehabilitation and intelligent control[5]. However, EEG signal decoding remains challenging due to the high dimensionality, dynamic nature, and strong temporal dependencies of EEG signals, making feature extraction more difficult[6,7]. Additionally, EEG signals are highly susceptible to noise. While traditional methods have made progress, they struggle to balance computational efficiency and accuracy in large-scale data processing. Therefore, developing efficient and highly generalizable decoding models has become a key research focus[6].

In the field of electroencephalography (EEG) signal decoding, traditional methods primarily relied on handcrafted feature extraction techniques, such as power spectral density, time-frequency analysis, and wavelet transforms[8]. While these approaches are straightforward, their capacity to model complex patterns remains limited. With the advent of machine learning and deep learning, researchers employed algorithms such as support vector machines (SVM), random forests, and neural networks like multilayer perceptrons (MLP) to extract EEG features[3]. Although these methods enhanced feature representation, they remained dependent on manual feature design. Subsequently, convolutional neural networks (CNNs) emerged as a pivotal tool for EEG feature extraction, leveraging their strengths in processing spatial structures and sequential data[6]. CNN applications diverge into two categories: one transforms EEG signals into heatmaps, feeding them into CNNs

[1]Suzhou Institute of Biomedical Engineering and Technology,Chinese Academy of Sciences, Suzhou 215000, Jiangsu, China. [2]School of Information Science and Engineering, Hohai University, Changzhou 213200, Jiangsu, China. [3]School of Biomedical Engineering (Suzhou), Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei 230000, Anhui, China. [4]These authors contributed equally to this work: Tianhua Miao and Liansen Sha. ✉email: liubin@sibet.ac.cn

as images to exploit image processing techniques for spatiotemporal information[9]; the other treats EEG directly as signals, utilizing convolutional operations to capture local spatiotemporal features, significantly improving decoding performance over traditional and early machine learning approaches[10–14]. Notably, EEGNet[12] integrates temporal convolution and depthwise separable convolution, optimizing computational efficiency and achieving widespread use in EEG tasks. However, constrained by their local receptive fields, CNNs struggle to model long-range dependencies, exhibiting limitations when processing complex long-sequence EEG signals[13].

In recent years, attention-based models such as Transformer, inspired by their success in natural language processing, have been increasingly applied to EEG signal processing[15]. Their global attention mechanism effectively captures long-range dependencies, demonstrating strong performance in motor imagery classification[16], emotion recognition, and other EEG tasks[17–21]. Compared to RNNs and LSTMs, Transformers offer higher computational efficiency[22]. However, their weakness in local feature extraction makes it challenging to capture fine-grained patterns in EEG signals[23].

To address the respective limitations of CNNs and Transformers, their integration has emerged as a prevailing trend. For instance, EEG Conformer enhances decoding performance by jointly modeling local and global features[24].However, traditional Transformers rely on global attention mechanisms with a computational complexity of $O(N^2)$ (where N denotes the sequence length), which leads to computational redundancy when modeling EEG-MI sequences. This limitation hinders their ability to effectively capture long-range temporal dependencies, resulting in insufficient exploitation of latent information[25]. Previous studies have shown that, under appropriate task settings and temporal resolutions, leveraging longer-range features can more comprehensively extract useful information from EEG-MI signals and thus improve decoding performance[26]. To tackle these Transformer-related challenges, researchers have proposed various optimization strategies leveraging sparse attention mechanisms to reduce redundant computations and enhance long-sequence modeling capabilities[27–29]. For instance, Informer employs ProbSparse self-attention, dynamically selecting highly relevant keys via KL-divergence scoring to minimize computational overhead and focus on critical information, thereby facilitating the extraction of insights from extended sequences[29]..

In summary, this study proposes a novel deep learning model, SATrans-Net. Firstly, EEG-MI signals are processed using 2D depthwise separable convolution layers to extract spatiotemporal features, providing an efficient and enriched feature representation for subsequent analysis. These extracted features are then fed into a Transformer module with a Top-k sparse attention mechanism, which not only effectively integrates global and local features but also reduces the computational complexity of the traditional multi-head attention mechanism through sparse selection. This enables the model to more efficiently capture long-range dependencies and enhance its capability in modeling temporal dependencies. Finally, the optimized features are passed to the classification module for EEG-MI signal decoding, further improving classification accuracy and model generalization. Through this design, SATrans-Net achieves significant improvements in classification accuracy and generalization performance, offering an innovative solution for efficient EEG-MI signal decoding. In addition, to enhance the interpretability of the model, this study incorporates a Grad-CAM-based Class Activation Topography (CAT) visualization method to reveal the model's spatial attention distribution over cortical regions, thereby providing an intuitive explanation of the decision-making process. Our main contributions are:

- Tailored to the long-sequence characteristics of EEG-based motor imagery (EEG-MI), we propose SA-Trans-Net, a deep learning framework that employs 2D depthwise separable convolutions to extract spatiotemporal EEG features, enhances a Transformer architecture by incorporating a Top-K sparse attention mechanism to optimize long-range dependency modeling, and achieves accurate decoding through a classification layer. This approach significantly improves decoding performance, demonstrating superior capability in processing complex EEG signals.
- SATrans-Net delivers state-of-the-art performance in cross-session EEG-MI classification, achieving accuracies of 84.72% on the BCI IV-2a dataset, 89.76% on the BCI IV-2b dataset, and 96.79% on the High Gamma dataset, outperforming existing top-tier methods.
- We conducted ablation studies and sensitivity analysis on key factors, including Top-K sparse attention, data augmentation, convolution parameters, and module selection. Further insights were gained through matrix and t-SNE visualizations.
- We introduce a Grad-CAM-based visualization approach to project the model's attention during classification onto EEG topographic maps, providing an intuitive representation of its focus across different brain regions. This approach verifies the model's ability to capture key spatial patterns and further enhances the interpretability of the decoding process.

Related works

Deep learning methods for EEG decoding

Previous CNN-based EEG decoding studies have laid a solid foundation for the field, accumulating valuable technical insights. Pan et al.[10] proposed ShallowConvNet, which employs a shallow convolutional architecture to extract temporal EEG features while avoiding the computational complexity of deep networks. This approach ensures efficient feature representation and serves as a paradigm for lightweight model design. Schirrmeister et al.[11] introduced DeepConvNet, which utilizes a deep convolutional structure to focus on frequency-domain feature extraction, effectively capturing the spectral characteristics of EEG signals and demonstrating the potential of deep learning models. Lawhern et al.[12] proposed EEGNet, which emerged as a state-of-the-art (SOTA) framework at the time by integrating temporal convolution and depthwise separable convolution. This design significantly enhanced time-frequency feature extraction while reducing model parameters and computational burden, EEGNet achieved outstanding performance on benchmark datasets such as BCI IV-2a and BCI IV-2b, setting a standard for subsequent research. Building upon this foundation, the integration of CNN and

Transformer architectures has further advanced EEG decoding. Song et al.[24]proposed EEG Conformer, which leverages the lightweight design of ShallowConvNet[10] and the global modeling capabilities of Transformer. By enabling collaborative extraction of local and global features, EEG Conformer demonstrated superior performance on BCI IV-2a (78.66%), BCI IV-2b (84.63%), and SEED (95.30%), validating the effectiveness of hybrid architectures. Zhao et al.[30] proposed CTNet, which enhances EEG decoding performance by incorporating a Transformer encoder into the previously introduced EEGNet[12]. Therefore, the aforementioned deep learning methods have provided important inspiration for the design of SATrans-Net in this study. However, the high computational complexity of traditional multi-head attention still limits its decoding accuracy in long-sequence EEG-MI tasks, underscoring the necessity for targeted optimization in this study.

Long-sequence optimization in transformers

The computational burden of Transformer in long-sequence processing has driven the development of various optimization strategies, which serve as critical technical foundations for SATrans-Net. Wang et al.[27] proposed Linformer, which approximates the attention matrix using low-rank projections, significantly reducing memory and computational costs while demonstrating the feasibility of complexity reduction. Choromanski et al.[28] introduced Performer, leveraging random feature mapping and kernel methods to approximate Softmax attention, reducing the complexity from $O(N^2)$ to $O(N \cdot d)$, where $O(N \cdot d)$ is the dimensionality of the low-dimensional feature space. To further enhance long-sequence feature modeling, Zhou et al.[29] proposed Informer, which employs ProbSparse self-attention to dynamically select highly relevant keys via KL-divergence scoring, reducing computational redundancy and focusing on critical information. However, its probabilistic selection may overlook subtle spatiotemporal patterns in EEG-MI signals. In contrast, Chen et al.[31] introduced DRSformer, whose Top-K sparse attention (TKSA) deterministically selects the K most relevant keys and integrates multi-ratio fusion, capturing the dynamic characteristics of EEG-MI long sequences with greater precision. These methods have significantly improved the applicability of Transformer for long-sequence tasks, providing both theoretical and practical foundations for SATrans-Net in EEG-MI decoding.

Method

Overview

This paper presents SATrans-Net, an innovative deep learning architecture for EEG-MI signal classification (Fig. 1). The model uses a 2D Depthwise Separable Convolution Module (DSCM) to extract temporal and spatial features from preprocessed EEG-MI data, followed by a Top-K Transformer Encoder Module (TKTEM) to model long-range dependencies and enhance spatiotemporal relationship capture. Finally, a Fully Connected Classification Module (FCCM) consolidates the features for accurate classification. This model effectively improves decoding accuracy and generalization of EEG-MI signals.

Depthwise Separable Convolution Module (DSCM)

The Depthwise Separable Convolution Module (DSCM) designed in this study is optimized based on the EEGNet structure[12] to more effectively extract inter - channel information from the raw EEG-MI signals, thereby enabling the efficient fusion of temporal and spatial features. This module consists of three convolutional

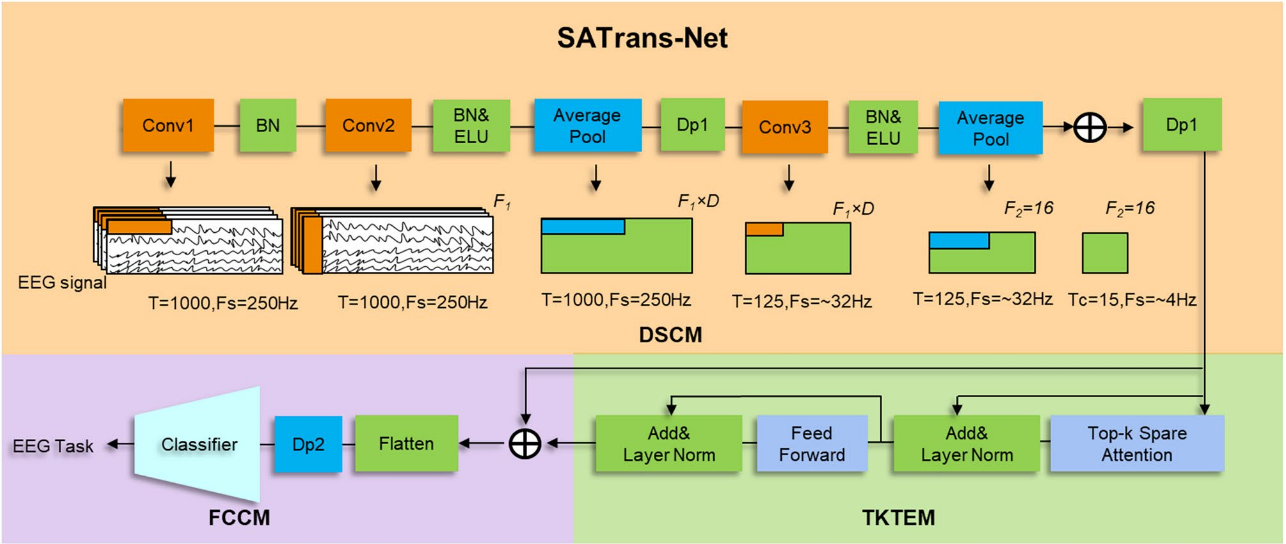


Fig. 1. illustrates the overall framework of the innovative deep learning architecture proposed in this paper, SATrans - Net. The architecture consists of three main modules: the 2D Depthwise Separable Convolution Module (DSCM), the Top - K Transformer Encoder Module (TKTEM), and the Fully Connected Classification Module (FCCM).

layers, combined with pooling, normalization, and dropout operations to enhance feature extraction and fusion capabilities.

