MoEEG: A Sparse Mixture-of-Experts Transformer for Universal EEG Representation Learning

Yuxuan Wang[1], **Sha Yang**[1], **Jiecheng Ren**[2], **Yusha Liu**[2], **Tianyuan Mei**[1], **Huixing Gou (gouhx@mail.ustc.edu.cn)**[2] **& Xiaochu Zhang (zxcustc@ustc.edu.cn)**[1,2]
[1]Institute of Advanced Technology, University of Science and Technology of China
[2]Division of Life Science and Medicine, University of Science and Technology of China

Abstract

Electroencephalography (EEG) provides a non-invasive and temporally precise window into neural dynamics, yet learning transferable representations across subjects, paradigms, and electrode layouts remains challenging. Existing EEG foundation models typically rely on patch-wise tokenization and dense Transformer blocks, which may weaken fine-grained temporal continuity and provide limited adaptability to heterogeneous neural patterns. We propose MoEEG, a sparse Mixture-of-Experts (MoE) Transformer for generalizable EEG representation learning. MoEEG combines global-local temporal embedding, electrode-aware channel embedding, a dual-path EEGAttention module for temporal and spatial dependency modeling, and sparse expert routing for adaptive representation learning. Linear-probing evaluations on P300, ERN, and motor imagery benchmarks show that MoEEG consistently outperforms BIOT, LaBraM, and EEGPT, with balanced-accuracy improvements over EEGPT of 2.94, 1.69, 0.72, and 1.43 percentage points on PhysioP300, KaggleERN, BCIC-2A, and BCIC-2B, respectively. These results suggest that sparse expert routing and EEG-specific attention can improve cross-task EEG representation learning. Our open source code is available at https://github.com/YuuSenW/MoEEG.git.

Keywords: EEG; EEG Representation Learning; EEG Foundation Model; MoE; EEG Classification

Introduction

Electroencephalography (EEG) provides a non-invasive modality for capturing cerebral activity. As reflections of underlying neural dynamics, EEG signals are essential in clinical diagnostics and cognitive neuroscience. Successful applications range from seizure detection (Boonyakitanont et al., 2020) to brain-computer interfaces (Amin et al., 2019) and affective analysis (Suhaimi et al., 2020). The functional significance of EEG lies in complex neural oscillations that manifest as synchronized spatiotemporal patterns across diverse brain regions (Müller-Putz, 2020). Effectively representing these intricate dynamics bridges raw electrophysiology with sophisticated interpretations of complex cognition. This synergy establishes the foundation for universal EEG representation learning (Jiang et al., 2024).

The Transformer architecture has catalyzed a paradigm shift in Large Language Models (LLMs) (Ouyang et al., 2022; Vaswani et al., 2017). This evolution demonstrates a powerful synergy between self-supervised pretraining and scalable attention mechanisms. Inspired by these advancements, the neuroscience community has developed foundational models for universal EEG representation learning. For example, BIOT (Yang et al., 2023) utilizes a linear Transformer to capture complex token interactions. Similarly, EEG2Vec (Bethge et al., 2022) integrates self-supervised contrastive and reconstruction objectives for high-dimensional feature learning. To enhance generalization, LaBraM (Jiang et al., 2024) introduces neural tokenization via vector-quantized spectrum prediction. EEGPT (G. Wang et al., 2024) employs an ensemble of Transformer blocks within a hybrid reconstruction framework to achieve state-of-the-art (SOTA) performance. Despite the progress of existing EEG models, a critical limitation persists in current modeling. Most architectures discretize EEG sequences via patch partitioning along the temporal dimension (Jiang et al., 2024; G. Wang et al., 2024). Such patch-level tokenization conflicts with the intrinsic continuous nature of neural oscillations and event-related activities. In addition, generic self-attention cannot explicitly disentangle intra-electrode temporal dynamics and inter-electrode spatial correlations, restricting the extraction of physiologically plausible spatiotemporal patterns. This drawback becomes more severe under low signal-to-noise ratios (Cole & Voytek, 2019).

Another limitation of dense Transformer backbones is that the same feed-forward parameters are applied to all samples regardless of subject, session, or task paradigm. Given the strong inter-subject variability and paradigm-specific spectral characteristics of EEG, a single dense pathway may lead to feature interference when trained on heterogeneous datasets (Tran et al., 2026).

