Grounding Global Attention: A Prior-Guided Fusion Framework for Cross-Subject EEG Decoding

Xiao CHEN[1,2], Xin Zhang[2], Shiwei Guo[3], Guodong Li[1*], and Jixuan Kang[2] ,
[1]Central China Normal University, Wuhan, China
[2]Harbin Institute of Technology, Harbin, China
[3]University of Chinese Academy of Sciences, Beijing, China
Email: chenxiao21@mails.ucas.ac.cn, xin928560@gmail.com, guoshiwei18@mails.ucas.ac.cn,
liguodong@ccnu.edu.cn*, 18045154515@163.com

***Abstract*—Electroencephalography (EEG) offers millisecond-scale temporal resolution for probing brain dynamics. However, accurate decoding is hindered by pronounced cross-subject variability and the inherent non-stationarity of neural oscillations. Existing hybrid methods often treat local feature extraction and global modeling as decoupled stages, failing to ground the attention mechanism in biologically meaningful local structures. To address this, we propose DMTA-PGT, a unified framework that tightly couples adaptive feature weighting with prior-guided context modeling. Specifically, the Dynamic Multi-Scale Temporal Attention (DMTA) module adaptively fuses multi-resolution temporal features to mitigate non-stationarity. Crucially, the Prior-Guided Transformer (PGT) injects convolution-derived inductive biases directly into the self-attention mechanism via a gating function, ensuring global dependencies are anchored to local temporal continuity. Extensive cross-subject evaluations on two benchmarks demonstrate that DMTA-PGT significantly outperforms state-of-the-art CNN and Transformer baselines , offering a superior trade-off between decoding accuracy and stability.**

***Index Terms*—brain-computer interface, deep learning, electroencephalography, EEG decoding**

I. Introduction

Electroencephalography (EEG) has become a central modality for brain–computer interfaces (BCIs) due to its millisecond-level temporal resolution, portability, and suitability for real-world cognitive monitoring [1]. Despite substantial progress, achieving robust cross-subject EEG decoding remains a major challenge. Neural responses vary widely across individuals and fluctuate over time within the same recording session, causing severe non-stationarity in temporal patterns [2]. These characteristics reduce the generalization ability of many existing decoding models.

Deep neural architectures have been widely explored to mitigate these issues. Convolutional networks such as EEGNet [3] and DeepConvNet [4] extract discriminative spatial–temporal features but typically rely on a single temporal kernel, making them sensitive to temporal-scale drift across subjects. Multi-branch [5], [6] designs alleviate this limitation by incorporating multiple temporal kernels or window sizes, yet their fusion strategies often use static or hand-crafted weights that ignore trial-specific variability. Graph-based EEG models [7], [8] encode spatial topology but offer limited flexibility in modeling temporal dynamics. Recent Transformer-based approaches [9]–[11] introduce global receptive fields but lack inherent inductive biases to respect local temporal continuity, leading to unstable training and susceptibility to spurious correlations in noisy EEG recordings.

A key limitation shared across hybrid CNN–Transformer solutions [11]–[13] is the structural decoupling between local feature extraction and global contextual modeling. Convolutions capture short-range temporal patterns that are biologically grounded, whereas self-attention excels at modeling long-range dependencies. However, when these processes are treated as sequential and independent, the Transformer lacks guidance from the local neural structure, and the network fails to fully exploit the complementary strengths of both components.

To bridge this gap, we propose DMTA–PGT, a unified local-to-global framework for robust EEG decoding. First, the Dynamic Multi-Scale Temporal Attention (DMTA) module captures temporal features at multiple receptive fields and adaptively fuses them on a per-trial basis, thereby mitigating the impact of non-stationary temporal variations. Second, the Prior-Guided Transformer (PGT) integrates a lightweight, data-driven temporal prior—derived from convolutional representations—into the self-attention mechanism by modulating the Query and Key projections through a gating function. This operation embeds locally coherent temporal structure directly within the attention computation, helping the model maintain more stable long-range interactions under noisy EEG conditions. While several recent EEG models combine CNNs with Transformers, these hybrid designs typically treat CNNs purely as external feature extractors and leave the attention mechanism unaltered, offering limited ability to regulate attention over unstable EEG tokens. In contrast, PGT injects an explicit inductive bias inside the attention layer, establishing a convolution-informed local structural constraint that more tightly couples local representations with global reasoning. This provides a more tightly coupled form of local-to-global modeling compared with prior hybrid approaches.

Our contributions are as follows:

Xiao CHEN is an intern at Central China Normal University.
*Corresponding author.