First convolution layer: local feature extraction
The first convolution layer of the DSCM is designed to receive the preprocessed EEG-MI signals and utilizes F_1 filters of size $(1, K_{c1})$ to extract local features along the temporal dimension. By sliding the convolutional kernel along the time axis, this layer effectively captures time - spatial features beyond the 4 Hz frequency range. To ensure this property, K_{c1} is set to one - quarter of the sampling rate (64), ensuring the extraction of key time - spatial information.

To further compress the temporal dimension, the layer is followed by an average pooling layer with a size of $(1, P_1)$, which down - samples the sampling rate by a factor of 8, reducing the final frequency to approximately 32 Hz. Additionally, this layer incorporates the following key operations: Batch Normalization (BN): Accelerates training convergence and alleviates overfitting. Dropout mechanism: A dropout probability of 0.2 - 0.3 is applied to prevent overfitting.

The number of feature maps output by this layer is calculated using the following formula, where F_1 is the number of filters in the first convolution layer and *D* is the depth parameter of the deep convolution.

$$F_{out} = F_1 \times D \tag{1}$$

Second convolution layer: spatiotemporal feature learning
The second convolutional layer is primarily designed to learn the spatiotemporal features of EEG-MI signals, utilizing F_2 filters of size (*C*, 1), where *C* represents the number of EEG-MI signal channels (e.g., for the BCI IV - 2a dataset, $C = 22$; for the BCI IV - 2b dataset, $C = 3$; for the High Gamma dataset, $C = 44$).

This layer captures spatial patterns (e.g., electrode activity) for each time-domain feature map, extracting specific spatial filters across conditions. With $D = 2$, each feature map produces two spatial filters, yielding $F_1 \times D$ spatiotemporal feature maps. Like the first layer, it uses batch normalization (BN) and the ELU activation to enhance nonlinearity and address gradient vanishing.

Third convolution layer: cross - feature fusion and dimensionality reduction
The primary task of the third convolution layer is to fuse the spatiotemporal features, creating a joint representation across different features. This layer employs F_2 filters of size $(1, K_{c2})$, where K_{c2} is set to 16 to accommodate the motor imagery (MI) task within a 500ms time window.The convolution operation in this layer integrates temporal and channel - wise information, providing high - quality input for the subsequent pooling layer. Additionally, this layer is followed by an average pooling layer of size $(1, P_2)$, which controls the length of the final feature sequence (token size) to ensure that the input to the Transformer module remains of fixed length.To prevent overfitting, a dropout mechanism is also employed in this layer. The pooling operation performs downsampling on the temporal dimension, with the length of the resulting feature sequence computed as follows:

$$L_{out} = \frac{T}{P_1 \times P_2} \tag{2}$$

After the three convolution and pooling operations, the Depthwise Separable Convolution Module (DSCM) extracts high - dimensional spatiotemporal features, providing input for the subsequent Top - K Transformer Encoder Module (TKTEM) to further model long - range dependencies.

Top - K Transformer Encoder Module (TKTEM)

In the SATrans - Net architecture, the Top - K Transformer Encoder Module (TKTEM) employs the Top - K Sparse Attention mechanism (TKSA) and a position - wise Feed - Forward Network (FFN), combined with residual connections and Layer Normalization (LN), to enhance the model's training efficiency and robustness. The TKTEM is composed of *L* stacked layers of Transformer encoders, each containing two core sub - modules: TKSA and FFN (Fig. 2).

Top - K Sparse Attention Mechanism (TKSA)
The TKSA computes the attention weights between the query (*Q*) and key (*K*) by retaining only the *K* most informative attention connections, thereby reducing computational complexity and preserving the most important Spatiotemporal feature dependencies[31]. For a given input $x \in \mathbb{R}^{\text{batch} \times \text{sequence length} \times \text{embedding size}}$, the query (*Q*), key (*K*), and value (*V*) matrices are first extracted through 1D convolutional layers:

$$\begin{aligned} Q = \text{Conv1D}(x, W_Q),\ \ K = \text{Conv1D}(x, W_K), \\ V = \text{Conv1D}(x, W_V) \end{aligned} \tag{3}$$

Where W_Q, W_K and W_V represent the convolution kernel weights, which are used to extract the *Q*, *K*, and *V* structural information, respectively. Subsequently, the scaled dot - product attention scores are computed:

$$A = \frac{QK^T}{\sqrt{d_k}} \cdot \text{Temperature} \tag{4}$$

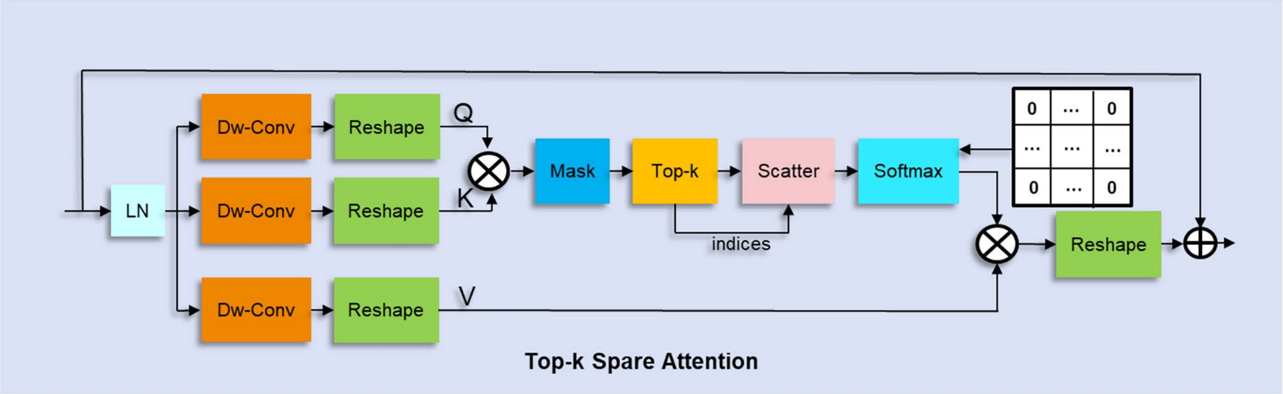


Fig. 2. First, the TKSA applies Dw-conv (depthwise separable convolution) to the input data, converting it into query (Q), key (K), and value (V) matrices. Then, the dot product between the query and key matrices is computed to obtain the attention scores. Next, a Top-K selection operation is performed, retaining only the K most relevant keys for each query, while setting the attention scores for the remaining positions to negative infinity. The filtered attention scores are then normalized using softmax to obtain the weighted value matrix. Finally, the output is mapped through a 1D convolution and optimized using residual connections and Layer Normalization (LN), resulting in the final output of the model.

In this context, d_k denotes the dimension of each attention head, and *temperature* is a learnable parameter that dynamically modulates the sensitivity of the attention distribution. This sparsity reduces computational complexity from $O(N^2)$ to $O(N \cdot K)$, retaining dominant dependencies while filtering out less informative connections[30].

Unlike the standard Transformer, which computes attention scores for all positions, we select the top *K* most relevant keys for each query point, forming a sparse attention matrix:

$$A_{i,j} = \begin{cases} A_{i,j}, & j \in \text{Top} - K(A_i) \\ -\infty, & \text{otherwise} \end{cases} \tag{5}$$

Here, the Top-K selection is based on the pre-softmax attention scores A_i, i.e., the keys corresponding to the largest dot-product values with the query are retained, while others are ignored.

Here, the value of *K* is determined by multiple ratios *r*, as follows:

$$K = r \times C \tag{6}$$

where *C* is the number of input channels. In the model, multiple different Top - K ratios are set, and the sparse attention weights are computed independently for each.

After the Top - K selection, the attention distribution for different values of *K* undergoes softmax normalization and is then weighted and fused:

$$Z = \sum_i w_i \cdot \text{softmax}(A_i)V \tag{7}$$

where w_i are learnable weight parameters used to adjust the influence of different *K* selections. Finally, we apply a 1D convolutional layer for projection and use the Dropout mechanism to prevent overfitting:

$$O = \text{Dropout}(\text{Conv1D}(Z)) \tag{8}$$

Computational complexity analysis

In the standard multi-head attention (MHA), the computation of attention scores involves all query–key pairs, leading to a complexity of $O(N^2 \cdot d)$, where *N* is the sequence length and *d* is the embedding dimension per head. In contrast, the proposed Top-K Sparse Attention (TKSA) restricts each query to interact with only the *K* most relevant keys. This reduces the complexity to $O(N \cdot K \cdot d)$, where typically $K \ll N$. As a result, TKSA not only preserves the most informative dependencies but also achieves a theoretical reduction from quadratic to linear complexity with respect to *N*, making it more efficient for long EEG sequences. Importantly, the additional overhead introduced by the fusion of multiple Top-K ratios is marginal, as it scales linearly with the number of ratios and remains significantly more efficient than standard MHA.

Residual connection and layer normalization

The output of TKSA is added to the input features via a residual connection, followed by layer normalization (LN) to standardize the feature distribution and improve the stability of gradient flow. The formula is:

$$O = \mathrm{LN}(Z + x) \tag{9}$$

Position - wise Feedforward Network (FFN)
After the computation of TKSA, the position - wise feedforward network (FFN) independently applies two linear transformations at each time - spatial position, combined with the GELU activation function and Dropout operation, to enhance the non - linear expressiveness of the features:

$$\mathrm{FFN}(x) = W_2 \cdot \mathrm{GELU}(W_1 \cdot x) + b \tag{10}$$

Where W_1 and W_2 are the weight matrices for the linear transformations. The output of FFN is then processed again with a residual connection and layer normalization, as described by:

$$O = \mathrm{LN}(\mathrm{FFN}(x) + x) \tag{11}$$

Positional encoding
To preserve the positional information of the input sequence, the module employs learnable positional encoding. The positional encoding is implemented in a parameterized form, and the encoding vector is added to the input features, as expressed by the following formula:

$$x = x + \mathrm{PositionalEncoding} \tag{12}$$

Encoder module
The overall architecture of the TKTEM consists of multiple stacked Top - K Transformer encoder layers. Each layer contains the Top - K Sparse Attention (TKSA) submodule, the position - wise Feedforward Network (FFN) submodule, and residual connections with layer normalization operations. These components are designed to capture global temporal dependencies, enhance nonlinear modeling capabilities, and improve training stability. After feature extraction and optimization through multiple encoder layers, the output global spatiotemporal features are flattened and directly passed to the linear classification head, achieving efficient integration from feature modeling to classification.

Fully Connected Classification Module (FCCM)

The Fully Connected Classification Module (FCCM) is responsible for mapping the global features extracted by the preceding modules to the target classes. Initially, FCCM flattens the output features from the TKTEM and applies a Dropout operation to effectively mitigate overfitting and enhance the model's generalization ability. Subsequently, the processed features are input into a fully connected layer consisting of *N* units, where *N* represents the number of classes in the classification task. During training, cross - entropy is used as the loss function, defined as follows:

$$L_{CE} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{N} y_{ij}\log(\hat{y}_{ij}) \tag{13}$$

where *M* is the total number of samples; *N* is the total number of classes; y_{ij} is the true label of the *j*-th class in the *i*-th sample; and $\hat{y}_{ij}$ is the predicted probability of the *j*-th class in the *i*-th sample. After processing through the FCCM, the model is able to efficiently generate the final classification results based on the optimized features.