To address these limitations, we propose MoEEG, a self-supervised framework integrating a specialized EEGAttention mechanism with a sparse Mixture-of-Experts (MoE) (Du et al., 2022) architecture to capture intricate spatio-temporal dependencies. Within this framework, EEGAttention employs a dual-path design to simultaneously capture comprehensive temporal information and refined spatial dependencies, effectively disentangling non-stationary neural patterns. This mechanism is integrated with a sparse MoE structure, enabling adaptive reconciliation of subject-specific and task-related heterogeneity through dynamic expert routing. Using a self-supervised masked-reconstruction strategy, MoEEG aligns latent representations with the intrinsic temporal and spatial structure of EEG data. We conducted extensive pretraining across five heterogeneous datasets, including SEED (Zheng & Lu, 2015), M3CV (Huang et al., 2022), THU-Benchmark (Y. J. Wang et al., 2017), PhysioNet-MI (Gold-

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

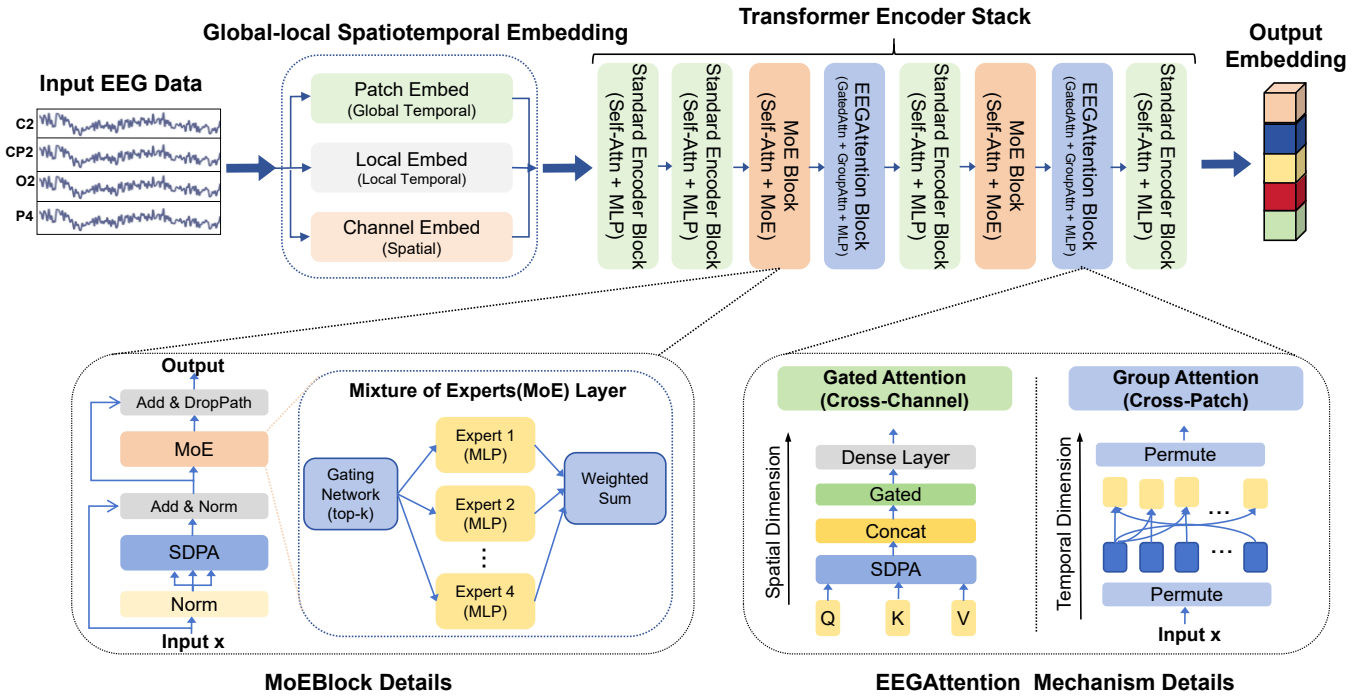

Figure 1: Architecture of the MoEEG framework.

berger et al., 2000), and a laboratory-built Cue-Reactivity dataset, to distill robust neural patterns. Comprehensive cross-validation shows that MoEEG achieves generalization and effectiveness across diverse neurophysiological downstream tasks. Our main contributions are summarized as follows:

1. Propose Global-local Spatiotemporal Embedding, which combines patch-level temporal abstraction, local waveform modeling, and electrode-aware channel encoding.
2. Design EEGAttention, a dual-path attention module that separately models within-channel temporal dependencies and cross-channel spatial dependencies.
3. Introduce sparse MoE routing into EEG foundation modeling. Extensive experiments and ablation studies verify it improves cross-task and cross-subject generalization.

Methods

Model Architecture

As illustrated in Fig. 1, MoEEG maps raw EEG signals $\mathbf{X}_{raw} \in \mathbb{R}^{B\times C\times T}$ (where B denotes the batch size, C is the number of EEG electrodes, and T is the time points) into latent feature. While the embedding module initiates this process by partitioning signals into P patches of dimension D, the computational core resides in a hierarchical Transformer stack. This architecture centers on the integration of specialized EEGAttention and MoE modules, which collectively capture multi-scale neural dynamics to extract robust neurophysiological features from high-dimensional representations.

