AwareNet: A Topography-Informed Spatio-Temporal Network for Motor Imagery EEG Classification

1st Mingjie Gao
School of Information Science
Yunnan University
Kunming, China
0000-0003-4962-3500

2nd Zifeng Yao
Institute of Brain and Psychological Sciences
Sichuan Normal University
Sichuan, China
0009-0009-8139-2117

***Abstract*—Accurate decoding of motor imagery electroencephalography (MI-EEG) is fundamental to brain–computer interface applications, yet remains challenging due to low signal-to-noise ratio and amplitude fluctuations. Existing approaches often underutilize spatial relationships and struggle to model long-range temporal dependencies. We propose AwareNet, an end-to-end spatio-temporal EEG decoding framework that employs a Bidirectional Temporal Convolution layer to capture both past and future contextual information, enriching temporal feature representations. Next, we design a lightweight Coordinate-aware Spatial Gating module that leverages prior knowledge of actual electrode positions to model spatial relationships. Furthermore, we introduce the Channel-aware Transformer module to capture global dependencies. CAT is an eeg-friendly attention mechanism which replaces conventional dot-product attention with cosine similarity and incorporates per-head learnable temperature scaling, enabling more stable and adaptive temporal modeling under noisy, small-sample conditions. Extensive experiments demonstrate that AwareNet consistently outperforms state-of-the-art baselines on multiple datasets. Visualization analyses further confirm that AwareNet yields more discriminative and neurophysiologically consistent representations, with our CAT module surpassing conventional MSA in training stability, decoding performance, and interpretability. Overall, AwareNet provides accurate, stable, and interpretable MI-EEG decoding, demonstrating strong potential for practical BCI deployment.**

***Index Terms*—brain-computer interface (BCI), motor imagery (MI), transformer, cosine similarity, electrode spatial prior**

I. Introduction

Brain Computer Interface (BCI) enables direct communication between the brain and external devices [1]. Owing to its non-invasiveness, high temporal resolution, and low cost, electroencephalography (EEG) remains one of the most widely used technologies for practical BCI systems [2]. Among EEG-based paradigms, motor imagery (MI), the mental rehearsal of movement without execution [3], has become a cornerstone for assistive control and neurorehabilitation applications [4]. Despite steady progress, accurate MI-EEG decoding in real-world settings is still challenging due to low signal-to-noise ratio (SNR), large training variability, and contamination from artifacts such as muscle activity and eye blinks.

Early MI-EEG decoding pipelines typically relied on hand-crafted feature extraction followed by shallow classifiers [5]. Methods such as Common Spatial Pattern (CSP) and time–frequency transforms can be effective [6], yet they often depend on task-specific assumptions and require careful tuning, limiting scalability across subjects and datasets.

Recent deep learning approaches shift toward end-to-end optimization and automatic representation learning from raw or minimally processed EEG data [7]. Convolutional neural networks (CNNs), including Shallow/Deep ConvNet [8] and EEGNet [9], have demonstrated strong performance by learning local temporal patterns and inter-channel dependencies. However, CNN-based methods still face two fundamental limitations: First, EEG signals are temporal sequences recorded from spatially discrete and relatively independent electrodes, making it difficult to fully exploit spatial relationships. Second, due to the limited size of convolutional kernels, CNNs have a restricted receptive field and can only capture local features.

A. Challenges in Electrode-based Spatial Representation

EEG signals are inherently spatiotemporal, consisting of time-series recordings from spatially distributed but anatomically discrete electrode sites [10]. A common approach is to apply spatial convolutions with kernel shapes such as (electrode, 1) to learn inter-channel representations [9]. However, such designs often fail to capture the complex spatial dependencies among electrodes. Recent efforts incorporate graphs [11], pooling-driven spatial attention [12], or channel re-weighting mechanisms [13]. While beneficial, these approaches frequently rely on manually defined connectivity or static heuristics, leaving the rich geometric priors of the electrode layout underutilized and limiting neurophysiological interpretability.

B. Global modeling in MI-EEG decoding

Attention-based models have been introduced to capture global dependencies beyond local convolutions. Hybrid CNN+Transformer architectures, such as ATCNet [14], Conformer [15], and CTNet [16], typically use CNN backbones for local feature extraction and self-attention for long-range modeling. However, Transformer-based models often require large-scale data and often suffer from training instability on MI-EEG decoding. In particular, dot-product attention is sensitive to the magnitude of feature vectors, which may lead to gradient saturation or even explosion, especially in noisy and small-sample situations.

To this end, we propose AwareNet, an attention-based end-to-end deep learning framework. The framework consists of three main components: a feature extraction module that applies Bidirectional Temporal Convolution (BTC) to capture temporal dynamics from both past and future contexts, followed by a Coordinate-Aware Spatial Gate (CSG) that leverages electrode positions to modulate spatial features in a topology-informed manner. A Channel-aware Transformer (CAT) module that replaces conventional multi-head self-attention (MSA) with cosine-similarity-based attention combined with channel-based temperature scaling, and finally, a classifier.

The main contributions of this work are as follows:

- We propose AwareNet, an end-to-end spatio-temporal EEG decoding framework. AwareNet achieves state-of-the-art performance across multiple datasets in both subject-dependent and subject-independent settings.
- We design a Bidirectional Temporal Convolution for contextual temporal modeling and a lightweight Coordinate-aware Spatial Gating module that exploits electrode-coordinate priors to enhance spatial modeling and interpretability.
- We introduce CAT, an EEG-tailored attention mechanism based on cosine similarity with per-head learnable temperature scaling, improving robustness and training stability compared to conventional dot-product attention.
- We conduct extensive ablations, parameter analyses, and visualizations to quantify the contribution of each component and to support model interpretability.