Experimental results

The experiments in this study were conducted using three publicly available datasets for model evaluation, including the BCI IV - 2a dataset, the BCI IV - 2b dataset, and the High Gamma Dataset (HGD). These datasets encompass multiple motor imagery tasks and various experimental settings. The experiments were performed under a cross - session setup, which aligns more closely with the research needs of personalized healthcare.

Dataset

BCI IV - 2a dataset[32]
The BCI Competition IV-2a dataset, released by Graz University of Technology, is a widely used open-access benchmark for evaluating motor imagery (MI)-based brain-computer interface (BCI) systems. It contains EEG recordings from 9 subjects (A01–A09), each performing four MI tasks: imagination of left hand, right hand, both feet, and tongue movements. EEG signals were recorded using 22 Ag/AgCl electrodes at a sampling rate of 250 Hz. The data were preprocessed with a 0.5–100 Hz bandpass filter and a 50 Hz notch filter to suppress power-line interference. Each subject participated in two sessions (training and testing), with each session consisting of 288 trials-72 trials per class. In this study, we extracted the 2–6 s interval from each trial as the analysis window, and each EEG segment was represented as a (22, 1000) matrix (channels × time points).. Available at: https://www.bbci.de/competition/iv/

BCI IV - 2b dataset[32]
The BCI Competition IV-2b dataset is another publicly available benchmark released by Graz University of Technology, commonly used for evaluating binary motor imagery (MI) classification. It contains EEG recordings from 9 subjects (B01–B09), each performing two MI tasks: imagination of left hand and right hand movements.

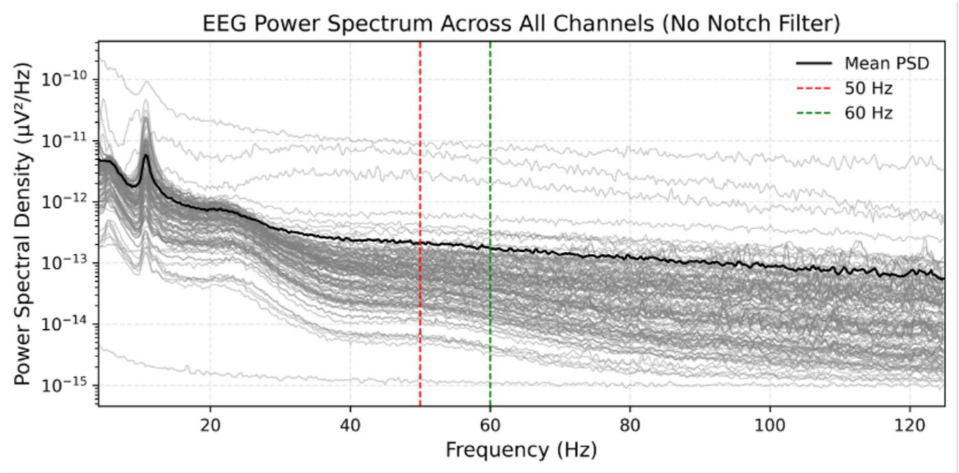


Fig. 3. Power Spectral Density (PSD) visualization of the HGD dataset.

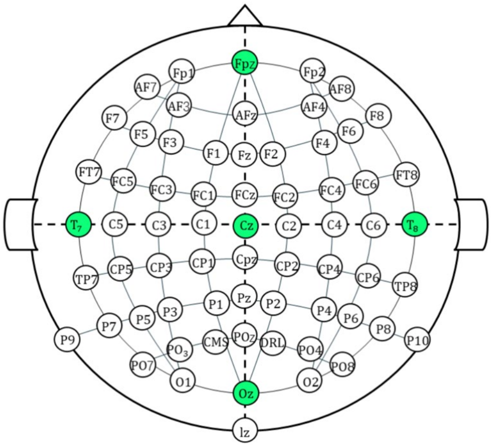


Fig. 4. Topographic Layout of the Standard 64-Channel Electrode System.

EEG signals were recorded using 3 bipolar channels (C3, Cz, C4) at a sampling rate of 250 Hz. All signals were preprocessed using a 0.5–100 Hz bandpass filter and a 50 Hz notch filter to remove power-line noise. Each subject completed five sessions, resulting in approximately 400 trials for training and 320 trials for testing. In this study, we extracted the 3–7 s interval from each trial as the analysis window. Each EEG segment was represented as a (3, 1000) matrix (channels × time points). Available at: https://www.bbci.de/competition/iv/

High Gamma Dataset (HGD)[11]

The High Gamma Dataset (HGD) comprises 14 participants (S01–S14, 6F/8M, mean age 27.2 ± 3.6) performing minimal movements: left-hand tapping, right-hand tapping, toe clenching, and resting state. Unlike traditional motor imagery (MI) tasks that rely on imagined movements and low-frequency signals (μ/β rhythms), HGD focuses on high-gamma EEG components (> 30 Hz) associated with executed actions. The signals were recorded with 128 electrodes (44 selected from the motor cortex) at 500 Hz, then downsampled to 250 Hz and bandpass-filtered (4–124.99 Hz), without applying notch filtering. As shown in Fig. 3, we visualized the power spectral density (PSD) of the HGD dataset and found minimal contamination near 50/60 Hz, indicating that additional notch filtering was unnecessary. The dataset includes 13 runs: 11 for training (880 trials) and 2 for testing (160 trials). We analyzed segments from -0.5 to 3.996 s, represented as (44, 1125) matrices. In real-world medical applications, patients such as those with amyotrophic lateral sclerosis (ALS) often retain partial motor abilities, including distal muscle movements, making HGD's focus on high-frequency signals an ideal choice for evaluating EEG decoding performance in such populations. Available at: https://braindecode.org/stable/generated/braindecode.datasets.HGD.html

Electrode selection topography

To provide a more intuitive illustration of the electrode configurations used across different datasets, we present the electrode topography based on a standard 64-channel layout (Fig. 4). For the BCI2a dataset, a total of 22 channels were selected, including FC3, FC1, FCz, FC2, FC4, C5, C3, C1, Cz, C2, C4, C6, CP3, CP1, CPz, CP2, CP4, P1, Pz, P2, and POz. In contrast, the BCI2b dataset is more compact, utilizing only three channels-C3, Cz, and C4-which correspond to the left, central, and right motor cortex areas. The High-Gamma Dataset, however, employs 44 channels, divided into two groups: standard 10–20/10-10 system electrodes (FC5, FC1, FC2, FC6, C3, C4, CP5, CP1, CP2, CP6, FC3, FCz, FC4, C5, C1, C2, C6, CP3, CPz, CP4) and high-density electrodes from the 128-channel cap positioned between standard locations (FFC5h, FFC3h, FFC4h, FFC6h, FCC5h, FCC3h, FCC4h, FCC6h, CCP5h, CCP3h, CCP4h, CCP6h, CPP5h, CPP3h, CPP4h, CPP6h, FFC1h, FFC2h, FCC1h, FCC2h, CCP1h, CCP2h, CPP1h, CPP2h).

Data preprocessing

In the data preprocessing stage, we applied Z-score normalization to all datasets to ensure consistency in EEG signal analysis across different experimental conditions[33]. The Z-score normalization formula is as follows:

$$z = \frac{x - \mu}{\sigma} \tag{14}$$

where x represents the raw EEG signal, μ is the dataset mean, and σ is the standard deviation. After normalization, the data has a mean of 0 and a standard deviation of 1, reducing inter-subject variability and signal amplitude differences, thereby improving model robustness and generalization.

Data augmentation

EEG signal acquisition is time-consuming and costly, yielding small datasets prone to overfitting[34]. Data augmentation, particularly segmentation and reconstruction (S&R), addresses this by increasing sample size and enhancing generalization[35]. We used a time-domain S&R approach, dividing samples into segments, randomly rearranging them while preserving intra-segment order, and generating new samples. Unlike traditional methods (e.g., Gaussian noise), S&R maintains signal features and temporal consistency, improving augmentation efficacy. In our implementation, two hyperparameters, number_seg and number_aug, are introduced to control the S&R-based augmentation process. Specifically, number_seg denotes the number of temporal segments each EEG trial is divided into, while number_aug represents the augmentation factor - the number of additional reconstructed samples generated from each original trial. Their specific values were set with reference to the original Segmentation and Reconstruction (S&R) paper[35]. Augmented data matching the batch size was generated per iteration, expanding the dataset and boosting model performance[36]. The details of the augmentation procedure, including segment length, number of segments, and the resulting sample counts per class, are summarized in Table 1. We summarize the procedure of the S&R-based augmentation strategy in Algorithm 1.

```
Input: $X \in \mathbb{R}^{C \times T}$: EEG trial with $C$ channels and $T$ time points
       $N_{\text{seg}}$: Number of segments to split
Output: $X_{\text{aug}} \in \mathbb{R}^{C \times T}$: Augmented EEG trial
Divide $X$ into $N_{\text{seg}}$ non-overlapping segments along the time axis;
    $\{X_1, X_2, \ldots, X_{N_{\text{seg}}}\}$ = Split($X$, axis=time);
Randomly shuffle the order of segments;
    $\{X'_1, X'_2, \ldots, X'_{N_{\text{seg}}}\}$ = Shuffle($\{X_1, \ldots, X_{N_{\text{seg}}}\}$);
Concatenate shuffled segments along the time axis;
    $X_{\text{aug}}$ = Concat($\{X'_1, \ldots, X'_{N_{\text{seg}}}\}$, axis=time);
return $X_{\text{aug}}$;
```

Algorithm 1. Segmentation & Reconstruction (S&R) for EEG Data Augmentation

Experimental setup

This study trained models on a Windows 11 platform with an AMD Ryzen Threadripper 7965X and Nvidia RTX 4090 GPU (24 GB), using PyTorch. It focused on within-subject EEG-MI classification for personalized

Dataset	#Classes	Original samples/class	Augmented samples/class	Segment length (time points)
BCI2a	4	72	$72 \times$ number_aug	1000/number_seg
BCI2b	2	200	$200 \times$ number_aug	1000/number_seg
HGD	4	220	$220 \times$ number_aug	1125/number_seg

Table 1. Number of samples per class before and after data augmentation for each dataset.

medicine, leveraging the BCI IV-2a (1000 epochs), BCI IV-2b, and High Gamma (600 epochs each) datasets. The Adam optimizer (learning rate 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$) and cross-entropy loss were used, tuned for optimal performance. Within-subject evaluation captured individual patterns, enhancing classification and supporting precision medicine[37]. The Threadripper's 24-core, 48-thread design and GPU accelerated training. The Adam setup, validated previously, balanced convergence and stability[38].

Evaluation metrics

To comprehensively evaluate the model's performance in EEG-MI signal classification, this study employs five metrics: Accuracy[39], Precision[40], Recall[41], F1-Score[42], and Cohen's Kappa[43]. These metrics collectively provide a robust quantitative analysis across multiple dimensions-global classification performance (Accuracy), class discrimination (Precision and Recall), balanced performance under class imbalance (F1-Score), and prediction consistency (Cohen's Kappa)-and are particularly suited to the complexity and imbalance inherent in EEG-MI tasks. Widely validated as standard evaluation tools in EEG-related studies, they enable a thorough assessment of the model's strengths and potential limitations without requiring additional metrics.

Accuracy

Accuracy measures the overall correctness of the model's classification, defined as the ratio of correctly predicted samples to the total number of samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

where *TP* is the number of true positives, *TN* is the number of true negatives, *FP* is the number of false positives, and *FN* is the number of false negatives.