Global-local Spatiotemporal Embedding The embedding block synthesizes global-local features and spatial topology

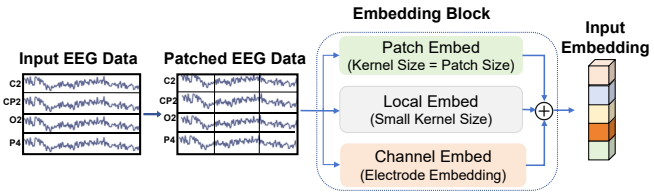

Figure 2: Global-local Spatiotemporal Embedding.

into a unified latent space (Fig. 2). For an input signal $\mathbf{X}_{raw} \in \mathbb{R}^{B\times C\times T}$, the representation is formulated as an additive fusion:

$$\mathbf{X} = \mathbf{E}_{patch} + \mathbf{E}_{local} + \mathbf{E}_{chan} \tag{1}$$

Here, $\mathbf{X} \in \mathbb{R}^{BP\times C\times D}$. The Patch Embedding layer discretizes the signal into $P = \lfloor T/S \rfloor$ non-overlapping segments of size S to ensure temporal consistency. A learnable 2D convolution $\mathcal{F}_{patch}(\cdot)$, with stride S, projects these segments into a D-dimensional manifold:

$$\mathbf{E}_{patch} = \mathcal{F}_{patch}(\mathbf{X}[:, :, 0 : P \cdot S]) \tag{2}$$

Concurrently, Local Embed ($\mathbf{E}_{local}$) captures transient morphological details via 1D convolutions, while Channel Embed ($\mathbf{E}_{chan}$) utilizes learnable electrode-specific embeddings $\mathbf{W}_{chan} \in \mathbb{R}^{C\times D}$ to preserve anatomical topology. This multi-path fusion enables the subsequent Transformer blocks to reconcile global trends with micro-level neural dynamics effectively.

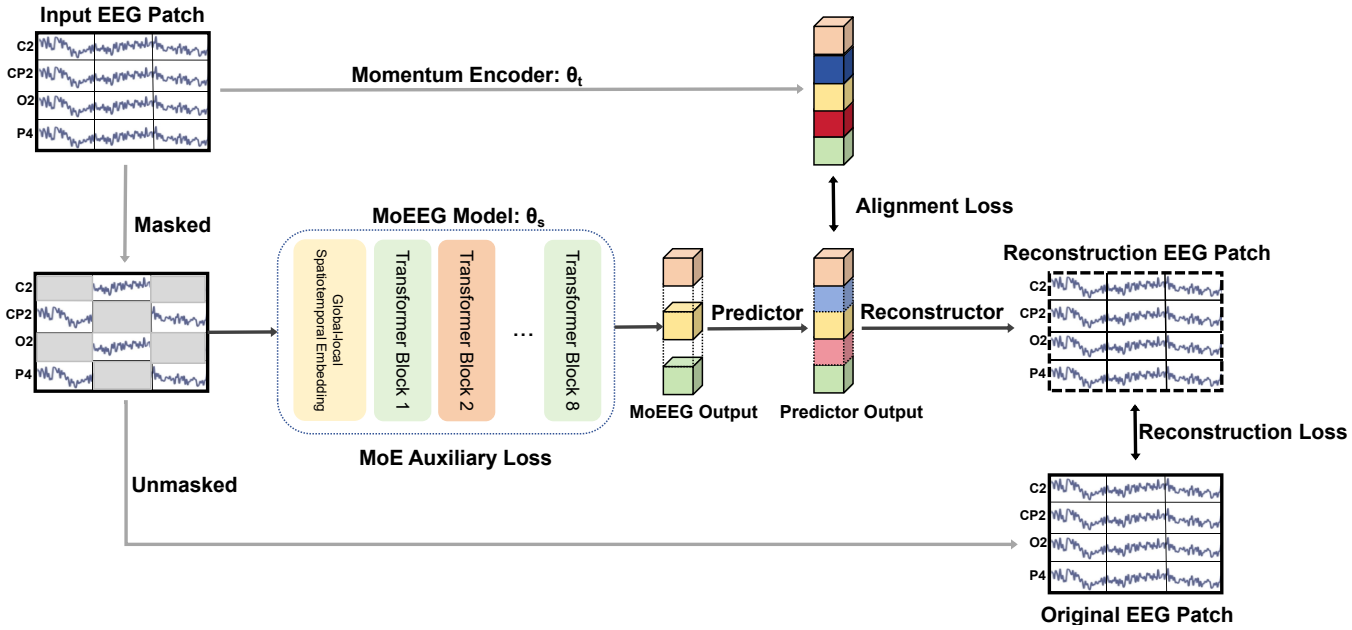

Figure 3: Pre-training of MoEEG.

