BRAINMAMBA: A SCALABLE EEG FOUNDATION MODEL FOR COUPLED MULTI-CHANNEL SPATIOTEMPORAL EVOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale representation learning from heterogeneous electroencephalography (EEG) data is essential for building generalizable EEG foundation models across diverse brain states, subjects, and downstream tasks. However, building scalable EEG foundation models faces two fundamental challenges: (1) data heterogeneity, as diverse channel configurations, acquisition protocols, and task settings hinder unified large-scale pretraining and cross-dataset transfer; and (2) neural dynamics modeling, as existing backbones struggle to capture the coupled spatial interactions and temporal evolution intrinsic to multi-channel brain activity. To address these challenges, we propose BrainMamba, a scalable foundation model that formulates large-scale EEG modeling as structured neural state evolution. At its core, BrainMamba models multi-channel EEG as a coupled dynamical system, where latent neural states evolve over time through structured state-space transitions while being continuously shaped by cross-channel interactions. By jointly capturing temporal evolution and inter-channel coupling, this formulation enables coherent modeling of distributed brain dynamics over long temporal contexts. To further accommodate the substantial heterogeneity of EEG recordings, the state evolution is adaptively specialized through expert routing, allowing different neural patterns across subjects, tasks, and recording conditions to be captured within a unified backbone. This state-centric design naturally scales across heterogeneous channel configurations and sequence lengths, providing a unified foundation for large-scale EEG pretraining. BrainMamba adopts a large-scale mask-reconstruction-based pretraining paradigm, where masked spatiotemporal segments are reconstructed to learn EEG contextual semantic representations. For evaluation, we propose a novel validation paradigm for EEG foundation models, comprising task-constrained pretraining, channel-consistent transfer, and self-evaluation, to rigorously evaluate model generalization and robustness. Extensive experiments on more than 10 datasets across emotion, cognition, and disorder benchmarks demonstrate that BrainMamba consistently outperforms existing EEG foundation models, exhibiting superior cross-scenario generalization and strong scalability, highlighting its potential for real-world deployment.

1 INTRODUCTION

Multi-channel electroencephalogram (EEG) signals provide a non-invasive window into brain activity, capturing the collective electrophysiological responses of cortical neuronal populations via scalp electrodes. Specifically, these signals originate from postsynaptic potentials in pyramidal neurons driven by external stimuli or internal modulation, which propagate through neural and volume conduction to be synchronously recorded by electrode arrays Jackson & Bolger (2014). Due to their high temporal resolution and non-invasive nature, multi-channel EEG encodes rich neurophysiological data, characterizing the spatial distribution of neural activity while capturing rapid temporal dynamics, thus attracting increasing attention in neuroscience and intelligent sensing Gu et al. (2021). Early neuroscientific research primarily relied on the analysis of event-related potentials (ERPs), such as P300 Turnip & Hong (2012) and N1 Bruyns-Haylett et al. (2017), where characteristic channel–time waveform patterns were manually identified for specific cognitive responses.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

With the rapid advancement of machine learning and deep learning, EEG analysis has evolved toward two primary paradigms. The first focuses on hybrid frameworks that integrate handcrafted neurophysiological features, such as Power Spectral Density (PSD) and Differential Entropy (DE), with deep architectures to improve discriminative modeling Duan et al. (2013); Alsolamy & Fattouh (2016). The second paradigm comprises fully end-to-end models that directly learn latent representations from raw signals Dose et al. (2018). Representative approaches include Convolutional Neural Networks (CNNs) Zhou et al. (2018); Dai et al. (2020), for spatiotemporal filtering, Recurrent Neural Networks (RNNs/LSTMs) Alhagry et al. (2017) for modeling sequential dynamics, and more recently, Transformers Song et al. (2022); Xie et al. (2022) and State Space Models (SSMs) Zhou et al. (2024); Gu & Dao (2023) for capturing long-range dependencies in brain signals. These approaches have substantially advanced EEG modeling paradigms, forming a fundamental basis for real-world brain–computer interface (BCI) systems Lu et al. (2025).