Precision

Precision focuses on the accuracy of positive class predictions, representing the proportion of actual positive samples among all samples predicted as positive:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{16}$$

Recall

Recall measures the model's ability to capture positive class samples, representing the proportion of correctly predicted positive samples among all actual positive samples:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{17}$$

F1 - score

The F1 - Score is the harmonic mean of Precision and Recall, providing a combined evaluation of the model's classification performance, especially important in cases of class imbalance:

$$F1 - Score = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{18}$$

Cohen's Kappa

Finally, Cohen's Kappa measures the consistency between the model's predictions and the true labels by comparing it to random predictions. It is calculated as:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{19}$$

where p_0 is the observed accuracy:

$$p_0 = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

and p_e is the expected accuracy of random predictions, calculated as:

$$p_e = \frac{(TP + FP)(TP + FN) + (FN + TN)(FP + TN)}{(TP + TN + FP + FN)^2} \tag{21}$$

Baseline comparison

Metric comparison

We validated SATrans-Net against ShallowConvNet[10], DeepConvNet[11], EEGNet[12], EEGEncoder[44], Conformer[24], and CTNet[30] on BCI IV-2a, BCI IV-2b, and High Gamma datasets, using accuracy, precision, recall, F1-score, and Kappa. ShallowConvNet extracts spatiotemporal features with low cost, DeepConvNet boosts time-spatial extraction, EEGNet uses lightweight convolutions, EEGEncoder employs modified Transformers and Temporal

Methods	A01	A02	A03	A04	A05	A06	A07	A08	A09	Avg ACC (%)	Kappa
ShallowConvNet	82.64	55.21	92.01	74.31	72.92	59.72	81.60	83.33	79.51	75.69	0.68
DeepConvNet	82.29	44.79	90.63	76.04	77.43	68.06	92.01	83.33	85.42	77.78	0.70
EEGNet	88.19	56.94	93.06	71.18	70.49	62.85	87.15	82.64	84.03	77.39	0.70
Conformer	87.15	56.25	89.23	73.61	61.81	61.46	90.97	81.60	84.72	76.31	0.70
CTNet	**90.97**	**72.92**	92.01	82.98	76.04	64.93	90.28	87.50	**88.54**	82.90	0.77
EEGEncoder	84.03	72.57	93.06	81.06	**84.03**	73.61	**95.83**	89.24	87.50	84.55	–
SATrans-Net	87.15	69.44	**95.93**	**84.38**	80.21	**73.61**	94.44	**89.93**	87.5	**84.72**	**0.80**

Table 2. Baseline Comparison on the BCI IV-2a Dataset.

Methods	B01	B02	B03	B04	B05	B06	B07	B08	B09	Avg ACC (%)	Kappa
ShallowConvNet	77.81	61.79	83.13	97.50	93.13	83.44	92.50	91.88	85.00	85.13	0.70
DeepConvNet	75.00	67.50	81.56	97.81	91.56	82.50	90.31	93.13	87.50	85.21	0.70
EEGNet	78.75	67.50	85.94	97.50	94.69	90.00	93.13	92.50	89.38	87.71	0.75
EEGEncoder	73.13	67.50	79.06	97.19	96.88	78.75	91.56	90.00	87.50	84.62	-
Conformer	**82.50**	65.71	63.75	98.44	86.56	**90.31**	87.81	94.38	**92.19**	84.63	0.70
CTNet	78.75	71.70	84.38	97.19	97.81	73.96	94.06	94.69	90.63	88.49	0.77
SATrans-Net	**82.50**	**72.86**	**87.81**	**99.06**	**100**	88.13	**95.00**	**95.00**	87.50	**89.76**	**0.80**

Table 3. Baseline Comparison on the BCI IV - 2b Dataset.

	CTNet					SATrans-Net				
	ACC (%)	Precision (%)	Recall (%)	F1 (%)	Kappa	ACC (%)	Precision (%)	Recall (%)	F1 (%)	Kappa
H01	**100**	**100**	**100**	**100**	**1.00**	**100**	**100**	**100**	**100**	**1.00**
H02	91.25	91.48	91.25	91.23	0.88	**93.13**	**93.46**	**93.13**	**93.06**	**0.88**
H03	99.38	99.39	99.38	99.37	0.99	**100**	**100**	**100**	**100**	**1.00**
H04	99.38	99.39	99.38	99.37	0.99	**100**	**100**	**100**	**100**	**1.00**
H05	**100**	**100**	**100**	**100**	**1.00**	**100**	**100**	**100**	**100**	**1.00**
H06	95.00	95.45	95.00	94.98	0.93	**97.50**	**97.65**	**97.50**	**97.51**	**0.97**
H07	86.25	86.75	86.25	86.27	0.82	**87.50**	**88.48**	**87.50**	**87.40**	**0.83**
H08	91.25	91.31	91.25	91.25	0.88	**96.88**	**97.11**	**96.88**	**96.87**	**0.96**
H09	98.75	98.75	98.75	98.75	0.98	**100**	**100**	**100**	**100**	**1.00**
H10	95.00	94.98	95.00	94.91	0.93	**96.88**	**96.90**	**96.88**	**96.83**	**0.96**
H11	97.50	97.57	97.50	97.51	0.97	**99.38**	**99.39**	**99.38**	**99.37**	**0.99**
H12	97.50	97.50	97.50	97.49	0.97	**98.75**	**98.75**	**98.75**	**98.75**	**0.98**
H13	95.00	95.05	95.00	94.99	0.93	**97.50**	**97.53**	**97.50**	**97.50**	**0.97**
H14	78.13	83.36	78.13	78.19	0.71	**87.50**	**90.11**	**87.50**	**87.50**	**0.83**
Avg	94.60	95.07	94.60	94.59	0.93	**96.70**	**97.03**	**96.70**	**96.77**	**0.96**

Table 4. Baseline Comparison on the High - Gamma Dataset.

Convolutional Networks to capture temporal and spatial features, Conformer combines CNN and Transformer, and CTNet adds Transformer and channel attention. SATrans-Net, fusing time-spatial features, channel attention, and multi-head attention, outperforms baselines in EEG-MI tasks, showing robustness in signal decoding.

To evaluate the performance of SATrans - Net on the BCI IV - 2a and IV - 2b datasets, we directly refer to the experimental results of ShallowConvNet, DeepConvNet, and EEGNet reported in CTNet[30]. These experimental results have been rigorously validated, and the data processing and experimental settings are consistent with those in this study, ensuring a high level of comparability and reference value. The experimental results for the BCI IV - 2a dataset are shown in Table 2, and the experimental results for the BCI IV - 2b dataset are shown in Table 3.

In the experiment on the High - Gamma dataset, due to the distinct nature of this task compared to traditional resting motor imagery tasks, we chose to compare with the most similar model (CTNet) to more accurately assess the performance of SATrans - Net in decoding tasks. The experimental results are shown in Table 4. It is worth noting that although the High - Gamma dataset task includes motor imagery components, it also involves a mix of distal muscle movement tasks. Therefore, the focus of this study is to evaluate the model's ability to decode high - frequency EEG signals from this dataset, rather than making a direct comparison with traditional resting

	BCI IV-2a					BCI IV-2b				
Subject	ACC (%)	Precision (%)	Recall (%)	F1 (%)	Kappa	ACC (%)	Precision (%)	Recall (%)	F1 (%)	Kappa
S01	76.04	76.36	76.04	75.96	0.68	80.00	80.54	80.00	79.91	0.60
S02	54.17	55.09	54.17	53.63	0.38	73.38	73.84	73.38	73.25	0.46
S03	85.24	85.47	85.24	85.00	0.80	68.06	70.67	68.06	67.01	0.36
S04	60.24	65.27	60.24	59.16	0.47	82.16	86.38	82.16	81.63	0.64
S05	55.38	50.09	55.38	46.46	0.41	86.62	86.62	86.62	86.62	0.73
S06	56.42	58.24	56.42	56.21	0.42	86.39	86.42	86.39	86.39	0.73
S07	78.47	78.96	78.47	78.48	0.71	83.89	84.04	83.89	83.87	0.68
S08	76.91	79.91	76.91	76.76	0.69	76.71	81.56	76.71	75.78	0.53
S09	70.31	71.30	70.31	70.15	0.60	79.58	79.97	79.58	79.52	0.59
Mean	68.13	68.96	68.13	66.87	0.58	79.64	81.12	79.64	79.33	0.59
Std	11.73	12.40	11.73	13.32	0.16	6.13	5.69	6.13	6.43	0.12

Table 5. LOSO cross-validation performance of SATrans-Net on BCI IV-2a and BCI IV-2b datasets.

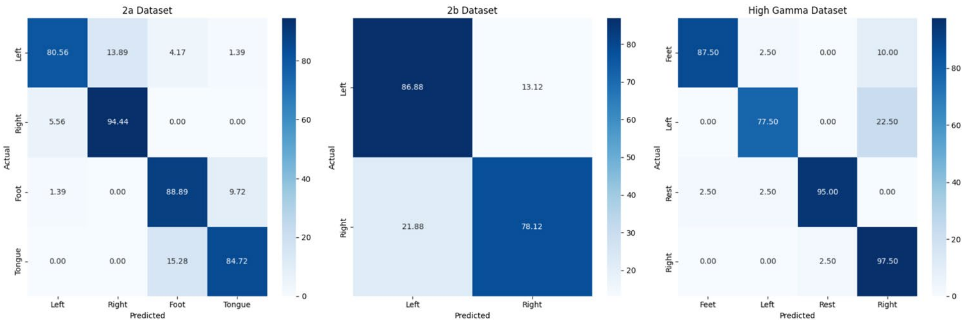


Fig. 5. Confusion Matrix Visualization Results for BCI IV - 2a, BCI IV - 2b, and HGD.

motor imagery tasks. In some subject tests, SATrans - Net achieved 100% classification accuracy, which could be attributed to the clarity of task signals in the dataset and the influence of muscle movement components. Nevertheless, the experimental results still demonstrate SATrans - Net's excellent capability in handling high - frequency brain activity and further validate the model's adaptability when dealing with different task types.

Cross-subject generalization evaluation
To verify the generalization ability of the proposed model across subjects, we conducted experiments on the BCI IV-2a and BCI IV-2b datasets using the Leave-One-Subject-Out (LOSO) cross-validation strategy. In this setting, data from $k - 1$ subjects were used for training while the remaining subject was reserved for testing in each iteration. This protocol effectively evaluates the robustness and generalization of the model under inter-subject EEG variability. The results are summarized in Table 5.

Confusion matrix visualization
We visualized SATrans-Net's classification performance using confusion matrices for the last subject of BCI IV-2a, BCI IV-2b, and High Gamma datasets (Fig. 5). For BCI IV-2a (four-class), right-hand and foot motor imagery achieved 94.00% and 88.89% accuracy, while left-hand was 80.56%, often confused with right-hand. For BCI IV-2b (two-class), left and right-hand accuracies were 86.88% and 78.12%. For High Gamma (four-class), right-hand and resting state reached 97.50% and 95.00%, but left-hand was 77.50%, frequently mistaken for right-hand. Performance varies by category, likely due to signal distribution and interference.