The rest of this paper is organized as follows. Section II presents the proposed method. Section III details datasets and experimental settings. Section IV reports results, and Section V provides visualizations and discussion. Section VI concludes the paper.

II. Methods

This section provide an overview of the proposed model, see Figure 1.

A. EEG signal input

The raw EEG signals are used as input, formatted as time series data across electrodes. Each sample is represented by a 2D matrix $X \in \mathbb{R}^{ch \times T}$, where ch represents the number of electrode channels and T the number of time points.

B. Network architecture

AwareNet consists of three key components: a convolution module, a Channel-aware Transformer module, and a classifier.

1) Convolution module: The convolution module integrates four main components: Bidirectional Temporal Convolution, Coordinate-aware Spatial Gate, spatial convolution, and channel convolution.

a) BTC: Temporal convolution is commonly applied as the first layer in EEG decoding models to extract time-domain features, serving as a temporal filtering mechanism. For instance, EEGNet utilizes a kernel size corresponding to half the sampling rate (125 Hz for data sampled at 250 Hz), effectively capturing frequency components above 2 Hz [9]. Similarly, CTNet adopts a more aggressive filtering strategy with a kernel size of one-fourth the sampling rate (64 Hz for data sampled at 250 Hz) [16]. Conformer, on the other hand, employs small-scale convolutions (e.g., kernel size 25) to extract fine-grained temporal features in the early stage [15]. This strategy allows the model to learn diverse temporal patterns with varying receptive fields. However, conventional temporal convolutional approaches typically rely on fixed-size sliding windows that aggregate information from preceding time points only, implicitly modeling temporal dependencies in a unidirectional (past-to-present) manner. This may limit their ability to capture bidirectional neural dynamics, as the brain processes information through both feedforward and feedback pathways [17]. While feedforward activity reflects stimulus-driven responses, feedback signals, such as attention or expectation, often manifest in later time points, which are not accessible to past-only models. To better exploit contextual information along time, we employ a Bidirectional Temporal Convolution (BTC) with two parallel paths (forward and time-reversed) to capture temporal patterns in both directions. Let $X \in \mathbb{R}^{B \times 1 \times ch \times T}$ be the input. The forward path is

$$F_{\text{fwd}} = \text{Conv2D}_{\text{fwd}}(X) \in \mathbb{R}^{B \times \frac{f_1}{2} \times ch \times T}, \tag{1}$$

and the backward path applies convolution on the time-flipped signal and flips back:

$$F_{\text{bwd}} = \text{Flip}\left(\text{Conv2D}_{\text{bwd}}(\text{Flip}(X, T)), T\right) \in \mathbb{R}^{B \times \frac{f_1}{2} \times ch \times T}. \tag{2}$$

The outputs are concatenated along the channel dimension:

$$F = \text{Concat}(F_{\text{fwd}}, F_{\text{bwd}}, \text{dim} = 1) \in \mathbb{R}^{B \times f_1 \times ch \times T}. \tag{3}$$

In practice, each path uses $\frac{f_1}{2}$ filters with kernel size 64, followed by batch normalization.

This bidirectional design enables the model to exploit contextual information from both past and future time points while remaining computationally efficient, this is followed by a batch normalization layer, which standardizes the feature distributions.

b) CSG: To incorporate spatial priors derived from electrode placement, we propose a coordinate-aware spatial gating mechanism that modulates electrode activations based on their topographic positions. This module learns a soft attention mask over electrode channels, guided by their 2D spatial coordinates on the scalp.

Let $X \in \mathbb{R}^{B \times C \times ch \times T}$ be the input feature map, where B is the batch size, C is the number of convolutional filters, ch is the number of electrodes, and T is the number of time samples.