- We present a principled architecture that tightly couples adaptive multi-scale temporal analysis with prior-guided global reasoning.
- We introduce DMTA and PGT, two complementary modules that address temporal non-stationarity and structural grounding.
- Extensive evaluations show that our method outperforms a range of baselines in cross-subject generalization and stability on two EEG benchmarks.

II. Related Work

Early deep learning approaches for EEG decoding primarily relied on Convolutional Neural Networks (CNNs). While models like EEGNet [3] successfully introduced channel-wise and depthwise convolutions, their reliance on single-scale temporal kernels made them inherently sensitive to the cross-subject frequency drifts and temporal non-stationarity prevalent in EEG data. Attempts to increase robustness, such as DeepConvNet [4] and multi-branch designs like TSception [14] and MSCFormer [15], introduced multiple parallel kernels. However, these methods typically suffer from a key limitation: they employ static, trial-shared weighting or simple concatenation for feature fusion. This static approach prevents the model from dynamically adapting the emphasis on specific temporal scales to align with the heterogeneous and non-stationary characteristics of individual subjects' neural patterns, hindering optimal cross-subject generalization.

The emergence of Transformer models provided a mechanism for capturing long-range global dependencies. Pure Transformer methods, like EEGVIT [17], treat the EEG signal as disconnected patches, inherently neglecting the crucial local temporal continuity and inductive bias essential for robust time-series processing. To leverage both local and global features, recent research introduced CNN-Transformer hybrid models [11]–[13], such as EEG-Conformer [9]. These models typically function by treating the CNN as an external feature extractor whose output is flattened and then fed into a standard, unmodified self-attention layer. This common design results in a structural decoupling: the self-attention operation remains unconstrained by the local neural structure, making the global modeling process susceptible to overfitting on noise and spurious long-range correlations, particularly in low Signal-to-Noise Ratio EEG datasets.

Our proposed DMTA-PGT framework directly addresses the limitations outlined above through two synergistic innovations. First, the Dynamic Multi-Scale Temporal Attention (DMTA) module resolves the issue of static weighting by implementing an adaptive fusion mechanism that dynamically emphasizes the most informative temporal scales on a per-sample basis, significantly improving resilience to non-stationarity. Second, and most critically, our Prior-Guided Transformer (PGT) moves beyond simple component stacking. The PGT injects the convolutional-derived local temporal prior directly into the self-attention mechanism. By using this prior to modulate the Query ($\tilde{Q}$) and Key ($\tilde{K}$) vectors via a gating function, we establish an explicit, structure-aware constraint inside the attention layer. This tight integration achieves a more robust and interpretable form of local-to-global modeling compared to existing decoupled hybrid approaches.

III. Method

We propose DMTA–PGT, an end-to-end hybrid architecture designed to address the temporal non-stationarity and cross-subject variability inherent in EEG signals. The framework consists of two complementary components: (1) a Dynamic Multi-Scale Temporal Attention (DMTA) module that adaptively extracts temporal patterns across multiple receptive fields; and (2) a Prior-Guided Transformer (PGT) that incorporates convolution-derived temporal priors to modulate global attention. An overview of the proposed DMTA-PGT architecture is illustrated in Fig. 1.

A. Dynamic Multi-Scale Temporal Attention

1) Motivation: EEG signals embed information across multiple temporal resolutions due to the coexistence of fast transient events and slower oscillatory dynamics. Models that rely on a single temporal kernel size are inherently limited [18], as they cannot adapt to subject-specific temporal-scale drift or non-stationarity across sessions [19], [20]. Multi-branch architectures partially mitigate this issue but commonly use static or manually designed fusion rules that cannot adjust to trial-level variability. The DMTA module addresses this by performing adaptive multi-scale fusion, enabling sample-specific weighting of temporal features for robust EEG decoding.

2) Multi-Scale Feature Extraction and Pooling: Given the input EEG epoch $X \in \mathbb{R}^{B\times C\times T}$, where B denotes the batch size, C the number of channels, and T the temporal steps. We employ three parallel temporal convolutional branches with kernel sizes $\{k_1 = 7,\ k_2 = 11,\ k_3 = 15\}$ to capture features at different receptive fields:

$$F_m = Conv2D_{k_m}(BN(ELU(X))) \tag{1}$$

The outputs of the three branches are concatenated along the channel dimension to form a unified multi-scale representation:

$$F_{ms} = Concat(F_1, F_2, F_3) \tag{2}$$

Next, to transition from temporal to spatial modeling, we apply a spatial convolution across the EEG channels to model inter-channel dependencies, producing spatially enriched feature maps F_s:

$$F_s = Conv2D_{k_s}(F_{ms}) \tag{3}$$

To further capture temporal dynamics at different temporal resolutions and prepare tokens for the downstream Transformer, we apply adaptive temporal pooling with ratios $r_m \in \{0.25, 0.125, 0.0625\}$, producing multi-resolution temporal sequences:

$$\tilde{F}_m = AdaptiveAvgPool(F_s, r_m) \tag{4}$$

The multi-resolution pooling generates a dense sampling of temporal features across multiple receptive fields, enabling the