Sensitivity and specificity heatmap
To provide a more intuitive illustration of classification performance across subjects and classes, we generated sensitivity and specificity heatmaps for the BCI2a dataset (Fig. 6), BCI2b dataset (Fig. 7), and HGD dataset (Fig. 8). From these heatmaps, it can be observed that in the BCI2a dataset, sub2 exhibits the lowest sensitivity for foot motor imagery (0.43), while sub3 achieves the highest sensitivity for right-hand motor imagery, and sub7 reaches 0.99 for tongue motor imagery. Regarding specificity, sub2 records the lowest value for left-hand motor imagery (0.75), whereas sub7 attains perfect specificity (1.00) for the same task. In the BCI2b dataset, sub1 shows both the lowest sensitivity for right-hand motor imagery and the lowest specificity for left-hand motor imagery. In the HGD dataset, although the overall recognition accuracy is relatively high, sub14 demonstrates a markedly reduced sensitivity (0.70) for the left-hand class, while maintaining strong performance across other classes and specificity. These observations suggest that even with balanced class distributions, individual variability or signal quality issues may still lead to substantial performance degradation in specific cases. Overall, this heatmap-based

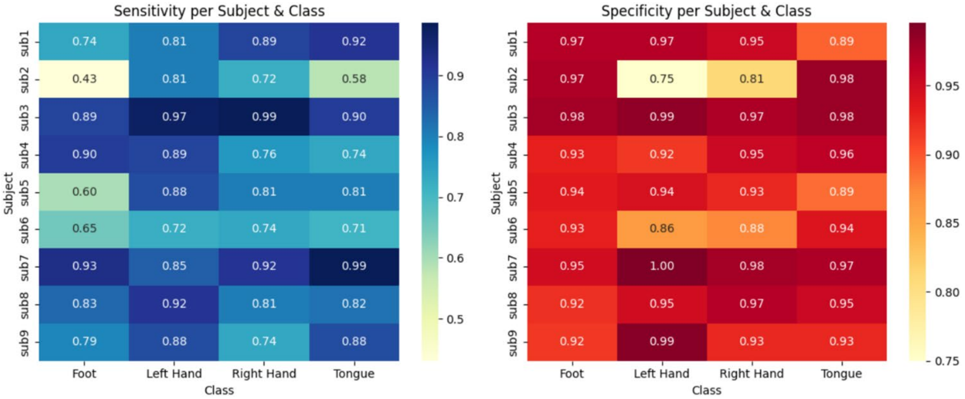


Fig. 6. BCI2a Dataset Sensitivity & Specificity.

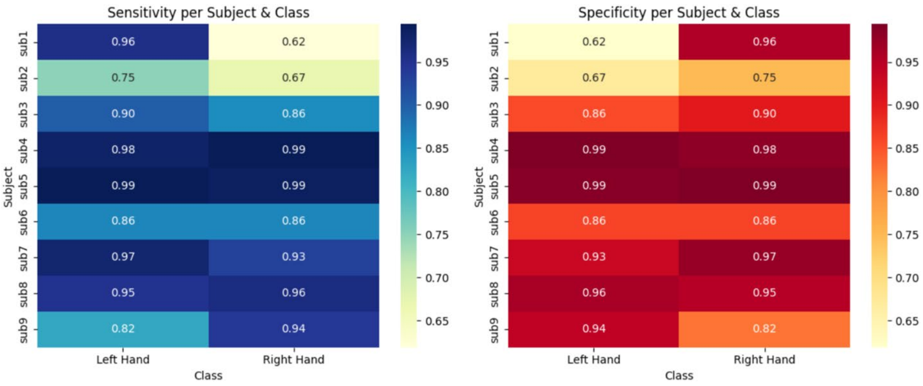


Fig. 7. BCI2b Dataset Sensitivity & Specificity.

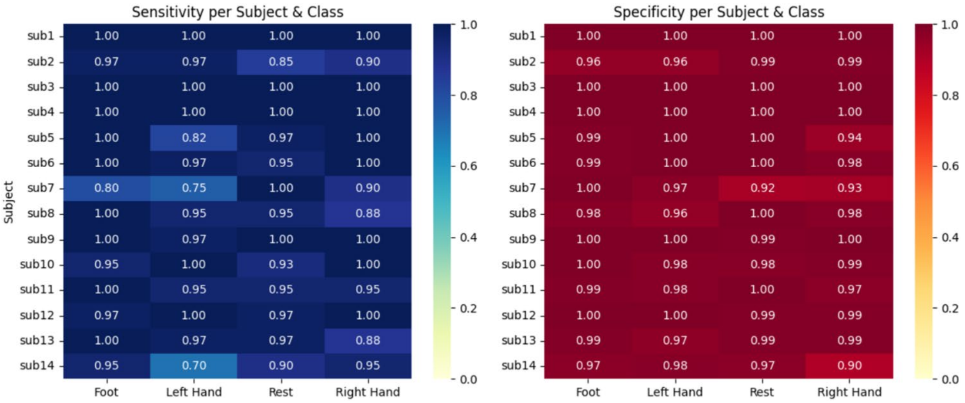


Fig. 8. HGD Dataset Sensitivity & Specificity.

visualization not only provides a clear overview of performance differences but also offers valuable insights into individual variability and model adaptability.

Training process

In image processing tasks, Transformer models typically rely on large-scale pretraining data to achieve superior performance in downstream tasks. However, the SATrans-Net proposed in this study does not utilize a pre-trained model, as the data available for calibration is relatively limited. Figure 9 illustrates the trend of loss and accuracy during the training process.We visualized the training and validation loss and accuracy. Under the cross-session paradigm, we selected sub3 from the BCI2a dataset, sub5 from the BCI2b dataset, and sub5 from the HGD dataset as representative examples. After introducing the lightweight Top-K sparse attention

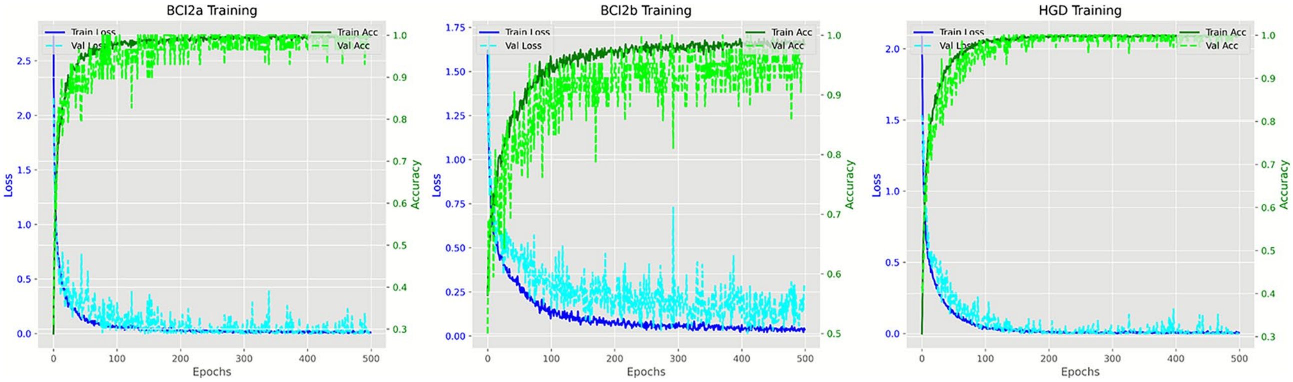


Fig. 9. Visualization of Loss and Accuracy During the Training Process.

Transformer module, the training process remained stable, with the BCI2a dataset converging at around 100 epochs, the BCI2b dataset converging at around 200 epochs, and the HGD dataset converging at around 100 epochs, indicating the model's strong learning ability.

Ablation study

Ablation verification

Ablation experiments were conducted to assess the contributions of SATrans-Net modules: (1) removing the TKTEM to evaluate its role in modeling global spatiotemporal dependencies; (2) removing data augmentation to examine its effect on mitigating overfitting and enhancing model generalization; and (3) removing both components to analyze their combined impact. The results (Fig. 10) indicate that the simultaneous removal of TKTEM and data augmentation leads to the largest drop in accuracy: 3.97% for BCI IV-2a, 2.16% for BCI IV-2b, and 2.5% for the HGD dataset. Removing data augmentation alone reduces accuracy by 3.82% (BCI IV-2a), 1.79% (BCI IV-2b), and 2.11% (HGD), while the removal of TKTEM also significantly degrades performance. These findings demonstrate that both TKTEM and data augmentation substantially improve decoding accuracy and enhance model robustness.

t - SNE visualization

In the ablation experiments, we employed t-SNE to perform dimensionality reduction and visualize the high-dimensional features extracted by the model (see Fig. 11). The complete model (Fig. 11a) displays compact and well-separated clusters, reflecting strong class discriminability. After removing data augmentation (Fig. 11b), the feature distribution becomes more scattered and the class boundaries appear blurred, indicating a reduction in robustness. When the Top-K sparse attention mechanism is replaced with standard multi-head attention (Fig. 11c), the overall feature distribution remains relatively compact; however, the separability among right hand, foot, and tongue classes deteriorates. Notably, a comparison between Fig. 11a and c shows that standard attention yields denser feature representations, which may result from the Top-K mechanism's sparsification strategy suppressing redundancy while inadvertently weakening the expression of marginal features.

When both the Top-K sparse attention and data augmentation modules are removed (Fig. 11d), the feature distribution becomes disorganized and class-wise clustering disappears. Furthermore, removing the Transformer encoder module (TKTEM) alone (Fig. 11e) results in sparse feature distributions, with indistinct class boundaries and reduced inter-class distance. In the absence of both TKTEM and data augmentation (Fig. 11f), the feature representations are highly dispersed, exhibiting poor intra-class compactness and severely degraded inter-class separability. These findings demonstrate the critical role of the TKTEM module in modeling high-dimensional representations and highlight the contribution of data augmentation in enhancing the model's robustness and generalization under input perturbations.

Parameter sensitivity

We evaluated five hyperparameters-Top-K Transformer Encoder depth, multi-head attention heads, average pooling kernel size, Top-K attention kernel size, and Top-K selection ratio-analyzing their impact on model performance. Experiments on subjects with robust accuracies provide optimization guidance and highlight module roles.

Top - K Transformer encoder depth

The experiment first evaluates the impact of the depth of the Top - K Transformer Encoder, ranging from 1 to 10, on the model's classification accuracy, as shown in Fig. 12. Overall, the model's decoding accuracy fluctuates between 86.45% and 89.93%, with an average accuracy of 88.12%. Specifically, when the depth is set to 4, the model achieves the highest accuracy of 89.93%, significantly outperforming other depth configurations, indicating that this depth strikes an optimal balance between feature extraction and model complexity. In contrast, when the depth is set to 2, 3, or 6, the accuracy is relatively lower, likely due to insufficient feature extraction capability in shallower networks or overfitting and optimization difficulties in deeper networks.