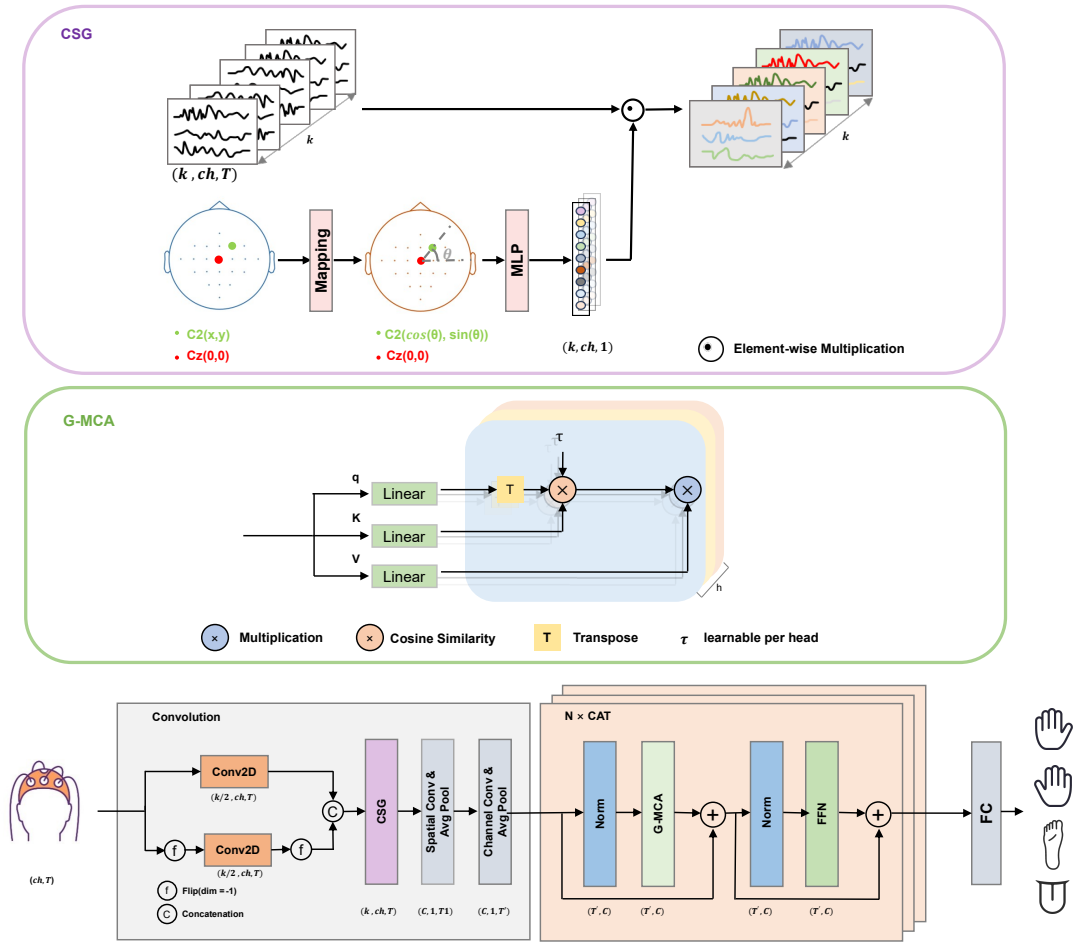

Fig. 1: Overall architecture of AwareNet.

We begin by extracting the 2D Cartesian coordinates $P = [x_i, y_i] \in \mathbb{R}^{ch\times 2}$ of each selected electrode based on the standard montage corresponding to the dataset [18]. To regularize their distribution and reduce angular irregularity, we apply an equiangular mapping. Specifically, we transform each position to its polar angle:

$$\theta_i = \text{atan2}(y_i, x_i) \tag{4}$$

where atan2 returns the arctangent of (y_i, x_i) in the range $(-\pi, \pi)$, providing the correct angular orientation for each point in polar coordinates. We then sort the electrodes by θ_i and remap them to uniformly spaced angular positions:

$$\theta_i^{\text{uniform}} = \text{linspace}(-\pi, \pi,\ ch) \tag{5}$$

where ch denotes the number of selected electrodes. Finally we reproject these angles to 2D unit circle coordinates:

$$P_i = \left[\cos\theta_i^{\text{uniform}},\ \sin\theta_i^{\text{uniform}}\right] \tag{6}$$

where P_i represents the remapped coordinates of the i-th electrode on the unit circle. This ensures that all selected electrodes are evenly distributed in angular space, facilitating regularized spatial gating.

The mapped coordinates P are passed through a lightweight MLP to generate electrode-wise gates, the MLP consists of two fully connected layers: the first layer projects the 2D input to a hidden representation of dimension d using a linear transformation followed by a ReLU activation, the second layer maps the hidden features to C output channels, followed by a sigmoid activation:

$$G = \sigma\left(\text{MLP}(P)\right) \in \mathbb{R}^{ch\times C} \tag{7}$$

where σ denotes the sigmoid function, ch is the number of electrodes and C is the number of feature channels. The resulting gating matrix is then reshaped to match the input dimensions:

$$G_{\text{reshaped}} = G^{\top} \in \mathbb{R}^{1\times C\times ch\times 1} \tag{8}$$

The final spatial gating is applied via element-wise multiplication:

$$X_{\text{gated}} = X \odot G_{\text{reshaped}} \tag{9}$$

This gating mechanism allows the model to emphasize or suppress feature responses based on electrode spatial context, allowing the network to learn electrode-specific importance based on their physical arrangement.

c) Spatial and channel convolutions: After CSG, we apply a depthwise spatial convolution across electrodes with kernel size $(ch, 1)$ to extract spatial patterns, followed by BN and ELU activation. Then, a temporal convolution with kernel size $(1, 16)$ is applied to refine temporal features, also followed by BN and ELU. Two average pooling layers with kernel sizes $(1, 8)$ and $(1, 6)$ are used to reduce temporal redundancy. The resulting feature map is $X \in \mathbb{R}^{B\times C\times 1\times T'}$. We then transpose and reshape it into a token sequence:

$$X \in \mathbb{R}^{B\times T'\times C}, \tag{10}$$

where each time step corresponds to a token for the subsequent CAT module.

2) CAT module: Self-attention is effective for modeling long-range dependencies in EEG sequences [15]. However, traditional dot-product attention can be sensitive to feature magnitude in small-sample settings. We therefore adopt a cosine-similarity multi-head attention with learnable temperature scaling (G-MCA) to improve stability.