Sparse Mixture of Experts MoE replaces specific feed-forward networks to capture inter-subject and intra-paradigm variability while maintaining computational efficiency (Du et al., 2022). A gating network routes inputs $\mathbf{X}$ to $\mathbf{N}$ independent experts via a softmax-normalized linear projection, $P(\mathbf{x}) = \text{Softmax}(\mathbf{W}_{\text{gate}}\mathbf{x})$, employing a Top-$k$ selection strategy to activate the most relevant modules. The final output is computed as a weighted summation of these sparse expert contributions:

$$\mathbf{Y}_{MoE} = \sum_{i \in \text{Top-}k} P(\mathbf{x})_i \cdot \text{Expert}_i(\mathbf{x}) \tag{3}$$

To prevent expert collapse, an auxiliary load-balancing loss $\mathcal{L}_{aux} = N \cdot \sum_{j=1}^{N} \mathcal{I}_j \cdot \mathcal{L}_j$ penalizes non-uniform utilization based on expert importance $\mathcal{I}$ and selection load $\mathcal{L}$ (L. Wang et al., 2024). This constraint encourages uniform engagement across the MoE framework, maximizing the representational power for diverse neurophysiological patterns.

EEGAttention Characterized by a dual-path architecture, EEGAttention integrates Group and Gated (Qiu et al., 2025) attention mechanisms to capture non-stationary spatiotemporal dependencies. We propose Group Attention, which permutes the input into $\hat{\mathbf{X}} \in \mathbb{R}^{BC\times P\times D}$ to facilitate attention computation across patches, enabling cross-patch information interaction within individual channels. Computation via Scaled Dot-Product Attention (SDPA) (Vaswani et al., 2017):

$$\mathbf{Y}_{group} = \text{SDPA}(\hat{\mathbf{X}}\mathbf{W}^Q, \hat{\mathbf{X}}\mathbf{W}^K, \hat{\mathbf{X}}\mathbf{W}^V) \tag{4}$$

The resulting representation is permuted to restore the original dimensionality. Parallel to this, the Gated Attention branch (Qiu et al., 2025) captures cross-channel spatial dependencies by instilling input-dependent sparsity via a sigmoid gate $\mathbf{G} = \sigma(\mathbf{X}\mathbf{W}_{gate})$ applied to the attention output, while suppressing pervasive physiological artifacts. Given an embedded tensor $\mathbf{X} \in \mathbb{R}^{BP\times C\times D}$:

$$\mathbf{Y}_{gate} = \text{SDPA}(\mathbf{X}\mathbf{W}^Q, \mathbf{X}\mathbf{W}^K, \mathbf{X}\mathbf{W}^V) \odot \mathbf{G} \tag{5}$$

These concurrent streams are integrated to produce the final normalized representation:

$$\mathbf{Y}_{EEGAttention} = \text{Norm}(\mathbf{X} + \mathbf{Y}_{group} + \mathbf{Y}_{gate}) \tag{6}$$

This synthesis yields temporally comprehensive and spatially salient features, establishing a foundation for high-fidelity neural representation across heterogeneous subject populations and complex neurophysiological scenarios.

Transformer Encoder Stack MoEEG comprises an eight-layer Transformer encoder stack, utilizing Q learnable summary tokens along the channel dimension to distill complex neurophysiological dynamics. To balance representational capacity with computational efficiency, sparse MoE blocks are strategically positioned at layers 3 and 6, while EEGAttention modules are integrated at layers 4 and 7 for targeted feature refinement. A selective distillation process subsequently isolates these summary tokens to filter representations.

Pre-training MoEEG

Inspired by EEGPT (G. Wang et al., 2024), MoEEG adopts a masked self-supervised strategy for universal representation learning. As shown in Fig. 3, the architecture reconstructs occluded signals via the synergy of a latent-space Predictor, Momentum Encoder, and generative Reconstructor. This process captures complex spatiotemporal dependencies by optimizing a denoising objective where input $\mathbf{X}_{raw}$ is corrupted

by a stochastic mask $\mathbf{M}$ to yield $\mathbf{X}_{vis} = \mathbf{X}_{raw} \odot \mathbf{M}$. Such sparsity compels MoEEG to bypass transient noise and internalize essential neurophysiological signatures.