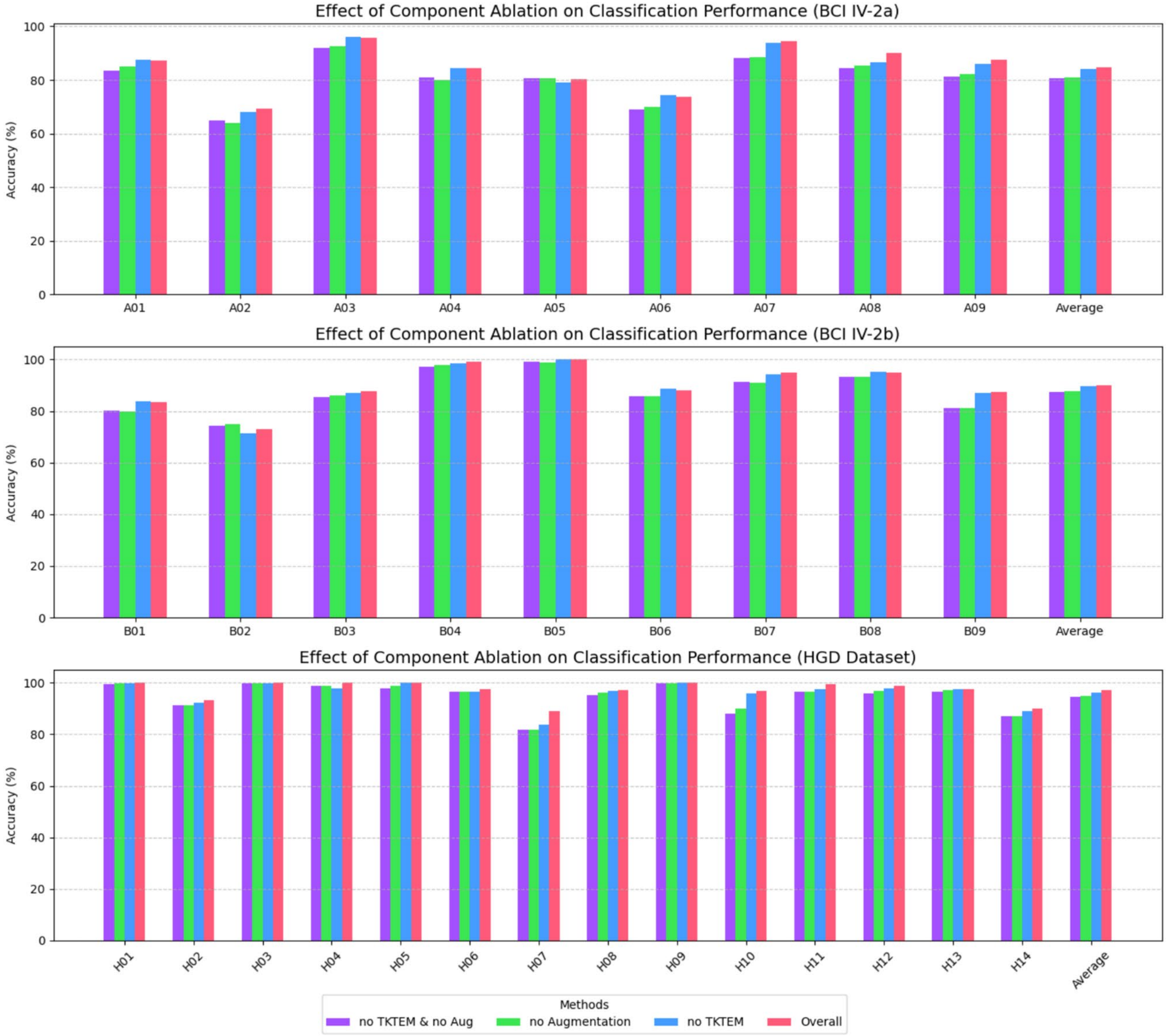


Fig. 10. Bar Chart Visualization of Ablation Study on BCI IV-2a, BCI IV-2b and HGD Datasets.

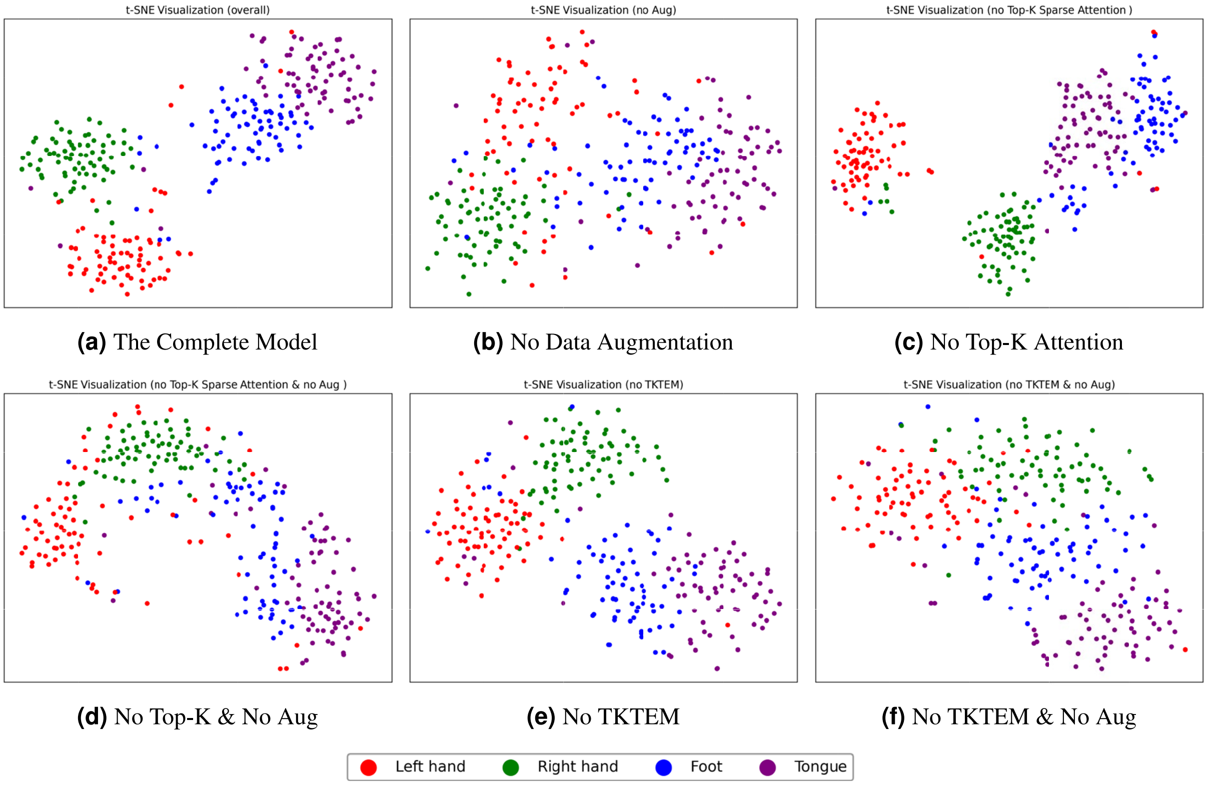
The variations in model performance across different depths further suggest that the depth of the Top - K Transformer Encoder is highly sensitive to the decoding task. For example, increasing the depth from 3 to 4 results in a 3.48% accuracy improvement, whereas increasing the depth from 4 to 5 leads to a 1.05% accuracy decline. This significant fluctuation indicates that too shallow an encoder fails to capture global information in the sequence adequately, while a deeper structure may introduce redundant features that reduce generalization ability.

In summary, a moderate depth (e.g., 4) strikes an optimal balance between model complexity and generalization ability, effectively improving classification performance. Additionally, the sensitivity analysis of the depth hyperparameter in this experiment provides important insights for the subsequent optimization of the model.

Head number in Top - K sparse attention mechanism

The head number in the Top - K sparse attention mechanism plays a crucial role in the model's decoding performance, directly influencing the model's ability to capture global dependencies and extract local features. As shown in Fig. 13, the model accuracy exhibits a certain fluctuation trend as the number of heads increases from 1 to 10. When the head number is 8, the model achieves its highest accuracy of 89.93%, and the Kappa index follows a similar trend, indicating that, at this configuration, the attention mechanism is able to more efficiently capture global information and optimize feature representation, thereby significantly improving classification performance.

In contrast, when the number of heads is smaller, the model's accuracy is relatively low and exhibits less fluctuation. This may be due to an insufficient number of heads, preventing the model from fully learning



(a) The Complete Model **(b)** No Data Augmentation **(c)** No Top-K Attention

(d) No Top-K & No Aug **(e)** No TKTEM **(f)** No TKTEM & No Aug

Fig. 11. t-SNE Visualization of Feature Distributions.

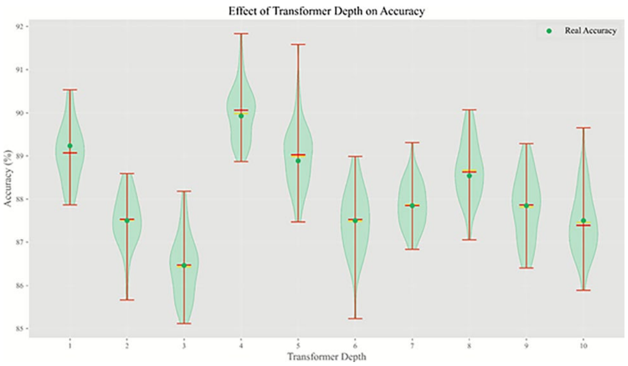


Fig. 12. Impact of Top - K Transformer Encoder Depth on Classification Accuracy.

information across different feature dimensions. When the head number is increased to 10, both accuracy and stability decrease, which could be attributed to an excess of heads leading to redundant feature decomposition, increasing the computational burden and thus affecting training efficiency and model performance.

In conclusion, a moderate head number (e.g., 8) strikes the optimal balance between global modeling capacity and computational complexity, enhancing the model's classification performance while avoiding information loss or redundancy issues associated with too few or too many heads.

Synergistic effect of heads and depth

We conducted experiments on subject 04 of the BCI IV-2a dataset to analyze the synergistic effect of the heads and depth hyperparameters, with the results shown in Fig. 14. It appears that the number of heads has a slightly more noticeable impact on the performance. Under a controlled setup, we adjusted the CNN components according to the most suitable parameter configuration. Although we also performed grid search experiments on other subjects, the model was found to be relatively insensitive to these parameters across different subjects.

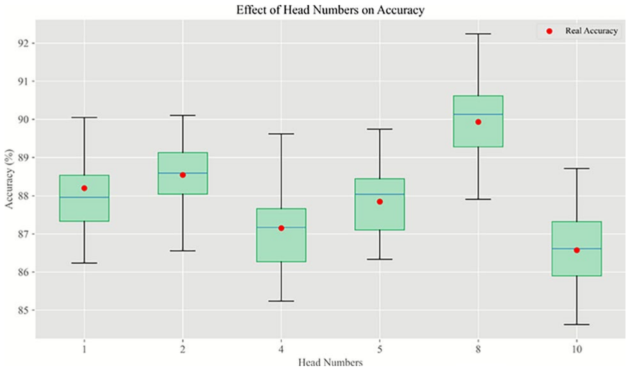


Fig. 13. Impact of Head Number on Model Performance in Top - K Sparse Attention Mechanism.

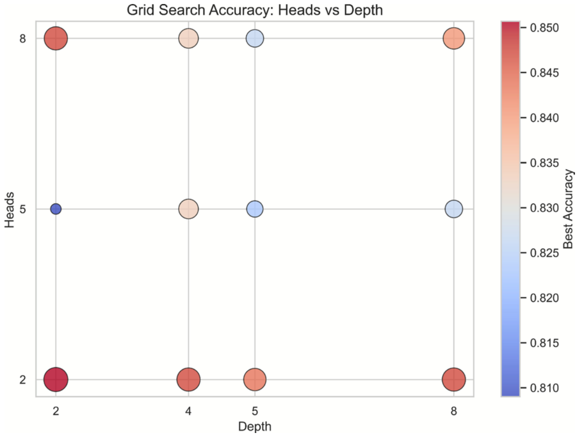


Fig. 14. Synergistic Effect of Heads and Depth.

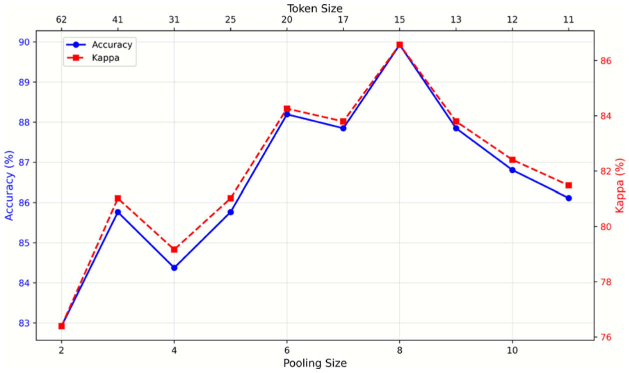


Fig. 15. Impact of Pooling Kernel Size on Model Performance.

Pooling Kernel size

SATrans-Net's convolution module has two pooling layers, with the second layer key to Top-K sparse attention Transformer performance. Pooling reduces dimensionality and complexity while affecting token size and feature integrity for the encoder.