Given $X \in \mathbb{R}^{B\times T'\times C}$, queries, keys, and values are obtained by linear projections:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V. \tag{11}$$

We normalize along the feature dimension to compute cosine similarity:

$$\tilde{Q} = \frac{Q}{\|Q\|}, \quad \tilde{K} = \frac{K}{\|K\|}. \tag{12}$$

Attention is computed as:

$$\mathrm{Attention}(\tilde{Q}, \tilde{K}, V) = \mathrm{Softmax}(\tilde{Q}^T\tilde{K}\cdot\tau)V, \tag{13}$$

where $\tau \in \mathbb{R}^{h\times 1\times 1}$ is a learnable temperature parameter per head (initialized to 1), adaptively controls the sharpness of the temporal attention distribution, which helps mitigate EEG-specific amplitude variability and improves training stability. h is the number of heads. The multi-head output is:

$$\mathrm{MCA}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)W^O, \tag{14}$$

$$\mathrm{head}_i = \mathrm{Attention}(\tilde{Q}W_i^Q, \tilde{K}W_i^K, VW_i^V). \tag{15}$$

Each attention block is followed by a feed-forward network:

$$\mathrm{FFN}(x) = \mathrm{GELU}(xW_1 + b_1)W_2 + b_2, \tag{16}$$

with dropout [19], layer normalization and residual connections are used for stable training [20]. The block is repeated N times according to the depth.

3) Classification: We use a fully connected (FC) layer followed by a softmax function to map the learned representations to class probabilities, producing the final prediction over all classes.

III. Experiments

A. Datasets

We evaluate AwareNet on three public MI-EEG benchmarks: BCIC-IV-2a, BCIC-IV-2b, and the High Gamma Dataset (HGD).

1) BCIC-IV-2a: BCIC-IV-2a [21] contains EEG from 9 subjects recorded with 22 channels at 250 Hz for 4-class motor imagery (left hand, right hand, feet, tongue). Each subject has 288 trials (72 per class) collected in two sessions; we follow the standard cross-session protocol (session 1 for training, session 2 for testing).

2) BCIC-IV-2b: BCIC-IV-2b [21] includes 9 subjects performing binary motor imagery (left vs. right hand) recorded from 3 channels (C3, Cz, C4) at 250 Hz. Each subject has five sessions (120 trials/session); The first three sessions are used for training, while the last two are reserved for testing.

3) HGD: HGD [8] consists of 14 subjects recorded with 128 channels at 500 Hz for 4-class motor imagery (left hand, right hand, feet, rest). Following common practice, we select 44 motor-area channels and downsample the data to 250 Hz for consistent hyperparameter settings.

For all datasets, we extract epochs from 0 to 4s after cue onset and apply no additional preprocessing.

B. Training configuration

Our model was trained on an NVIDIA GeForce RTX 4070s GPU using the PyTorch library in Python [22]. The training procedure strictly follows the official competition protocols. We selected 20% of the training set as a validation set and selected the checkpoint with the lowest validation loss.

The model was evaluated using both subject-dependent and cross-subject strategies. In the subject-dependent setting, the model was trained individually for each subject. For cross-subject evaluation, we adopted the leave-one-subject-out (LOSO) approach, where in each iteration, one subject was held out as the test set, and the data from all remaining subjects were used for training. This process was repeated for each subject so that every subject served as the test subject once. Both strategies employed the same training configuration. We used the Adam optimizer [23] with default parameters and a learning rate of 0.001. The loss function was set to cross-entropy. The batch size was fixed at 72, and the model was trained for 550 epochs. The hyperparameters used during training are summarized in Table I.

Convolution		**CAT**	
Forward Conv	10 kernels, 1 × 64	Head	8
Backward Conv	10 kernels, 1 × 64	Depth	4
Spatial Conv	40 kernels, ch × 1	Dropout	0.5
Pool1	8		
Channel Conv	40 kernels, 1 × 16		
Pool2	6		

TABLE I: Model Parameter Settings

We employed classification accuracy and Cohen's kappa coefficient to evaluate the performance of the proposed model.

C. Base models

We compare AwareNet with eight MI-EEG deep models, including CNN baselines: EEGNet [9], ShallowConvNet and DeepConvNet [8]; CNN + Transformer hybrids: Conformer [15] and CTNet [16]; Model with graph convolution: TVGNet

TABLE II: Subject-dependent results on 2a and 2b datasets

	2a-NoAug		2a-Aug		2b-NoAug		2b-Aug	
Model	Acc(%)	κ	Acc(%)	κ	Acc(%)	κ	Acc(%)	κ
ShallowConvNet	67.09±6.83	0.57	72.06±11.41	0.62	84.53±9.23	0.69	85.04±9.64	0.70
DeepConvNet	68.92±12.88	0.58	75.63±12.92	0.68	85.08±8.23	0.70	84.45±9.07	0.69
EEGNet	70.56±9.96	0.61	74.01±11.19	0.66	85.78±8.62	0.71	87.47±8.10	0.75
Conformer	65.20±10.94	0.53	73.60±15.90	0.65	82.73±10.49	0.65	85.53±9.36	0.70
SST-DPN	76.90±10.05	0.69	80.08±9.30	0.74	84.50±9.25	0.68	85.21±10.44	0.70
FA-STTM	**77.54**±8.61	**0.71**	82.46±11.34	0.77	86.39±9.20	0.73	88.26 ±**8.42**	0.77
CTNet	75.30±9.46	0.67	80.12±8.97	0.73	85.65±**8.46**	0.71	86.82±9.30	0.73
TVGNet	76.18±10.23	0.69	81.46±9.47	0.74	86.75±9.28	0.72	87.49±8.96	0.75
AwareNet	77.42±**7.96**	0.70	**82.85±8.62**	**0.77**	**87.06**±9.17	**0.74**	**88.37**±8.56	**0.77**

TABLE III: Subject-dependent results on HGD.