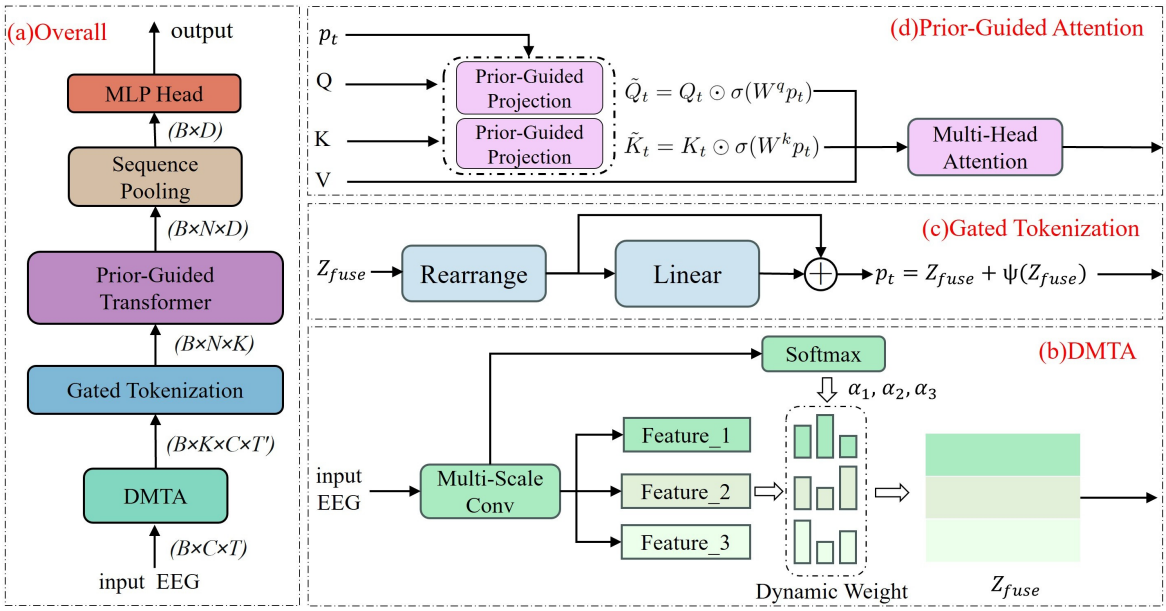

Fig. 1. Overall architecture of DMTA–PGT. An input EEG window is processed by the Dynamic Multi-Scale Temporal Attention (DMTA) module for spatial integration and adaptive temporal-scale fusion. The fused representation is tokenized and fed into the Prior-Guided Transformer (PGT), where local temporal priors modulate Query/Key interactions via gated attention. Global features are then aggregated by average pooling and classified. The framework follows a local-to-global modeling strategy for non-stationary EEG signals.

model to capture both short-lived and slowly evolving neural dynamics. This dense temporal representation reduces inter-subject scale mismatch and stabilizes downstream Transformer modeling.

3) Adaptive Learnable Fusion: Since different tasks, subjects, and trials rely on different temporal scales, DMTA introduces learnable scale attention weights α_m, normalized via softmax:

$$\alpha_m = \frac{\exp(\alpha_m)}{\sum_{i=1}^{3} \exp(\alpha_i)} \tag{5}$$

The final fused representation is obtained as a weighted sum across scales:

$$z_{fuse} = \sum_{m=1}^{3} \alpha_m \tilde{F}_m \tag{6}$$

This adaptive fusion mechanism allows the model to emphasize the most informative temporal resolutions for each sample, enhancing robustness to temporal non-stationarity and cross-subject variability.

B. Prior-Guided Transformer

1) Motivation: Transformers have shown strong capacity for modeling long-range temporal dependencies in EEG signals [21], but their self-attention mechanism lacks inherent inductive bias toward local temporal continuity. As a result, vanilla attention [22] often overfits noise, attends to spurious correlations, and struggles with cross-subject generalization. To mitigate these issues, we introduce the Prior-Guided Transformer (PGT), which injects a data-driven local temporal prior into the attention mechanism. This prior modulates the Query and Key projections, ensuring that global attention remains grounded in locally coherent neural structure.