Figure 15 shows tokens decreasing from 62 to 11 as second pooling size rises from 2 to 11. Accuracy and Kappa improve from 62 to 15 tokens, then decline. At 15 tokens (pooling size 8), accuracy peaks at 89.93%, Kappa at 86.574%.

This highlights the need for an optimal token size. Large sizes add redundancy, weakening feature aggregation, while small sizes limit expression, reducing performance. A moderate size (15) balances global and local features, boosting Top-K Transformer decoding.

Kernel Size	ACC (%)	Kappa (%)
1	88.19	84.25
3	**89.93**	**86.57**
5	87.85	83.80
7	87.15	82.87

Table 6. Impact of Kernel Size on Classification Accuracy and Kappa Value in the Top - K Attention Mechanism.

Top-K Ratios	ACC (%)	Kappa (%)
0.25	87.50	83.33
(0.25, 0.5)	87.84	83.70
(0.25, 0.5, 0.75)	88.19	84.25
(0.25, 0.5, 0.75, 0.9)	**89.93**	**86.57**

Table 7. Effect of Top-K Selection Ratios on Model Performance.

Kernel size of the Top - K Attention Mechanism

In Top-K sparse attention, convolution extracts local spatiotemporal features for the Transformer encoder. Kernel size sets the local information range, with smaller sizes capturing fine features and larger sizes covering broader sequences for global dependencies.We tested kernel sizes (1, 3, 5, 7) in Top-K attention for decoding impact (Table 6). Size 3 achieves the highest accuracy (89.93%, Kappa 0.8657), balancing local and global modeling. Small kernel size (e.g., 1) limits local perception, reducing feature extraction and decoding. Large size (e.g., 7) captures more context but adds redundancy and complexity, lowering generalization.Kernel size 5 best balances local patterns and global dependencies, enhancing Top-K attention decoding.

Top - K selection ratio (Top - K ratios)

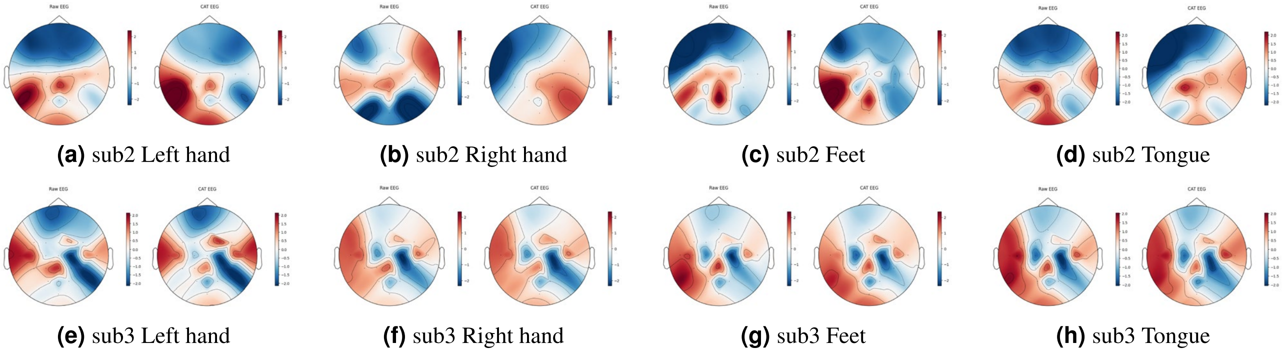
In the Top-K sparse attention mechanism, the selection ratio determines the number of key values retained for each query. In this study, multiple Top-K ratios were employed and the resulting attention outputs were fused using learnable weights[45], as shown in Table 7. Specifically, each ratio independently computes a sparse attention matrix by selecting the most relevant keys according to the pre-softmax dot-product attention scores. The outputs of these multiple Top-K computations are then combined through a learnable linear weighted sum, where the weights are trainable parameters that adaptively adjust the contribution of each ratio. Single ratios can limit performance, while the combined approach (e.g., 0.25, 0.5, 0.75, 0.9) improves adaptability and classification accuracy (89.93%, Kappa 0.8657) compared to using a single small ratio (e.g., 0.25, 87.50%). This weighted fusion enables the model to balance information from different sparsity levels, enhancing robustness and performance across tasks.

Class Activation Topography (CAT) visualization analysis

Although deep learning-based EEG decoding models have achieved significant performance improvements, their internal decision-making processes remain difficult to interpret. To enhance the interpretability of the proposed model, we applied the Grad-CAM method to project the model's spatial attention patterns during classification onto EEG topographic maps. This approach provides an intuitive visualization of the model's focus across different brain regions[46].

Based on the trained model using the representative and widely-used BCI 2a dataset, we extracted Grad-CAM weights corresponding to four motor imagery tasks-left hand, right hand, feet, and tongue. These weights were then combined with the original EEG-MI signals of each respective class to construct the Class Activation Topography (CAT) maps[47]. To reduce individual trial noise, trials within each class were first averaged. Both the original EEG topographic map (Raw EEG) and the Grad-CAM weighted CAT map were generated. Standardized raw EEG signals were then pointwise multiplied with the standardized CAM weights to produce weighted spatial EEG representations. Finally, averaging across channels yielded class-specific cortical activation distributions for each motor imagery task[24].

To better illustrate the model's spatial attention characteristics, we visualized two subjects from the test set: sub2, which had the lowest classification accuracy, and sub3, which had the highest. For sub2, the CAT maps consistently showed more focused and distinct activation patterns compared to the Raw EEG maps across all tasks. In the left-hand task (Fig. 16a), the lower-left region activation in the Raw map was further concentrated after Grad-CAM weighting. In the right-hand task (Fig. 16b), the CAT map revealed a strong focus in the lower-right region, while the Raw map showed more dispersed responses. For the feet task (Fig. 16c), the CAT map enhanced the left-lower region activations present in the Raw map. In the tongue task (Fig. 16d), the CAT map suppressed non-task-related activations, indicating a denoising effect.These overactive regions in the Raw map may reflect muscle artifacts or non-task-related brain activity, which are effectively attenuated in the CAT representation.



(a) sub2 Left hand **(b)** sub2 Right hand **(c)** sub2 Feet **(d)** sub2 Tongue

(e) sub3 Left hand **(f)** sub3 Right hand **(g)** sub3 Feet **(h)** sub3 Tongue

Fig. 16. Class Activation Topography (CAT) and Raw EEG topographies for sub2 and sub3 across four motor imagery tasks.

In contrast, the CAT maps of sub3 displayed further enhancement in task-relevant brain areas across all tasks. The left-hand (Fig. 16e) and right-hand tasks (Fig. 16f) showed intensified responses in the C4 region and broader cortical areas, respectively. The feet task (Fig. 16g) revealed reduced activation near the C3 electrode, while the tongue task (Fig. 16h) exhibited pronounced enhancement at both C4 and electrode 16 locations.

These results demonstrate that CAT maps effectively capture task-related and subject-specific attention patterns learned by the model. Compared with raw EEG topographies, CAT maps provide clearer spatial representations. The observed activation enhancement, suppression, and attention shifts indicate the model's ability to capture subtle cortical dynamics. As an intuitive visualization of the model's decision-making process, CAT maps offer valuable insights for neuroscience and biomedical research, supporting the interpretability and practical application of deep learning models in EEG decoding.

Discussion

This study proposes SATrans-Net, a novel end-to-end deep learning architecture capable of efficiently modeling long-sequence EEG-MI signals without the need for pre-training. By incorporating a Top-K sparse attention mechanism to enhance global dependency modeling and integrating 2D separable convolutions to reinforce local feature extraction, SATrans-Net effectively addresses the challenges of computational redundancy and insufficient representation in high-dimensional, long-sequence EEG decoding. Experimental results demonstrate that SATrans-Net outperforms state-of-the-art methods in cross-temporal decoding tasks. Additionally, t-SNE visualizations reveal that the features learned by the model exhibit strong class discriminability, and ablation studies confirm the synergistic contributions of each component (DSCM, TKTEM, FCCM) to the overall performance.

Importantly, beyond performance improvements, SATrans-Net also emphasizes interpretability. To this end, we introduce Class Activation Topography (CAT), a Grad-CAM-based visualization approach that maps the model's spatial attention patterns onto standardized EEG topographic layouts. This method reveals how the model allocates attention across cortical regions during classification, providing an intuitive explanation of its decision-making process. More critically, CAT offers clinically meaningful insights by linking model predictions to physiologically relevant brain areas, thereby supporting medical professionals in analyzing motor imagery tasks and promoting the application of neural decoding models in neurorehabilitation and personalized intervention.

Although individual CAT maps, such as those for sub2 and sub3, clearly reveal subject-specific attention patterns, there is substantial inter-subject variability, making it challenging to determine which brain regions consistently contribute to classification. To address this limitation in future work, CAT maps could be aggregated across subjects by normalizing Grad-CAM weights, averaging attention values across channels, and performing statistical analyses (e.g., ANOVA or permutation tests) to identify electrodes or cortical regions consistently emphasized by the model. Moreover, comparing results across different datasets, such as BCI IV-2a, high-gamma EEG covering broader cortical areas, and BCI IV-2b with only C3, C4, Cz channels, could strengthen the conclusions. These analyses would enhance interpretability and provide more robust evidence of the neural representations associated with multi-class motor imagery classification. Although this study focuses primarily on within-subject and cross-session decoding, cross-subject validation was only conducted on the BCI IV-2a and BCI IV-2b datasets. We did not perform LOSO validation on the HGD dataset, as its high-gamma signals already demonstrated strong generalization under cross-session settings. However, the cross-subject performance of HGD remains an open question and could provide further insight into the robustness of SATrans-Net. We plan to explore this direction in future work to comprehensively assess the model's generalization ability across different subjects.

Additionally, we conducted a systematic parameter sensitivity analysis to elucidate the key drivers of model performance, with experimental evidence validating their roles. Encoder depth directly governs feature extraction capability: excessive depth risks overfitting, whereas insufficient depth results in inadequate features. The number of heads in the Top-K sparse attention mechanism regulates the balance between global dependency modeling and local feature extraction. The second pooling layer in the convolutional module optimizes token scale and information integrity, preserving core features while reducing computational complexity. The kernel size in the Top-K mechanism delineates the scope of local perception, while Top-K selection ratios precisely determine the number of critical features retained. Experiments further establish that varying Top-K ratios significantly influence performance, and dynamically integrating multi-ratio attention outcomes via learnable weights markedly enhances SATrans-Net's robustness.

Although SATrans-Net is not specifically designed for runtime efficiency, its Top-K sparse attention mechanism effectively reduces redundant computations and theoretically lowers the attention complexity from $O(N^2)$ to $O(N \cdot K)$, enhancing the modeling of long-range dependencies. However, experimental results indicate that the inference time remains comparable to that of standard multi-head attention (MHA) Transformers, showing no significant difference. The primary reason for this is the incorporation of depthwise separable convolution (dwconv) when computing query-key correlations. Compared to standard linear projections, dwconv performs independent convolutions within each channel followed by pointwise convolution for feature fusion, which significantly strengthens local feature extraction and noise robustness, benefiting the discrimination of EEG-MI signals. At the same time, dwconv introduces additional convolutional cost, with complexity approximately $O(N \cdot k)$, partially offsetting the computational savings from Top-K sparsification.Therefore, although the theoretical acceleration is not fully reflected in inference time, this design trade-off remains reasonable: Top-K attention ensures efficient modeling of long-range temporal dependencies, while dwconv provides robust local feature representations. Their complementarity enables SATrans-Net to achieve classification performance far exceeding baseline models within millisecond-level inference time. In other words, SATrans-Net prioritizes maximizing decoding accuracy without increasing overall computational cost, rather than purely focusing on runtime speed, which is crucial for practical BCI applications.