Methods	Acc(%)	Std	κ	p
ShallowConvNet	88.87	8.32	0.84	.001
DeepConvNet	88.98	8.71	0.85	.001
EEGNet	89.32	7.59	0.85	.001
Conformer	88.43	7.86	0.85	.001
SST-DPN	93.16	**3.50**	**0.92**	.480
CTNet	92.20	7.86	0.89	.060
AwareNet	**93.62**	4.87	**0.92**	–

TABLE IV: Subject-independent results on 2a and 2b datasets.

	BCIC-IV-2a		BCIC-IV-2b	
Model	Acc(%)	κ	Acc(%)	κ
ShallowConvNet	59.25±10.75	0.46	76.15±6.59	0.52
DeepConvNet	58.64±12.72	0.45	76.35±6.60	0.53
EEGNet	55.23±8.84	0.40	75.88±5.46	0.51
Conformer	54.74±11.10	0.39	75.43±6.73	0.51
AwareNet	**60.34±11.25**	**0.48**	**77.34±5.03**	**0.54**

[25]; CNN + Mamba hybrid: FA-STTM [26]; and a CNN model with spatial attention: SST-DPN [12]. We reimplemented all baselines using publicly available code and followed the original hyperparameter settings whenever possible. For fairness, all models were trained and evaluated under the same data splits, preprocessing, optimizer, learning rate, batch size, and number of epochs as in our training configuration.

IV. Results

We first evaluate AwareNet against representative baselines under subject-dependent and subject-independent settings. We further report ablation and parameter-sensitivity analyses. Statistical significance is assessed using the paired Wilcoxon signed-rank test.

A. Overall comparison

Given that Transformer-based models benefit more from data augmentation when the sample size is limited, we compared the performance of all models under subject-dependent settings both with and without data augmentation. For the subject-independent setting, data augmentation was not applied, as the amount of training data was already sufficient.

1) Subject-dependent classification:

a) BCIC-IV-2a: As shown in Table II, 2a-Aug, AwareNet achieves the best mean decoding accuracy (82.85%). It also yields the highest Cohen's kappa (0.77) and the lowest standard deviation (8.62). AwareNet significantly outperforms CNN baselines ($p < .01$) and also exceeds competitive methods such as CTNet and SST-DPN ($p < .05$).

As shown in Table II, 2a-NoAug, AwareNet attains the second highest mean accuracy (77.42%) and kappa score (0.70), significantly outperforming most baselines ($ps < .05$). The differences to CTNet ($p = .31$) TVGNet ($p = .45$) and SST-DPN ($p = .72$) are not statistically significant, while AwareNet still achieves higher mean accuracy by 2.01%, 1.24% and 0.41%, respectively.

Overall, AwareNet remains robust under limited training data. Data augmentation yields larger improvements for Transformer/Mamba-based models than for CNN-based baselines.

b) BCIC-IV-2b: Table II, 2b-Aug illustrates the decoding performance of each model on the 2b dataset with data augmentation. AwareNet achieves the best mean accuracy (88.37%), the highest kappa (0.77), and the lowest standard deviation (8.56). It significantly outperforms most baselines.

Even without data augmentation Table II, 2b-NoAug, our model still attained a strong average decoding accuracy of 87.06%, leading the second-best model by 0.67%, and achieved the highest kappa value (0.74).

c) HGD: We further evaluate AwareNet on the High Gamma Dataset (HGD). Given the non-normal distribution of decoding performance across 14 participants, we employed the Friedman test to assess the decoding performance across models. As shown in Table III, the differences among models are significant ($\chi^2(7) = 62.31, p < .001$). AwareNet achieves the best overall performance (93.62% accuracy, $\kappa = 0.92$), and performs competitively against strong baselines such as SST-DPN and CTNet.

2) Subject-independent classification: To assess cross-subject generalization, we conduct leave-one-subject-out evaluation on BCIC-IV-2a and BCIC-IV-2b. As summarized in Table IV, AwareNet achieves the highest mean accuracy on 2a

(60.34%, $\kappa = 0.48$). On 2b (Table IV), although the overall differences are small, AwareNet remains the top-performing method on average (77.34%, $\kappa = 0.54$), indicating stable generalization in the low-density 3-channel setting.

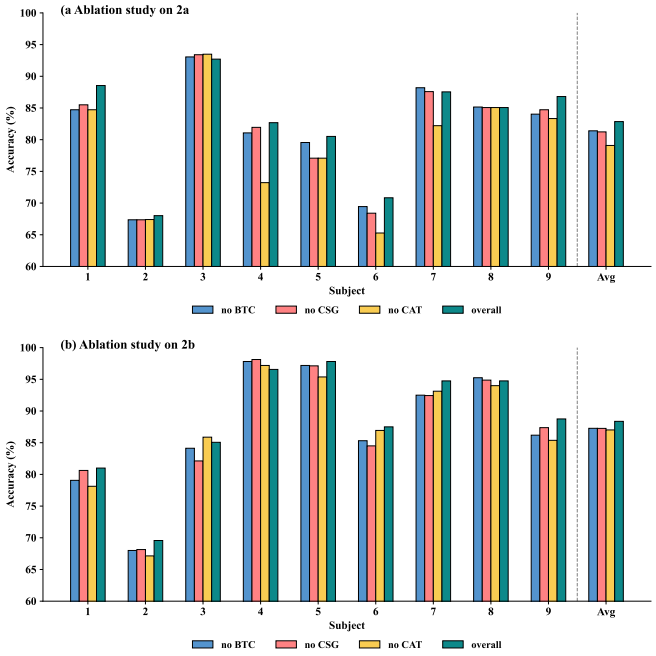

Fig. 2: Ablation results on BTC, CSG, and CAT modules on Dataset 2a (a) and Dataset 2b (b).