2) Gated Tokenization: The fused output of DMTA, $Z_{fuse} \in \mathbb{R}^{B\times T'\times K}$, is first reshaped (tokenized) into a sequence of tokens $\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_{T'}\}$ where $Z_t \in \mathbb{R}^K$ represents the rich feature vector for a specific time step t. For each time token Z_t, a local temporal prior p_t is computed using a residual connection:

$$p_t = Z_t + \Psi(Z_t) \tag{7}$$

where $\Psi(\cdot)$ is implemented as a lightweight Linear Projection followed by a non-linearity. This function serves to enhance and enrich the short-range temporal structure initially captured by the DMTA's convolutional front-end. The resulting prior p_t is, therefore, an explicit encoding of local temporal continuity and short-range dynamics, which CNNs reliably capture. By projecting this convolution-derived structural information into dynamic gating coefficients, PGT overcomes the structural decoupling of standard hybrid models. This mechanism ensures that the global token interactions are dynamically anchored to stable local EEG structure, drastically improving the robustness of the self-attention mechanism.

3) Prior-Guided Transformer: The core of our Prior-Guided Transformer (PGT) adapts the standard multi-head self-attention mechanism by incorporating the local prior p_t. This is the pivotal step where we inject the convolution-derived inductive bias directly into the attention mechanism. To bias the attention towards biologically meaningful local structures and stabilize long-range interactions, we modulate the Query (Q) and Key (K) vectors using a data-driven gating mechanism:

$$\tilde{Q}_t = Q_t \odot \sigma(W^q p_t), \quad \tilde{K}_t = K_t \odot \sigma(W^k p_t) \tag{8}$$

Here, σ denotes the sigmoid function, ensuring the gating coefficient $\sigma(W^q p_t)$ is bounded between 0 and 1, effectively acting as a soft mask. The learnable matrices W^q and W^k map the prior p_t into channel-wise gating coefficients. This element-wise multiplication ensures that each token's contribution to the global attention is dynamically constrained and stabilized by its own local temporal structure, thereby mitigating the influence of potentially noisy or spurious long-range dependencies common in EEG data.

The final attention is computed using the modulated vectors:

$$\mathrm{Attn}(Z) = \mathrm{softmax}\left(\frac{\tilde{Q}\tilde{K}^\top}{\sqrt{d_h}}\right)V \tag{9}$$

Multiple PGT layers are stacked to progressively refine global context, and the final representation is aggregated via global average pooling for classification.

IV. Datasets and Experiment Setting

A. Datasets

We evaluated our model on two public EEG benchmark datasets: Dataset I [23] for driver fatigue detection and Dataset II [24] for cognitive workload assessment during mental arithmetic tasks.

1) Dataset I: This dataset contains EEG recordings from a 90-minute VR-based sustained-attention driving simulation frequently used in fatigue research. EEG was acquired using a 32-channel 10–20 system at 500 Hz and band-pass filtered between 1–50 Hz, followed by ocular artifact removal (ICA-based inspection), Automatic Artifact Removal (AAR), and downsampling to 128 Hz. Each sample corresponds to a 3-second window immediately preceding a lane-departure event. Fatigue labels were derived from reaction-time (RT) statistics: trials with RTs exceeding both a global baseline RT and a subject-specific moving average were marked as fatigue (0), while the remaining trials were marked as non-fatigue (1). Following the protocol of prior work [7], 11 subjects with sufficient samples in both classes were included. To ensure reproducibility, we adopt the original segmentation rules and retain the class distribution without resampling or aggregation.

2) Dataset II: This dataset comprises 19-channel EEG recordings acquired from 36 subjects, following the 10–20 montage. Subjects performed a serial subtraction (mental arithmetic) task, with additional resting-state EEG used as baseline. Signals were filtered using a 1–30 Hz band-pass and a 50 Hz notch, followed by ICA to remove ocular and muscle artifacts. Continuous EEG was segmented into 4-second windows with a 2-second step, yielding overlapping trials representing different cognitive states. All subjects are retained for cross-subject evaluation.

B. Baselines

To comprehensively evaluate the performance of our proposed model, we compared it with the following representative baseline methods.

1) DeepConvNet: DeepConvNet [4] is a foundational convolutional neural network that employs broad temporal and spatial filtering to extract discriminative features from raw EEG signals.

2) LGGNet: LGGNet [7] leverages graph neural networks to explicitly model both the local and global functional connectivity among EEG channels, thereby enhancing the decoding of brain activity patterns.

3) EEGNet: EEGNet [3] provides a compact and efficient convolutional architecture that utilizes depthwise and separable convolutions for robust spatial-temporal feature learning.

4) TSception: TSception [14] captures multi-scale temporal and spatial dynamics from EEG signals through its unique multi-branch convolutional structure.