Predictor & Momentum Encoder The predictor serves as the core of the self-supervised alignment task. Comprising eight Transformer encoder blocks, this module maps masked MoEEG latent tokens $\mathbf{H}_{mask}$ into a predictive manifold. This process recovers high-level semantic signatures defined as $\hat{\mathbf{Z}} = \text{Predictor}(\mathbf{H}_{mask})$. To provide stable targets despite the inherent volatility of EEG signals, a momentum encoder (He et al., 2020) is utilized. Its parameters θ_t are derived through an Exponential Moving Average (EMA) (Tarvainen & Valpola, 2017) of the MoEEG weights θ_s. This mechanism ensures a consistent latent distribution across training iterations:

$$\theta_t \leftarrow m\theta_t + (1-m)\theta_s \tag{7}$$

Here, m represents the momentum coefficient. The framework subsequently enforces a spatio-temporal representation alignment by minimizing the Mean Squared Error (MSE) between the estimated features and the layer normalized (LN) (Ba et al., 2016) momentum encoder output $\mathbf{Z}_{target}$:

$$\mathcal{L}_{align} = \|\hat{\mathbf{Z}} - \text{LN}(\mathbf{Z}_{target})\|_2^2 \tag{8}$$

By synchronizing these latent distributions, the model effectively prioritizes subject-invariant neurophysiological patterns over transient artifacts.

Reconstructor Comprising eight Transformer encoder blocks, the reconstructor projects latent tokens $\hat{\mathbf{Z}}$ into the raw signal domain to preserve fine-grained temporal morphology. With $\mathbf{X}_R = \text{Reconstructor}(\hat{\mathbf{Z}})$:

$$\mathcal{L}_{rec} = \|\mathbf{X}_R - \mathbf{X}_{raw}\|_2^2 \tag{9}$$

The framework subsequently synchronizes semantic alignment with generative reconstruction by optimizing a joint objective that incorporates MoE load-balancing:

$$\mathcal{L}_{total} = \mathcal{L}_{align} + \alpha \cdot \mathcal{L}_{rec} + \beta \cdot \mathcal{L}_{aux} \tag{10}$$

Where α and β serve as hyperparameters to balance the different loss components. Such a multi-objective strategy effectively decouples invariant neurophysiological signatures from subject-specific noise.

Experiments

Datasets and Preprocessing

To establish a foundational universal neural representation, MoEEG was pre-trained on a diverse corpus comprising the SEED (Zheng & Lu, 2015), M3CV (Huang et al., 2022), THU-Benchmark (Y. J. Wang et al., 2017), PhysioNet-MI (Goldberger et al., 2000), and a laboratory-built Cue-Reactivity dataset (Table 1). This collection encompasses paradigms such as affective analysis, biometrics, SSVEP, and motor imagery. Offline preprocessing included average re-referencing, μV scaling, and 256 Hz resampling. To resolve spatial discrepancies, diverse electrode layouts were mapped to a standardized 58-channel 10-20 system as illustrated in Fig. 4. Signals were segmented into 4-second non-overlapping windows to construct uniform spatiotemporal grids $\mathbf{X} \in \mathbb{R}^{58\times1024}$, facilitating the extraction of subject-invariant features while mitigating transient experimental noise.

Table 1: Pre-training and Downstream Datasets

Datasets	Paradigms	Rate	Subjects	Targets
PhysioNetMI	MI&ME	160	109	4
THU Benchmark	SSVEP	250	35	40
M3CV	MULTI	250	106	-
SEED	EMO	200	15	3
Cue-Reactivity	CR	250	120	2
PhysioP300	P300	250	9	2
KaggleERN	ERN	250	26	2
BCIC-2A	MI	250	10	4
BCIC-2B	MI	200	10	2

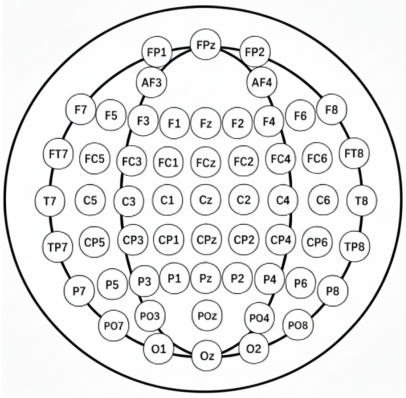

Figure 4: Electrode Location.

Model performance and generalizability were evaluated on the PhysioP300 (Goldberger et al., 2000), Kaggle-ERN (Perrin et al., 2012), BCIC-2A (Tangermann et al., 2012), and BCIC-2B (Steyrl et al., 2016), covering both event-related potentials and motor imagery paradigms. While signals were resampled to 256 Hz and scaled to μV for alignment with the pre-trained feature space. Furthermore, a 0–38 Hz bandpass filter was applied specifically to motor imagery datasets to isolate pertinent sensorimotor rhythms and suppress unrelated physiological artifacts.