B. Ablation study

We performed ablation experiments on BCIC-IV-2a and BCIC-IV-2b to quantify the contribution of the three core components: BTC, CSG, and CAT. Each variant was trained and validated under the same protocol, and the subject-averaged accuracies are reported in Figure 2.

BTC provides consistent but modest gains, improving accuracy by 1.07% on 2a and 1.24% on 2b, indicating the benefit of bidirectional temporal modeling.

Removing CSG leads to clear degradations (1.93% on 2a and 1.48% on 2b), suggesting that incorporating electrode-coordinate priors improves spatial feature learning, even in low-density recordings (2b).

CAT yields the largest performance gain. Excluding CAT reduces accuracy by 3.53% on 2a and 1.67% on 2b, highlighting the importance of global dependency modeling. To further isolate the effect of our attention design, we compare CAT against removing CAT and replacing G-MCA with conventional dot-product MSA. As shown in Table V, the proposed CAT achieves higher kappa and lower variance across subjects, supporting the advantage of cosine attention with learnable temperature scaling.

C. Parameter sensitivity

We analyze the sensitivity of key hyperparameters, including pooling sizes in the convolutional embedding and the number of heads and depth in CAT (see Figure 3).

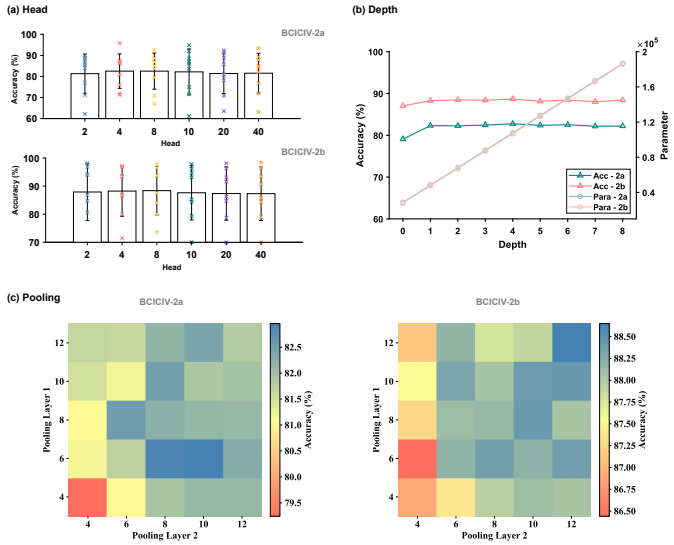

Fig. 3: Parameter sensitivity evaluation on 2a and 2b. (a) The effect of heads in the G-MCA layer. (b) The effect of depth in CAT module on accuracy and parameters. (c) The effect of pooling sizes.

Pooling sizes. Varying pooling sizes from 4 to 12 shows that overly small pooling (i.e., too many tokens) consistently degrades performance, whereas moderate-to-large pooling stabilizes accuracy on both datasets. The best performance is obtained at $(p1, p2) = (8, 6)$ for 2a and $(12, 12)$ for 2b, and performance remains within a narrow range once the overall downsampling is sufficiently large.

Number of heads. Increasing heads improves performance up to a peak around 8 heads, after which accuracy gradually declines. The overall fluctuation is limited (about 1.75% on 2a and 0.98% on 2b), suggesting that excessive head partitioning may introduce redundancy.

CAT depth. Increasing CAT depth from 0 to 1 yields a significant improvement ($p < .05$), while deeper stacks provide marginal gains ($p > .05$). Performance peaks around 4 to 6 layers, where the trade-off between accuracy and model complexity is most favorable.

V. Visualization and Discussion

We provide qualitative evidence of interpretability using t-SNE and CAM-based topographical visualizations (Grad-CAM) [27], [28].

A. Effect on training process

Figure 4 illustrates the validation loss curves of three model variants during training: the proposed model with CAT (red), the version without CAT (yellow), and a variant based on the standard Transformer architecture, where the MCA module is replaced by standard multi-head self-attention (blue). As shown, our model demonstrates the most rapid and stable convergence, consistently achieving the lowest validation loss

TABLE V: Comparison of classification accuracy (%) and kappa on BCIC-IV-2a among AwareNet (with CAT), a variant without CAT, and a standard Transformer variant (replacing CAT with MSA while keeping all other components unchanged).