Recently, inspired by the remarkable success of large-scale pretraining models in natural language processing Zhao et al. (2023) and computer vision Zhang et al. (2024), EEG foundation models have emerged as a promising paradigm for learning generalizable representations from large-scale neural recordings Lai et al. (2025); Rafiei et al. (2022); Weng et al. (2025). By exploiting rich spatiotemporal context in multi-channel EEG, these models aim to support unified transfer across subjects, datasets, and downstream tasks, including emotion recognition, sleep monitoring, and neurological disorder classification. However, scaling EEG foundation models remains fundamentally challenging due to two characteristics of real-world EEG data. First, **data heterogeneity** across channel configurations, acquisition protocols, subjects, and tasks makes it difficult to unify diverse EEG corpora for large-scale pretraining and cross-distribution transfer. Second, **coupled neural dynamics** require models to jointly characterize temporal evolution and cross-channel interactions, while existing EEG backbones often lack appropriate inductive biases for such structured spatiotemporal dynamics. These challenges limit the scalability and generalization of current EEG foundation models in real-world BCI applications.

The challenge of **EEG data heterogeneity** arises from substantial variations in channel configurations and signal distributions across subjects, recording devices, and experimental protocols. Many existing EEG foundation models implicitly assume fixed electrode layouts or predefined channel structures, tightly coupling learned representations to specific spatial configurations and making them sensitive to variations in electrode placement and channel availability. Meanwhile, models such as CBraMod Wang et al. (2024a) primarily adopt single-dataset pretraining, which limits exposure to the distributional diversity and domain shifts inherent in EEG recordings collected under different devices, protocols, and subject populations. Consequently, existing models tend to learn configuration- or dataset-dependent representations rather than transferable structures shared across heterogeneous EEG recordings, limiting their scalability and robustness under cross-subject, cross-device, and cross-dataset settings.

The challenge of **coupled neural dynamics modeling** stems from the inherently coordinated evolution of multi-channel EEG, where temporal state transitions are continuously intertwined with interactions across distributed channels. Existing EEG foundation models, however, predominantly treat EEG as spatiotemporal sequences and capture these dependencies through factorized or alternating operations Chen et al. (2025a; 2024). For example, ALFEE Xiong et al. (2025) employs separated spatial and temporal attention mechanisms, while CBraMod Wang et al. (2024a) alternates between independent spatial and temporal attention modules. Although effective for extracting spatial and temporal features, such formulations provide limited inductive bias for characterizing their *joint evolution*: how distributed neural states evolve over time while being shaped by cross-channel interactions. This mismatch becomes particularly restrictive over long temporal contexts, where coherent state evolution is essential for capturing complex brain dynamics. Therefore, moving beyond factorized sequence encoding toward a unified formulation of structured neural state evolution is critical for scalable EEG representation learning.

To solve the above challenges, we propose **BrainMamba**, a scalable Mamba-based foundation model framework for large-scale EEG pretraining. At its core, BrainMamba introduces a novel **State Mamba** module that leverages structured state-space dynamics to model the coupled evolution of multi-channel EEG signals, jointly capturing temporal state transitions and cross-channel interactions. By characterizing spatial and temporal dependencies within a unified dynamical formulation, State Mamba mitigates the limitations of factorized spatiotemporal modeling and enables effective learning of complex neural dynamics. To further improve representation capacity and

adaptability to heterogeneous EEG distributions, BrainMamba incorporates spatiotemporal Transformers and a Mixture-of-Experts (MoE) mechanism within a unified backbone, preserving structured cross-channel interactions while dynamically adapting to diverse channel configurations and data distributions. For large-scale pretraining, we adopt masked reconstruction of spatiotemporal EEG segments to learn transferable contextual representations, together with round-robin scheduling and temperature-controlled sampling to ensure balanced yet adaptive learning across heterogeneous datasets. We further introduce a comprehensive evaluation paradigm for EEG foundation models, covering task-constrained transfer, channel-consistent transfer, self-supervised evaluation, and real-world assessment, enabling systematic evaluation of model generalization and robustness across diverse scenarios. Extensive experiments on multiple EEG benchmarks and real-world BCI data demonstrate that BrainMamba consistently outperforms existing methods across a wide range of tasks, exhibiting strong cross-subject, cross-dataset, and cross-task generalization. Our main contributions are summarized as follows:

(1) **Novel EEG-based foundation model architecture.** We propose BrainMamba, a Mamba-based EEG foundation model that integrates State Mamba, spatiotemporal Transformers and a MoE mechanism within a unified architecture, enabling joint modeling of coupled spatiotemporal dependencies and latent neural state dynamics. BrainMamba effectively captures complex neural dynamics while improving generalization across heterogeneous EEG distributions, resulting in a scalable and robust foundation architecture for multi-channel EEG representation learning.