5) EEGViT: EEGViT [17] adapts the Vision Transformer to EEG by treating electrode signals as sequential patches[26], enabling global context modeling via self-attention.

6) EEG-Conformer: EEG-Conformer [9] is a hybrid model that integrates convolutional operations for local feature extraction with a self-attention mechanism to model global dependencies.

7) xEEGNet: xEEGNet [25] extends the classic EEGNet framework by incorporating intrinsic interpretability mechanisms.

C. Experiment Setting

To evaluate model generalization across individuals, we employed a strict cross-subject paradigm. Specifically, Leave-One-Subject-Out (LOSO) cross-validation was applied to all datasets. In each fold, the data from one subject was held out for testing. The data from the remaining subjects were then split into training (80%) and validation (20%) sets. We report the results using Accuracy and the F1-macro as the evaluation metrics, defined as:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

$$F1\text{-macro} = \frac{1}{N}\sum_{i=1}^{N}\frac{2 \times \mathrm{TP}_i}{2 \times \mathrm{TP}_i + \mathrm{FP}_i + \mathrm{FN}_i} \tag{11}$$

where the subscript i denotes the class, and TP, TN, FP, and FN represent the counts of samples that are true positives, true negatives, false positives, and false negatives with respect to that class.

D. Details

To ensure a fair and reproducible comparison, all baseline models were implemented using their official open-source repositories or re-implemented strictly following the original papers. For each method, we adopted a uniform training pipeline. We use the Adam optimizer with an initial learning rate of 1e-3 and a weight decay of 1e-5. The learning rate is scheduled using cosine annealing. A dropout rate of 0.3 is applied to prevent overfitting. The batch size is set to 64 for all experiments. Models are trained for 100 epochs, and the checkpoint achieving the highest validation accuracy was

TABLE I
CROSS-SUBJECT PERFORMANCE COMPARISON ON TWO BENCHMARKING DATASETS.PAIRED T-TESTS ACROSS LOSO FOLDS CONFIRM THAT DMTA–PGT OUTPERFORMS BASELINE MODELS.

Method	Dataset I: Driving Fatigue		Dataset II: Cognitive Workload	
	Accuracy	F1-macro	Accuracy	F1-macro
DeepConvNet	70.23%±15.27%	67.90%±14.69%	64.99%±11.12%	62.54%±12.83%
LGGNet	67.01%±14.95%	64.01%±14.20%	65.80%±11.74%	62.35%±15.11%
EEGNet	70.05%±12.46%	66.84%±13.72%	66.23%±14.21%	62.39%±17.89%
TSception	70.40%±17.06%	67.91%±17.76%	67.52%±16.99%	62.79%±21.45%
EEGViT	67.19%±12.84%	64.56%±12.53%	63.93%±14.52%	61.24%±16.79%
EEG-Conformer	74.04%±12.20%	70.47%±13.61%	68.53%±13.70%	64.92%±17.53%
xEEGNet	71.84%±12.15%	69.41%±12.49%	69.44%±15.59%	66.94%±18.08%
DMTA–PGT	**76.55%**±8.77%	**74.33%**±9.42%	**70.64%**±12.10%	**68.01%**±15.05%

selected for testing. The number of heads in the multi-head attention mechanism is set to 16. Model stability is assessed by conducting multiple runs with different random seeds and reporting the average performance and standard deviation. All models were trained using the same Leave-One-Subject-Out (LOSO) folds, same preprocessed data, and same hardware environment (NVIDIA L20), eliminating discrepancies from data splits or training schedules. This controlled evaluation allows us to isolate the effect of architectural differences and ensures that performance differences arise solely from model design rather than search-space advantages or hyperparameter tuning disparities.

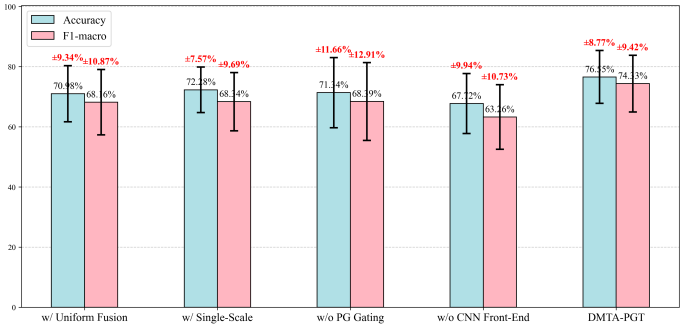

Fig. 2. Ablation study results on Dataset I.