Methods	A01	A02	A03	A04	A05	A06	A07	A08	A09	Avg.	Std.	Kappa
No CAT	84.33	67.62	**95.76**	72.88	76.38	65.86	80.99	85.98	84.06	79.32	9.57	0.72
With MSA	88.62	67.21	94.26	80.85	75.99	70.58	82.68	**86.32**	85.74	81.39	8.91	0.75
AwareNet	**89.03**	**67.14**	92.76	**82.98**	**79.84**	**71.39**	**89.04**	85.69	**87.78**	**82.85**	**8.62**	**0.77**

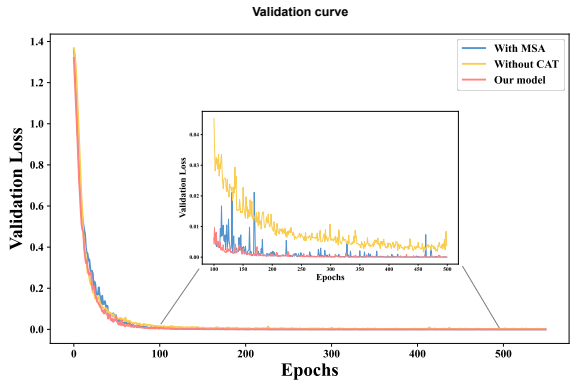

Fig. 4: Training process for different configurations: Our model (With CAT), Without CAT, and Replace CAT with MSA (With MSA).

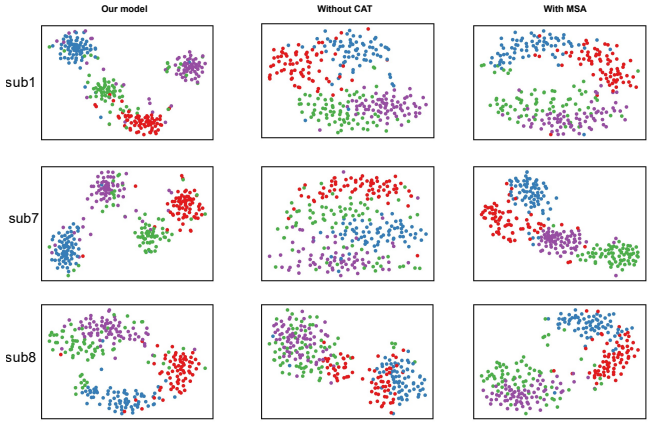

Fig. 5: t-SNE visualization of AwareNet, with comparisons to versions without CAT, and with MSA on BCICIV-2a.

throughout the training process. Model without CAT exhibits high variance and poor convergence, and the Transformer-based variant shows moderate fluctuations. These findings highlight the effectiveness of the CAT module in facilitating efficient optimization and enhancing the model's robustness.

B. Feature distribution

Figure 5 visualizes the learned representations. Features from AwareNet form more compact and better-separated class clusters than the w/o CAT and MSA variants, suggesting that CAT enhances discriminability by modeling long-range temporal dependencies within channel groups.

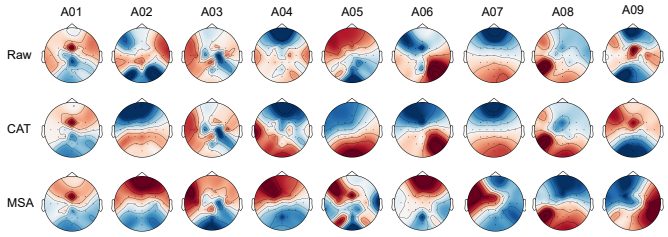

Fig. 6: Grad-CAM topographical visualizations on the BCICIV-2a dataset for each participant under different configurations. Top: patterns in the raw EEG topographies; Middle: activation maps generated by AwareNet with the proposed CAT module; Bottom: activation maps generated by the model using standard multi-head self-attention (MSA).

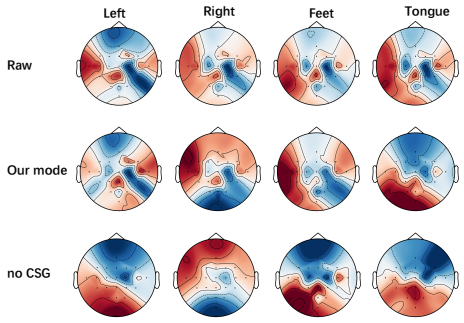

Fig. 7: Grad-CAM topographical visualizations on the BCICIV-2a dataset (A03), comparing raw EEG topographies with CAM-activated maps generated by AwareNet and by AwareNet without the CSG module across different motor imagery tasks.

C. Topographical interpretability

Figure 7 and Figure 6 report CAM-based activation maps for MI classes. With CSG, AwareNet produces more focused and physiologically plausible patterns that better align with the raw EEG topographies (e.g., contralateral sensorimotor activation for hand imagery and midline activation for feet). Removing CSG yields diffuse maps, indicating weakened spatial specificity. Moreover, replacing CAT with standard MSA leads to less consistent subject-wise activations, while CAT produces more stable and task-relevant topographies across subjects.

D. General discussion

EEG decoding remains challenging due to low SNR, limited training data, and pronounced inter-subject variability. In this

context, conventional dot-product multi-head self-attention can be sensitive to amplitude variations, which may amplify noise and destabilize optimization, partially explaining why some CNN–Transformer hybrids do not consistently outperform strong CNN baselines. CAT addresses this issue by using cosine-similarity attention with learnable temperature scaling, which reduces magnitude sensitivity while allowing each head to adapt its attention sharpness to EEG statistics. Across datasets, CAT contributes the largest performance gains and improves training stability, and the learned representations become more discriminative (e.g., better-separated clusters and more task-relevant activations), supporting its effectiveness for long-range temporal modeling in noisy MI-EEG.