Model Implementation Details

MoEEG comprises an 8-layer Transformer backbone with 40M parameters, utilizing a 512-dimensional hidden space, 8 attention heads, and 4 sparse experts per MoE layer with Top-2 selection strategy. Four learnable summary tokens are integrated along the channel dimension to facilitate high-level

Table 2: Results on Downstream Tasks

Datasets	Model	Balanced Accuracy	Cohen's Kappa	Weighted F1 / AUROC
PhysioP300	BIOT	0.5672±0.0247	0.1243±0.0533	0.5841±0.0290
	LaBraM	0.6446±0.0057	0.3021±0.0124	0.6988±0.0095
	EEGPT	0.6517±0.0038	0.3149±0.0047	0.7053±0.0052
	MoEEG	**0.6811±0.0082**	**0.3625±0.0075**	**0.7364±0.0087**
KaggleERN	BIOT	0.5203±0.0139	0.0853±0.0252	0.5567±0.0147
	LaBraM	0.5646±0.0170	0.1131±0.0129	0.5743±0.0053
	EEGPT	0.5967±0.0034	0.1935±0.0089	0.6678±0.0058
	MoEEG	**0.6136±0.0053**	**0.2304±0.0112**	**0.7213±0.0070**
BCIC-2A	BIOT	0.4353±0.0124	0.2578±0.0143	0.4198±0.0385
	LaBraM	0.4579±0.0046	0.2772±0.0610	0.4493±0.0062
	EEGPT	0.4569±0.0103	0.2792±0.0124	0.4502±0.0032
	MoEEG	**0.4641±0.0127**	**0.2859±0.0174**	**0.4561±0.0121**
BCIC-2B	BIOT	0.6510±0.0144	0.3527±0.0278	0.7543±0.0228
	LaBraM	0.6883±0.0094	0.3948±0.0093	0.7821±0.0117
	EEGPT	0.7215±0.0032	0.4357±0.0037	0.8063±0.0081
	MoEEG	**0.7358±0.0071**	**0.4631±0.0097**	**0.8129±0.0134**

feature distillation. During embedding, a patch size of $S = 64$ is utilized for global temporal partitioning via 2D convolutions, while the Local Embed branch employs 1×3 1D convolutions to capture micro-level transient features.

The model was pre-trained for 100 epochs on an RTX 4060 (8G) GPU, using mixed-precision and a batch size of 16. A stochastic masking strategy was applied to 50% of temporal patches and 80% of channel patches to compel the learning of essential neurophysiological signatures. Optimization followed the AdamW optimizer and OneCycleLR scheduler with a peak learning rate of 6×10^{-5} and loss weights $\alpha = 2$ and $\beta = 0.02$, while the momentum coefficient m increased from 0.96 to 1.0 to ensure latent stability.

Evaluation strategy

To facilitate a direct comparison with the EEGPT model, we adopted a standardized evaluation protocol across all downstream benchmarks. For the PhysioP300, BCIC-2A, and BCIC-2B datasets, performance was validated using a Leave-One-Subject-Out (LOSO) cross-validation scheme. In contrast, the Kaggle-ERN task was evaluated via 4-fold cross-validation, specifically utilizing 10 subjects for testing within each fold.

Across all scenarios, we employed a linear probing (Fig. 5) to assess the quality of the learned representations. During this process, the pre-trained MoEEG parameters remained frozen, and only the final linear classification layer was updated.

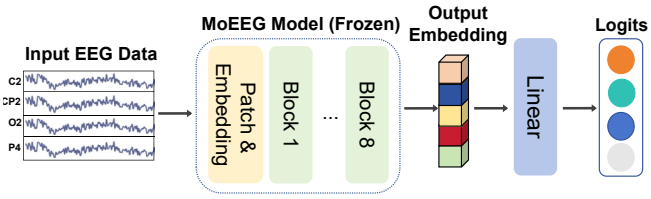

Figure 5: Linear Probing.

Baselines & Metrics

MoEEG's performance was benchmarked against leading architectures, including BIOT (Yang et al., 2023), LaBraM (Jiang et al., 2024), and EEGPT (G. Wang et al., 2024). All baseline comparison models are sourced from their officially released code repositories. Evaluation relied on a robust metric suite: Balanced Accuracy, AUROC, Weighted F1-score, and Cohen's Kappa. Specifically, AUROC served for binary classification, while the Weighted F1-score was utilized for multi-class paradigms.