V. RESULTS AND ABLATION STUDY

A. Classification Results

As summarized in Table I, the proposed DMTA–PGT framework achieves the highest overall performance in cross-subject evaluations on both the Driving Fatigue (Dataset I) and Cognitive Workload (Dataset II) datasets, outperforming all convolutional and Transformer-based baselines.

On Dataset I, DMAT–PGT achieves an accuracy of 76.55% (±8.77%) and an F1-score of 74.33% (±9.42%). This not only surpasses classical convolutional models like EEGNet and TSception by a significant margin but also outperforms the strongest Transformer baseline, EEG-Conformer, by +2.51% in accuracy, with concurrently lower prediction variance.

This performance advantage consistently generalizes to Dataset II, where our model yields 70.64% (±12.10%) accuracy and 68.01% (±15.05%) F1-score, achieving better performance than recent approaches, such as EEG-Conformer and xEEGNet. The consistently lower standard deviations across both datasets provide empirical evidence for the model's enhanced robustness to cross-subject variability, a key challenge in EEG decoding.

Collectively, these results validate the efficacy of the proposed local-to-global architecture. The DMTA module contributes to this strong performance by capturing discriminative multi-scale temporal features, while the PGT component enhances generalization by structuring attention around local, biologically plausible priors. Their synergistic integration leads to cross-subject EEG decoding performance.

B. Ablation Study

To systematically evaluate the functional contributions of the core components in DMTA–PGT, including multi-scale temporal modeling, adaptive fusion, and prior-guided global attention, we conducted a structured ablation study. The evaluation adhered to a consistent LOSO cross-subject protocol. Furthermore, each model variant was run with multiple random seeds to mitigate the influence of randomness. This rigorous setup ensures that any observed performance difference is attributable solely to the model architecture, rather than data partitioning or optimization stochasticity.

1) Ablation 1: Uniform Fusion Across Temporal Scales: To assess the benefit of adaptive temporal weighting, we replace the learnable fusion weights α_m with uniform averaging over the pooled multi-scale features. Uniform fusion yields inferior performance, particularly on Dataset I, which contains highly heterogeneous fatigue patterns. The adaptive fusion mechanism enables trial-dependent re-weighting of temporal resolutions, allowing the model to emphasize short- or long-range features according to each subject's neural response. This demonstrates that DMTA's dynamic weighting is critical for handling non-stationary temporal signatures in EEG.

2) Ablation 2: Single-scale Temporal Modeling: This ablation evaluates whether the multi-branch design in DMTA is essential. We replace the three temporal convolutional branches with a single convolution using the largest kernel (k=15), while keeping all subsequent components unchanged.

The consistent performance decline of the single-scale model highlights the limited effectiveness of a fixed receptive field. This approach is inadequate for capturing the rich and varied nature of EEG temporal dynamics, which blend rapid transients with slow oscillations. In contrast, multi-scale modeling successfully integrates complementary temporal information, proving crucial for handling cross-subject variability.

3) Ablation 3: Effect of Priors in Transformer: This ablation removes the prior-guided gating in the PGT module by replacing the gated projections with standard attention projections. Removing the gating mechanism leads to substantial degradation, accompanied by higher variance across folds. This indicates that the local temporal prior stabilizes self-attention by suppressing spurious long-range correlations and grounding global reasoning in short-range neural continuity. Without the prior, the Transformer becomes more sensitive to noise and cross-subject differences, aligning with known limitations of pure attention models on low-SNR EEG data.

TABLE II
ABLATION RESULT ON DATASET I.

Method	Dataset I: Driving Fatigue	
	Accuracy	F1-macro
w/ Uniform Fusion	70.98% ± 9.34%	68.16% ± 10.87%
w/ Single-Scale	72.28% ± 7.57%	68.34% ± 9.69%
w/o PG Gating	71.34% ± 11.66%	68.39% ± 12.91%
w/o CNN Front-End	67.72% ± 9.94%	63.26% ± 10.73%
DMTA–PGT	**76.55%**±8.77%	**74.33%**±9.42%

4) Removing the Convolutional Front-End: To examine whether the convolutional front-end is necessary, we construct a Transformer-only variant that directly tokenizes raw EEG windows without any convolutional modeling. This variant performs the worst among all ablations and exhibits unstable training. The result highlights the importance of establishing local structural representations before global attention. Convolutional processing provides inductive biases—such as local temporal continuity and spatial smoothness—that are indispensable for robust global dependency modeling in noisy EEG settings.