(2) **Large-scale Pretraining Framework**. We develop a unified pretraining framework tailored for large-scale and multi-dataset EEG learning, integrating mask-reconstruction with round-robin scheduling and temperature-controlled sampling, enabling efficient and balanced learning from heterogeneous EEG corpora and improving representation generalization across diverse distributions.

(3) **Realistic and comprehensive evaluation paradigm**. Innovatively, we propose a novel validation protocol for EEG foundation models, including task-constrained transfer, channel-consistent transfer, self-supervised evaluation, and real-world assessment, which simulates real-world deployment scenarios with heterogeneous data distributions, varying channel configurations, and cross-domain shifts, enabling systematic evaluation of generalization and robustness.

2 METHODOLOGY

2.1 STRUCTURE-AWARE TIME-FREQUENCY EMBEDDING

We construct structure-aware time-frequency embedding module that maps raw multi-channel EEG signals into a unified token space while preserving their time-frequency characteristics and structural relative positional information. Specifically, this module combines dual-branch time-frequency embedding with spatial relative positional encoding.

Dual-branch time-frequency embedding. Both temporal waveforms and spectral characteristics are essential for representing EEG activity. Therefore, we adopt a dual-branch embedding design to preserve both temporal and spectral information. In the **waveform branch**, multiple 1D convolutional layers with different kernel sizes are employed to capture temporal patterns at multiple scales. Convolutional layers with different receptive fields encode waveform patterns at multiple temporal scales into EEG tokens. In the **frequency branch**, each segment is transformed into the frequency domain via the Fast Fourier Transform (FFT), and the spectral magnitude is computed to capture energy distributions across frequency bands, which encode oscillatory patterns and overall rhythmic structures. The two branches are fused via element-wise addition to produce unified EEG tokens that encode time–frequency semantics, which can be formulated as follows:

$$\mathbf{z}_{i,j} = \underbrace{f_{\mathrm{w}}(\mathcal{C}(\mathbf{X}_{i,j}))}_{\text{Wave Token } \mathbf{z}^1_{i,j}} + \underbrace{f_{\mathrm{f}}\big(|\mathcal{F}(\mathbf{X}_{i,j})|\big)}_{\text{Frequency Token } \mathbf{z}^2_{i,j}} \tag{1}$$

where $\mathcal{C}(\cdot)$ denotes the multi-scale convolutional operator in the waveform branch, $\mathcal{F}(\cdot)$ represents the Fourier transform, $|\cdot|$ extracts the spectral magnitude, and $f_{\mathrm{w}}(\cdot)$ and $f_{\mathrm{f}}(\cdot)$ are learnable mapping functions that project waveform and frequency features into the embedding space, respectively. Details of the Structure-Aware Time–Frequency Embedding are illustrated in Figure 5 in the Appendix.

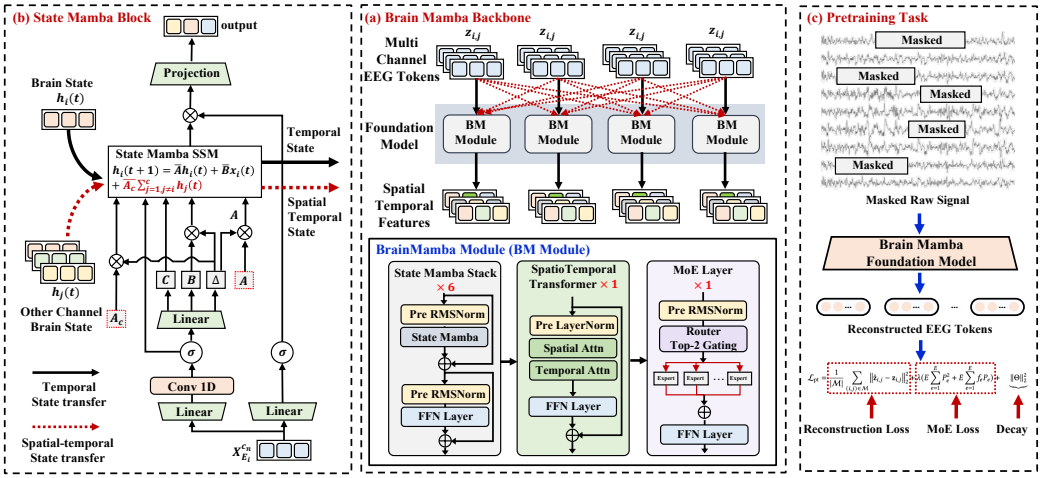


Figure 1: Overview of the proposed BrainMamba framework. (a) The BrainMamba backbone integrates stacked State Mamba layers with spatiotemporal Transformer modules and MoE layers, forming a unified architecture for scalable and expressive representation learning. (b) The State Mamba block captures the coupled spatiotemporal dynamics of multi-channel EEG signals through structured state space modeling with input-dependent parameterization. (c) A masked reconstruction pretraining objective is adopted to learn intrinsic EEG representations by recovering masked signal segments, facilitating large-scale pretraining and effective transfer to diverse downstream tasks.