Comparisons with Other Methods

Table 2 summarizes MoEEG's performance across various neurophysiological paradigms. The model effectively handles low signal-to-noise ratios in ERP tasks, consistently outperforming architectures such as BIOT, LaBraM, and EEGPT. For PhysioP300, MoEEG achieves a 0.6811 balanced ac-

Table 3: Core Component Ablation Results

Variants	$\mathcal{L}_{total}$	PhysioP300-BAC	BCIC-2B-BAC
A: w/o EA	1.3152	0.6499±0.0024	0.6835±0.0049
B: w/o MoE	**1.2253**	0.6560±0.0061	0.6849±0.0021
C: w/o $\mathcal{L}_{aux}$	1.3045	0.6566±0.0055	0.6703±0.0034
D: with all	1.2337	**0.6811±0.0082**	**0.7358±0.0071**

curacy. On KaggleERN, it reaches 0.6136 with a 0.2304 Cohen's Kappa. These gains suggest that the dual-path EEGAttention mechanism successfully isolates salient components from background noise through its integrated Gated and Group branches.

Generalizability is further demonstrated in MI paradigms where sensorimotor rhythms exhibit substantial variability. MoEEG establishes new standards for BCIC-2A and BCIC-2B, achieving Balanced Accuracies of 0.4641 and 0.7358, respectively. This stability highlights the capacity of the sparse MoE architecture to reconcile subject idiosyncrasies via dynamic expert routing. Achieving these results under linear probing with frozen backbone parameters validates the task-invariant nature and fidelity of the learned representations across disparate settings.

Ablation Experiments

Core Component Ablation Experiments Ablation experiments on PhysioP300 and BCIC-2B evaluated four MoEEG configurations (Table 3, Variants A-D). We replaced removed components with standard Transformer layers to maintain model depth while isolating modular impact. The complete MoEEG framework in Variant D yielded the highest balanced accuracy of 0.6811 and 0.7358 on the two datasets. These results validate that integrating spatiotemporal attention with dynamic routing is essential for universal EEG decoding. Specifically, the performance decline in Variant A suggests that standard layers fail to reconstruct the dependencies lost during patch discretization.

The results also indicate that these modules play complementary roles in model performance. Although Variant B recorded the lowest training loss of 1.2253, this resulted from removing the $\mathcal{L}_{aux}$ penalty rather than achieving superior convergence. The inferior accuracy of Variant B proves that static computational paths cannot handle high subject heterogeneity. Variant C confirms the importance of the load-balancing constraint, as disabling $\mathcal{L}_{aux}$ led to expert collapse and restricted representational capacity. Collectively, these findings show that EEGAttention extracts fidelity features while MoE structure ensures cross-subject alignment for EEG representation learning.

EEGAttention Ablation Experiments The contributions of Group and Gated Attention mechanisms are evaluated across five configurations (Table 4, Variants A–E). To maintain architectural consistency, the dual-branch (Variant B and Variant D) pair one specialized module with a standard attention mechanism instead of its specialized counterpart. The complete MoEEG (Variant E) achieves the highest BAC and lowest reconstruction loss, confirming that these mechanisms are not redundant.

Table 4: EEGAttention Ablation Results

Variants	$\mathcal{L}_{total}$	$\mathcal{L}_{rec}$	PhysioP300-BAC
A: Group Only	1.2454	0.8722	0.6733±0.0039
B: Group Dual	1.2402	0.8734	0.6784±0.0041
C: Gated Only	1.2611	0.8922	0.6659±0.0052
D: Gated Dual	1.2562	0.8859	0.6721±0.0057
E: with all	**1.2337**	**0.8683**	**0.6811±0.0082**

Comparative analysis reveals that Group Only (Variant A) outperforms Gated Only (Variant C) by 0.74 percentage points, indicating that global context restoration via patch-wise interaction is more fundamental for EEG representation than channel-wise gating alone. While the dual variants (Variant B and Variant D) improve upon single-path counterparts, they still underperform compared to the full model (Variant E). This confirms that the synergistic integration of Group and Gated mechanisms is the primary driver of MoEEG's robust feature extraction.

Conclusion

In this paper, we presented MoEEG, a sparse MoE Transformer for universal EEG representation learning. By combining global-local embedding, dual-path EEGAttention, and sparse expert routing, MoEEG improves linear-probing performance on several ERP and MI benchmarks. Ablation studies suggest that both EEGAttention and MoE routing contribute to the observed gains. However, the current evaluation is limited to a small set of downstream datasets and frozen-backbone probing. Future work will investigate larger-scale pre-training, fine-tuning and few-shot transfer, explicit electrode-geometry modeling, and more detailed analyses of expert specialization.

Acknowledgments

This work was supported by grants from National Key R&D Program of China (2024YFF0507600), The Chinese National Programs for Brain Science and Brain-like Intelligence Technology (2021ZD0202101), The National Natural Science Foundation of China (32571266, 32171080, 32400919, and 32200914), the Project of Guizhou Key Laboratory of Artificial Intelligence and Brain-inspired Computing QianKeHe Platform (ZSYS[2024]001), Natural Science Foundation of Anhui Province (2408085QC081), the Humanities and Social Science Fund of the Ministry of Education of China (24YJCZH014), Anhui Provincial Health Scientific Research Project (AHWJ2024Aa10016), and Shanghai Key Laboratory of Brain-Machine Intelligence for Information Behavior.