Future research can focus on three main directions. First, the optimization potential of SATrans-Net remains considerable, as ablation studies indicate that the contributions of different modules still have room for improvement. Inspired by CIACNet[48], which employs a dual-branch CNN to extract rich temporal features, further enhancements can be made in the convolutional modules to improve decoding performance and feature representation. Second, the model's generalization ability and cross-subject EEG-MI decoding present promising avenues for expansion. Although this study was validated only on the representative and widely-used BCI IV-2a and BCI IV-2b datasets, the proposed model demonstrates strong generalization potential and could be extended to EEG-based applications such as emotion recognition, sleep monitoring, and depression analysis. Moreover, as this work primarily focuses on within-subject EEG-MI decoding to meet precision medicine requirements, future research could explore cross-subject EEG-MI decoding. Incorporating techniques such as transfer learning, domain adaptation, or contrastive learning may enhance the model's stability and adaptability across individuals, thereby improving its industrial applicability and cost-efficiency. Third, parameter sensitivity optimization is essential, as SATrans-Net is highly dependent on key hyperparameters, providing a foundation for further improvements. Automated hyperparameter search methods, such as Bayesian optimization or evolutionary algorithms, can efficiently identify the optimal configuration, while meta-learning approaches may enable adaptive optimization across tasks, further enhancing model performance and robustness.

Conclusion

This study presents SATrans-Net, an innovative deep learning architecture that integrates 2D separable convolutions with a Top-K sparse attention Transformer to enhance EEG-MI decoding performance. To address the limitations of standard Transformers in modeling long-range dependencies, SATrans-Net introduces structural optimizations validated across multiple benchmark datasets. Ablation and sensitivity analyses confirm the significance of key components. To improve model interpretability, we incorporate Grad-CAM and construct Class Activation Topography (CAT) maps by integrating EEG topographic representations, providing spatial insights into the model's attention patterns. This approach offers intuitive evidence for understanding decision-making processes and supports potential applications in precision medicine.

Data availability

The datasets used in this study are publicly available. Specifically, the BCI IV-2a and IV-2b datasets are available at the BCI Competition IV repository (https://www.bbci.de/competition/iv/), and the High-Gamma Dataset (HGD) is accessible via the Braindecode platform (https://braindecode.org/stable/generated/braindecode.datasets.HGD.html). Accession codes: https://github.com/Jasmin-Tianhua/EEG-research_SATrans-Net.

Received: 24 May 2025; Accepted: 27 November 2025
Published online: 04 December 2025

References

1. Karikari, E. & Koshechkin, K. A. Review on brain-computer interface technologies in healthcare. *Biophys. Rev.* **15**, 1351–1358 (2023).
2. Alonso-Valerdi, L. M., Salido-Ruiz, R. A. & Ramirez-Mendoza, R. A. Motor imagery based brain-computer interfaces: An emerging technology to rehabilitate motor deficits. *Neuropsychologia* **79**, 354–363 (2015).
3. Altaheri, H. et al. Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: A review. *Neural Comput. Appl.* **35**, 14681–14722 (2023).
4. Gu, L. et al. Functional neural networks in human brain organoids. *BME frontiers* **5**, 0065 (2024).
5. Al-Qaysi, Z. et al. Systematic review of training environments with motor imagery brain-computer interface: coherent taxonomy, open issues and recommendation pathway solution. *Heal. Technol.* **11**, 783–801 (2021).
6. Zhang, P., Wang, X., Zhang, W. & Chen, J. Learning spatial-spectral-temporal eeg features with recurrent 3d convolutional neural networks for cross-task mental workload assessment. *IEEE Transactions on neural systems and rehabilitation engineering* **27**, 31–42 (2018).

7. Phan, H. et al. L-seqsleepnet: Whole-cycle long sequence modeling for automatic sleep staging. *IEEE J. Biomed. Heal. Informatics* **27**, 4748–4757 (2023).
8. Al-Fahoum, A. S. & Al-Fraihat, A. A. Methods of eeg signal features extraction using linear analysis in frequency and time-frequency domains. *Int. Sch. Res. Notices* **2014**, 730218 (2014).
9. Vidyasagar, K. C., Kumar, K. R., Sai, G. A., Ruchita, M. & Saikia, M. J. Signal to image conversion and convolutional neural networks for physiological signal processing: A review. IEEE Access (2024).
10. Pan, J., Sayrol, E., Giro-i Nieto, X., McGuinness, K. & O'Connor, N. E. Shallow and deep convolutional networks for saliency prediction. In Proceedings of the IEEE conference on computer vision and pattern recognition, 598–606 (2016).
11. Schirrmeister, R. T. et al. Deep learning with convolutional neural networks for eeg decoding and visualization. *Hum. brain mapping* **38**, 5391–5420 (2017).
12. Lawhern, V. J. et al. Eegnet: a compact convolutional neural network for eeg-based brain-computer interfaces. *J. neural engineering* **15**, 056013 (2018).
13. Amin, S. U., Alsulaiman, M., Muhammad, G., Mekhtiche, M. A. & Hossain, M. S. Deep learning for eeg motor imagery classification based on multi-layer cnns feature fusion. *Futur. Gener. computer systems* **101**, 542–554 (2019).
14. Gao, Y., Zhang, M., Wang, J. & Li, W. Cross-scale mixing attention for multisource remote sensing data fusion and classification. *IEEE Transactions on Geosci. Remote. Sens.* **61**, 1–15 (2023).
15. Vaswani, A. Attention is all you need. In Advances in Neural Information Processing Systems (2017).
16. Li, Y., Miao, N., Ma, L., Shuang, F. & Huang, X. Transformer for object detection: Review and benchmark. *Eng. Appl. Artif. Intell.* **126**, 107021 (2023).
17. Sun, J., Xie, J. & Zhou, H. Eeg classification with transformer-based models. In 2021 ieee 3rd global conference on life sciences and technologies (lifetech), 92–93 (IEEE, 2021).
18. Deny, P., Cheon, S., Son, H. & Choi, K. W. Hierarchical transformer for motor imagery-based brain computer interface. *IEEE J. Biomed. Heal. Informatics.* (2023).
19. Sun, B., Wang, Q., Li, S. & Deng, Q. Sctrans: Motor imagery eeg classification method based on cnn-transformer structure. In 2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT), 2001–2004 (IEEE, 2024).
20. Li, C. et al. Eeg-based emotion recognition via transformer neural architecture search. *IEEE Transactions on Industrial Informatics* **19**, 6016–6025 (2022).
21. Liu, R. et al. Ertnet: an interpretable transformer-based framework for eeg emotion recognition. *Front. Neurosci.* **18**, 1320645 (2024).
22. Reza, S., Ferreira, M. C., Machado, J. J. & Tavares, J. M. R. A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent neural networks. *Expert. Syst. with Appl.* **202**, 117275 (2022).
23. Ding, Y. et al. Eeg-deformer: A dense convolutional transformer for brain-computer interfaces. *IEEE J. Biomed. Heal. Informatics.* (2024).
24. Song, Y., Zheng, Q., Liu, B. & Gao, X. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Syst. Rehabil. Eng.* **31**, 710–719 (2022).
25. Zaheer, M. et al. Big bird: Transformers for longer sequences. *In Advances in neural information processing systems* **33**, 17283–17297 (2020).
26. Bin, Y. et al. Describing video with attention-based bidirectional lstm. *IEEE transactions on cybernetics* **49**, 2631–2641 (2018).
27. Wang, S., Li, B. Z., Khabsa, M., Fang, H. & Ma, H. Linformer: Self-attention with linear complexity. arXiv preprint arXiv:2006.04768 (2020).
28. Choromanski, K. et al. Rethinking attention with performers. arXiv preprint arXiv:2009.14794 (2020).
29. Zhou, H. et al. Informer: Beyond efficient transformer for long sequence time-series forecasting. *In Proceedings of the AAAI conference on artificial intelligence* **35**, 11106–11115 (2021).
30. Zhao, W., Jiang, X., Zhang, B., Xiao, S. & Weng, S. Ctnet: a convolutional transformer network for eeg-based motor imagery classification. *Sci. Reports* **14**, 20237 (2024).
31. Chen, X., Li, H., Li, M. & Pan, J. Learning a sparse transformer network for effective image deraining. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 5896–5905 (2023).
32. Tangermann, M. et al. Review of the bci competition iv. *Front. neuroscience* **6**, 55 (2012).
33. Henderi, H., Wahyuningsih, T. & Rahwanto, E. Comparison of min-max normalization and z-score normalization in the k-nearest neighbor (knn) algorithm to test the accuracy of types of breast cancer. *Int. J. Informatics Inf. Syst.* **4**, 13–20 (2021).
34. Mumtaz, W., Rasheed, S. & Irfan, A. Review of challenges associated with the eeg artifact removal methods. *Biomed. Signal Process. Control.* **68**, 102741 (2021).
35. Zhang, Z., Liu, Y. & Zhong, S.-H. Ganser: A self-supervised data augmentation framework for eeg-based emotion recognition. *IEEE Transactions on Affect. Comput.* **14**, 2048–2063 (2022).
36. Shorten, C. & Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *J. big data* **6**, 1–48 (2019).
37. Kingma, D. P. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).
38. Ahmed, Z., Mohamed, K., Zeeshan, S. & Dong, X. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. Database **2020**, baaa010 (2020).
39. Bekkar, M., Djemaa, H. K. & Alitouche, T. A. Evaluation measures for models assessment over imbalanced data sets. *J Inf Eng Appl.* **3** (2013).
40. Revaud, J., Almazán, J., Rezende, R. S. & Souza, C. R. d. Learning with average precision: Training image retrieval with a listwise loss. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 5107–5116 (2019).
41. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J. & Aila, T. Improved precision and recall metric for assessing generative models. In Advances in Neural Information Processing Systems, **32** (2019).
42. Chicco, D. & Jurman, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 1–13 (2020).
43. Delgado, R. & Tibau, X.-A. Why cohen's kappa should be avoided as performance measure in classification. *PloS One* **14**, e0222916 (2019).
44. Liao, W., Liu, H. & Wang, W. Advancing bci with a transformer-based model for motor imagery classification. *Sci. Reports* **15**, 23380 (2025).
45. Dong, Y., Liu, Q., Du, B. & Zhang, L. Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification. *IEEE Transactions on Image Process.* **31**, 1559–1572 (2022).
46. Cunlin, H., Ye, Y. & Nenggang, X. Self-supervised motor imagery eeg recognition model based on 1-d mtcnn-lstm network. *J. Neural Eng.* **21**, 036014 (2024).
47. Liang, Y. et al. Fetcheeg: a hybrid approach combining feature extraction and temporal-channel joint attention for eeg-based emotion classification. *J. Neural Eng.* **21**, 036011 (2024).
48. Liao, W., Miao, Z., Liang, S., Zhang, L. & Li, C. A composite improved attention convolutional network for motor imagery eeg classification. *Front. Neurosci.* **19**, 1543508 (2025).

Author contributions

T.M., L.S., K.H., Y.L., and B.L. contributed to the study. T.M. conceptualized the study, developed the methodology, performed software development, conducted formal analysis, wrote the original draft, and prepared visualizations. L.S. curated data, conducted investigations, validated results, and reviewed and edited the manuscript. K.H. and Y.L. reviewed and edited the manuscript. B.L. conceptualized the study, supervised the project, administered the project, acquired funding, and reviewed and edited the manuscript. All authors reviewed and approved the final manuscript.

Funding

This work was supported in part by the Basic Research Program of Suzhou under Grant SJC2022011, in part by the Special Project of Basic Research on Frontier Leading Technology in Jiangsu Province under Grant BK20232008, and in part by the Key Research and Development Program of Jiangsu under Grants BE2021012-3 and BE2021012.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to B.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.