Spatial information in EEG is structured by electrode geometry, yet many architectures treat channels as an unordered set. CSG injects coordinate priors via a lightweight gating mechanism, improving spatial specificity and yielding activation patterns that better align with physiologically plausible motor-area topographies, which is consistent with the observed accuracy improvements. In short, our CSG module act as a spatial prior that guides the network's attention toward task-relevant brain regions while suppressing irrelevant or noisy channels, addressing the limitation of conventional approaches that treat electrodes as an unordered set of channels and rely solely on a single convolutional operation for spatial processing.

VI. Conclusion

We proposed AwareNet, an end-to-end spatio-temporal framework for MI-EEG decoding that integrates Bidirectional Temporal Convolution (BTC), Coordinate-aware Spatial Gating (CSG), and a Channel-aware Transformer (CAT) with cosine attention and learnable temperature scaling. AwareNet consistently improves decoding performance across multiple benchmarks and evaluation settings, while producing more discriminative and physiologically plausible representations in visualization analyses. These results suggest that combining coordinate priors with robust long-range temporal modeling is a promising direction for practical EEG-based BCI systems.

References

[1] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "Bci2000: a general-purpose brain-computer interface (bci) system," *IEEE Transactions on biomedical engineering*, vol. 51, no. 6, pp. 1034–1043, 2004.

[2] C. Guger, G. Edlinger, W. Harkam, I. Niedermayer, and G. Pfurtscheller, "How many people are able to operate an eeg-based brain-computer interface (bci)?" *IEEE transactions on neural systems and rehabilitation engineering*, vol. 11, no. 2, pp. 145–147, 2003.

[3] J. Decety, "The neurophysiological basis of motor imagery," *Behavioural brain research*, vol. 77, no. 1-2, pp. 45–52, 1996.

[4] K. K. Ang and C. Guan, "Eeg-based strategies to detect motor imagery for control and rehabilitation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 4, pp. 392–401, 2016.

[5] S. Aggarwal and N. Chugh, "Review of machine learning techniques for eeg based brain computer interface," *Archives of Computational Methods in Engineering*, vol. 29, no. 5, pp. 3001–3020, 2022.

[6] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve bci designs: unified theory and new algorithms," *IEEE Transactions on biomedical Engineering*, vol. 58, no. 2, pp. 355–362, 2010.

[7] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (eeg) classification tasks: a review," *Journal of neural engineering*, vol. 16, no. 3, p. 031001, 2019.

[8] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[9] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces," *Journal of neural engineering*, vol. 15, no. 5, p. 056013, 2018.

[10] M. Teplan *et al.*, "Fundamentals of eeg measurement," *Measurement science review*, vol. 2, no. 2, pp. 1–11, 2002.

[11] X. Tang, J. Zhang, Y. Qi, K. Liu, R. Li, and H. Wang, "A spatial filter temporal graph convolutional network for decoding motor imagery eeg signals," *Expert Systems with Applications*, vol. 238, p. 121915, 2024.

[12] C. Han, C. Liu, Y. Wang, C. Cai, J. Wang, and D. Qian, "A spatial-spectral and temporal dual prototype network for motor imagery brain-computer interface," *arXiv preprint arXiv:2407.03177*, 2024.

[13] M. Wimpff, L. Gizzi, J. Zerfowski, and B. Yang, "Eeg motor imagery decoding: A framework for comparative analysis with channel attention mechanisms," *Journal of neural engineering*, vol. 21, no. 3, p. 036020, 2024.

[14] H. Altaheri, G. Muhammad, and M. Alsulaiman, "Physics-informed attention temporal convolutional network for eeg-based motor imagery classification," *IEEE transactions on industrial informatics*, vol. 19, no. 2, pp. 2249–2258, 2022.

[15] Y. Song, Q. Zheng, B. Liu, and X. Gao, "Eeg conformer: Convolutional transformer for eeg decoding and visualization," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 710–719, 2022.

[16] W. Zhao, X. Jiang, B. Zhang, S. Xiao, and S. Weng, "Ctnet: a convolutional transformer network for eeg-based motor imagery classification," *Scientific Reports*, vol. 14, no. 1, p. 20237, 2024.

[17] R. D. Seidler, D. C. Noll, and G. Thiers, "Feedforward and feedback processes in motor control," *Neuroimage*, vol. 22, no. 4, pp. 1775–1783, 2004.

[18] J. N. Acharya and V. J. Acharya, "Overview of eeg montages and principles of localization," *Journal of Clinical Neurophysiology*, vol. 36, no. 5, pp. 325–329, 2019.

[19] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[20] J. L. Ba, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[21] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. R. Müller-Putz *et al.*, "Review of the bci competition iv," *Frontiers in neuroscience*, vol. 6, p. 55, 2012.

[22] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] F. Lotte, "Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain–computer interfaces," *Proceedings of the IEEE*, vol. 103, no. 6, pp. 871–890, 2015.

[25] Z. Yao, X. Xing, and M. Gao, "Tvgnet: A novel hybrid eeg motor imagery decoding network," *Biomedical Signal Processing and Control*, vol. 116, p. 109601, 2026.

[26] M. Gao, C. Gao, Z. Yao, and W. Zhu, "Fa-sttm: A hybrid transformer-mamba network for motor imagery eeg classification," *Neurocomputing*, p. 132481, 2025.

[27] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.