References

Amin, S. U., Alsulaiman, M., Muhammad, G., Mekhtiche, M. A., & Hossain, M. S. (2019). Deep learning for eeg motor imagery classification based on multi-layer cnns feature fusion. *Future Generation Computer Systems*, *101*, 542–554.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bethge, D., Hallgarten, P., Groe-Puppendahl, T. A., Kari, M., Chuang, L., Ozdenizci, O., & Schmidt, A. (2022). Eeg2vec: Learning affective eeg representations via variational autoencoders. *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 3150–3157.

Boonyakitanont, P., Lek-Uthai, A., Chomtho, K., & Songsiri, J. (2020). A review of feature extraction and performance evaluation in epileptic seizure detection using eeg. *Biomedical Signal Processing and Control*, *57*, 101702.

Cole, S., & Voytek, B. (2019). Cycle-by-cycle analysis of neural oscillations. *Journal of neurophysiology*, *122*(2), 849–861.

Du, N., Huang, Y. P., Dai, A. M., Tong, S., Lepikhin, D., Xu, Y. Z., Krikun, M., Zhou, Y. Q., Yu, A. W., Firat, O., et al. (2022). Glam: Efficient scaling of language models with mixture-of-experts. *International Conference on Machine Learning*, 5547–5569.

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C. K., & Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, *101*(23), e215–e220.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Huang, G., Hu, Z. X., Chen, W. Z., Zhang, S. R., Liang, Z., Li, L. L., Zhang, L., & Zhang, Z. G. (2022). M3cv: A multi-subject, multi-session, and multi-task database for eeg-based biometrics challenge. *NeuroImage*, *264*, 119666.

Jiang, W. B., Zhao, L. M., & Lu, B. L. (2024). Large brain model for learning generic representations with tremendous EEG data in BCI. *The Twelfth International Conference on Learning Representations*.

Müller-Putz, G. R. (2020). Chapter 18 - electroencephalography. In N. F. Ramsey & J. D. R. Millán (Eds.), *Brain-computer interfaces* (Vol. 168). Elsevier.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Perrin, M., Maby, E., Daligault, S., Bertrand, O., & Mattout, J. (2012). Objective and subjective evaluation of online error correction during p300-based spelling. *Advances in Human-Computer Interaction*, *2012*(1), 578295.

Qiu, Z. H., Wang, Z. K., Zheng, B., Huang, Z. Y., Wen, K. Y., Yang, S. L., Men, R., Yu, L., Huang, F., Huang, S. Z., Liu, D. H., Zhou, J. R., & Lin, J. Y. (2025). Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Steyrl, D., Scherer, R., Faller, J., & Müller-Putz, G. R. (2016). Random forests in non-invasive sensorimotor rhythm brain-computer interfaces: A practical and convenient non-linear classifier. *Biomedical Engineering/Biomedizinische Technik*, *61*(1), 77–86.

Suhaimi, N. S., Mountstephens, J., & Teo, J. (2020). Eeg-based emotion recognition: A state-of-the-art review of current trends and opportunities. *Computational Intelligence and Neuroscience*, *2020*(1), 8875426.

Tangermann, M., Müller, K. R., Aertsen, A., Birbaumer, N., Braun, C., Brunner, C., Leeb, R., Mehring, C., Miller, K. J., Müller-Putz, G. R., et al. (2012). Review of the bci competition iv. *Frontiers in Neuroscience*, *6*, 55.

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, *30*.

Tran, X.-T., Vo, T.-N., Vu, S.-T., Tran, T.-T., Nguyen, M.-D., Do, T., & Lin, C.-T. (2026). Inter-and intra-subject variability in eeg: A systematic survey. *arXiv preprint arXiv:2602.01019*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, G., Liu, W., He, Y., Xu, C., Ma, L., & Li, H. (2024). Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. *Advances in Neural Information Processing Systems*, *37*, 39249–39280.

Wang, L., Gao, H. Z., Zhao, C. G., Sun, X., & Dai, D. M. (2024). Auxiliary-loss-free load balancing strategy for mixture-of-experts.

Wang, Y. J., Chen, X. G., Gao, X. R., & Gao, S. K. (2017). A benchmark dataset for ssvep-based brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *25*, 1746–1752.

Yang, C. Q., Westover, M. B., & Sun, J. M. (2023). Biot: Biosignal transformer for cross-data learning in the wild. *Thirty-seventh Conference on Neural Information Processing Systems*.

Zheng, W. L., & Lu, B. L. (2015). Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, *7*(3), 162–175.