The results of the ablation study on Dataset I, detailed in Table II and Fig. 2, provides evidence for the synergistic design of DMTA-PGT. Across all comparisons, we observed consistent performance degradation whenever any key component was removed or simplified, underscoring the complementary nature of local inductive priors and global contextual modeling in our architecture. Performance drops in Ablation 1 due to the loss of adaptive temporal selectivity. It declines in Ablation 2 due to an inability to capture the full range of temporal dynamics. A significant drop occurs in Ablation 3 due to the model's blindness to global cognitive contexts. Performance collapses in Ablation 4 due to the Transformer's inherent lack of inductive bias for local signal structure, leading to overfitting on noise. Thus, these results collectively verify that adaptive multi-scale fusion, convolutional priors, and long-range modeling are not merely beneficial but important components of DMTA-PGT.

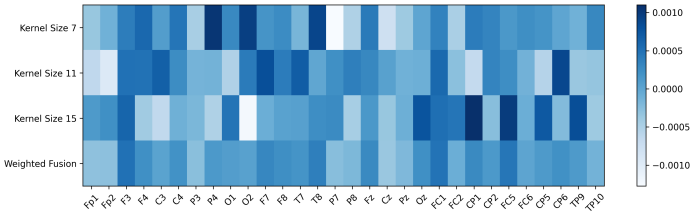

Fig. 3. Heatmap of spatial convolution weights for EEG channels. The color intensity indicates the relative weight of each kernel for each channel, with darker blue representing higher weight.

C. Computational Efficiency

We evaluate model efficiency through parameter quantification on Dataset I. Our DMTA-PGT maintains a compact footprint of approximately 283K parameters. While traditional CNNs like EEGNet are ultra-lightweight, they lack the capacity for global dependency modeling. Conversely, state-of-the-art hybrid models such as EEG-Conformer often exceed 3 million parameters. Our approach achieves significantly higher accuracy than both lightweight baselines and larger models. This demonstrates that DMTA-PGT's performance gains stem from the architectural synergy between multi-scale prior guidance and attention mechanisms, rather than a mere increase in model capacity, striking an optimal balance between complexity and decoding robustness.

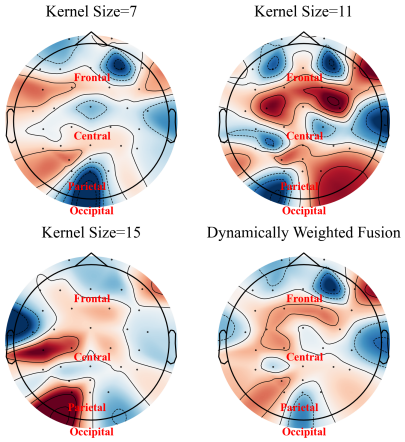

Fig. 4. Learned patterns for driver fatigue classification on Dataset I. Channels around frontal and central regions contribute more strongly during fatigue episodes, consistent with known neurophysiological findings.

D. Visualization

To elucidate the neurophysiological basis of our model's decisions, we visualized the learned weights of the spatial convolution layer on Dataset I. As presented in Fig.3 and Fig.4, the resulting topographic maps revealed distinct spatial attention patterns associated with different temporal contexts. Specifically, features extracted by the short-term temporal

kernel (k=7) were assigned higher weights in the frontal and occipital regions. The medium-term kernel (k=11) engaged a broader network, including central-parietal regions, while the long-term kernel (k=15) exhibited a more refined focus on sustained patterns in the occipital and frontal areas. The aggregate spatial sensitivity, averaged across all temporal scales, consistently highlighted the frontal and occipital cortices as the most critical regions for fatigue classification. Furthermore, these distinct topographic patterns for different temporal kernels validate the efficacy of the multi-scale architecture. This demonstrates that the model successfully captures the multi-faceted nature of the EEG signal, mirroring the complex spatiotemporal dynamics inherent in neural recordings.

VI. Discussion and Future Work

The DMTA–PGT framework demonstrates that coupling adaptive local feature extraction with globally coherent reasoning is essential for robust cross-subject EEG decoding. By grounding the self-attention mechanism in convolutional priors, our model effectively bridges the gap between high-level temporal dynamics and local oscillatory patterns. This synergy suggests that incorporating domain-specific inductive biases can significantly mitigate the challenges posed by EEG non-stationarity.

Future work will focus on two directions. First, we aim to integrate individualized structural priors, such as source-space brain connectivity, to further personalize the attention mechanism. Second, the framework's scalability to multimodal neural signals (e.g., fNIRS-EEG fusion) or alignment with natural language representations will be explored to build more generalized brain-computer interfaces.

References

[1] H. Li, X. Li, and J. R. del Millán, "Noninvasive EEG-based intelligent mobile robots: A systematic review," IEEE Trans. Autom. Sci. Eng., vol. 22, pp. 6291–6315, 2024.

[2] P. Guetschel, S. Ahmadi, and M. Tangermann, "Review of deep representation learning techniques for brain–computer interfaces," J. Neural Eng., vol. 21, no. 6, p. 061002, 2024.

[3] V. J. Lawhern, A. J. Solon, N. R. Waytowich, et al., "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," J. Neural Eng., vol. 15, no. 5, p. 056013, 2018.

[4] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," Hum. Brain Mapp., vol. 38, no. 11, pp. 5391–5420, 2017.

[5] S. Liu, Z. Wang, Y. An, et al., "DA-CapsNet: A multi-branch capsule network based on adversarial domain adaption for cross-subject EEG emotion recognition," Knowl.-Based Syst., vol. 283, p. 111137, 2024.

[6] R. Peng, Z. Du, C. Zhao, et al., "Multi-branch mutual-distillation transformer for EEG-based seizure subtype classification," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 32, pp. 831–839, 2024.

[7] Y. Ding, N. Robinson, C. Tong, et al., "LGGNet: Learning from local-global-graph representations for brain–computer interface," IEEE Trans. Neural Netw. Learn. Syst., vol. 35, no. 7, pp. 9773–9786, 2023.

[8] H. Gao, X. Wang, Z. Chen, et al., "Graph convolutional network with connectivity uncertainty for EEG-based emotion recognition," IEEE J. Biomed. Health Inform., vol. 28, no. 10, pp. 5917–5928, 2024.

[9] Y. Song, Q. Zheng, B. Liu, et al., "EEG conformer: Convolutional transformer for EEG decoding and visualization," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 31, pp. 710–719, 2022.

[10] Z. Wan, M. Li, S. Liu, et al., "EEGformer: A transformer–based brain activity classification method using EEG signal," Front. Neurosci., vol. 17, p. 1148855, 2023.

[11] L. Gong, M. Li, T. Zhang, et al., "EEG emotion recognition using attention-based convolutional transformer neural network," Biomed. Signal Process. Control, vol. 84, p. 104835, 2023.

[12] C. Li, X. Huang, R. Song, et al., "EEG-based seizure prediction via Transformer guided CNN," Measurement, vol. 203, p. 111948, 2022.

[13] J. Yin, A. Liu, C. Li, et al., "A GAN guided parallel CNN and transformer network for EEG denoising," IEEE J. Biomed. Health Inform., 2023, in press.

[14] Y. Ding, N. Robinson, S. Zhang, et al., "TSception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition," IEEE Trans. Affective Comput., vol. 14, no. 3, pp. 2238–2250, 2022.

[15] S. Wang, H. Wang, R. Liu, et al., "MSCFormer: Multi-Scale Circular Transformer for Image Deblurring," in 2024 IEEE Int. Conf. Visual Commun. Image Process. (VCIP), 2024, pp. 1–5.

[16] A. Demir, T. Koike-Akino, Y. Wang, et al., "EEG-GNN: Graph neural networks for classification of electroencephalogram (EEG) signals," in 2021 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), 2021, pp. 1061–1067.

[17] R. Yang and E. Modesitt, "ViT2EEG: leveraging hybrid pretrained vision transformers for EEG data," arXiv preprint arXiv:2308.00454, 2023.

[18] X. Ding, X. Zhang, J. Han, et al., "Scaling up your kernels to 31x31: Revisiting large kernel design in CNNs," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 11963–11975.

[19] J. Luo, W. Cui, S. Xu, et al., "A cross-scale transformer and triple-view attention based domain-rectified transfer learning for EEG classification in RSVP tasks," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 32, pp. 672–683, 2024.

[20] H. Zhang and H. Li, "Transformer-based EEG Decoding: A Survey," arXiv preprint arXiv:2507.02320, 2025.

[21] Y. Ma, Y. Song, and F. Gao, "A novel hybrid CNN-transformer model for EEG motor imagery classification," in 2022 Int. Joint Conf. Neural Netw. (IJCNN), 2022, pp. 1–8.

[22] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," Adv. Neural Inf. Process. Syst., vol. 30, 2017.

[23] Z. Cao, C. H. Chuang, J. K. King, et al., "Multi-channel EEG recordings during a sustained-attention driving task," Sci. Data, vol. 6, no. 1, p. 19, 2019.

[24] I. Zyma, S. Tukaev, I. Seleznov, et al., "Electroencephalograms during mental arithmetic task performance," Data, vol. 4, no. 1, p. 14, 2019.

[25] A. Zanola, L. F. Tshimanga, F. Del Pup, et al., "xEEGNet: Towards Explainable AI in EEG Dementia Classification," arXiv preprint arXiv:2504.21457, 2025.

[26] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.