Spatial relative position encoding. Inspired by the Conditional Positional Encoding in Wang et al. (2024a), BrainMamba adopts corresponding spatial encoding mechanisms for different channel configurations. The spatial relative positional encoding utilizes a 2D convolutional window to model interactions between each patch and its surrounding long-range spatial context and short-range temporal context, thereby enabling structure-aware relative positional encoding. Unlike absolute positional encoding relies on fixed positional indices, spatial relative positional encoding captures positional information through relational patterns among channels. This design enables dynamic encoding of channel ordering via relative dependencies, remaining effective under varying and heterogeneous channel configurations. Similarly, we adopt asymmetric convolutional kernels Wang et al. (2024a), where a larger receptive field is assigned to the spatial dimension to capture long-range cross-channel context, while a smaller receptive field is used along the temporal dimension to model short-range temporal dependencies. This design is motivated by the distinct characteristics of the spatial and temporal dimensions. Along the channel dimension, a larger receptive field is required to capture meaningful spatial relationships, as adjacent channels in the matrix are not necessarily semantically adjacent in the underlying neurophysiological space. In contrast, temporal patches exhibit high local similarity, making short-range context sufficient for accurate position encoding and fine-grained temporal modeling. Therefore, the relative positional encoding is generated in an input-conditioned manner, explicitly exploiting the local two-dimensional spatiotemporal structure to generate adaptive positional representations. This effectively accommodates heterogeneous EEG datasets, varying channel configurations. The detailed computation is given as: $z_{i,j} = z_{i,j} + CPE(z)_{i,j}$, where $CPE(\cdot)$ is the asymmetric convolutional network.

2.2 Unified BrainMamba Architecture

Different from conventional approaches that either directly adopt Transformer architectures from language models as the backbone or employ decoupled temporal and spatial Transformers in serial or parallel stacking paradigms, BrainMamba introduces a unified spatiotemporal modeling framework centered on structured state-space dynamics. It explicitly characterizes the collaborative evolution of multi-channel EEG signals, enabling efficient joint modeling of spatiotemporal dependencies and enhanced generalization across diverse and heterogeneous scenarios. Specifically, as illustrated in Figure 1, the BrainMamba architecture integrates the State Mamba module, spatiotemporal Transformer components, and a Mixture-of-Experts (MoE) mechanism within a unified framework.

2.2.1 STATE MAMBA MODULE.

The State Mamba module serves as the core computational unit of BrainMamba. It constructs a multi-channel state transition space and extends the conventional Mamba model to jointly model multi-channel time-series dynamics. In the standard Mamba state-space model, the continuous state update is described by the differential equation: $\mathbf{h}'(t) = \mathbf{A}\mathbf{h}(t) + \mathbf{B}\mathbf{x}(t), \mathbf{y}(t) = \mathbf{C}\mathbf{h}(t)$. In the multi-channel dynamical setting, the state evolution of each channel is influenced by other channels. Accordingly, the differential equation can be reformulated as:

$$\mathbf{h}_i'(t) = \mathbf{A}\mathbf{h}_i(t) + \mathbf{B}\mathbf{x}_i(t) + \sum_{j \neq i} \mathbf{A}_j^c \mathbf{h}_j(t) + \sum_{j \neq i} \alpha_j \mathbf{x}_j(t) \tag{2}$$

where $\mathbf{h}_i(t) \in \mathbb{R}^d$ denotes the hidden state of the i-th channel at time t, and $\mathbf{x}_i(t) \in \mathbb{R}^d$ represents the corresponding input signal. $\mathbf{A}, \mathbf{B}$ are channel-specific state transition matrices that govern the intra-channel state dynamics. In the state-space formulation, cross-channel state transitions are explicitly considered, where $\mathbf{A}_j^c$ models the spatial state transitions from channel j to channel i, and α_j is a scalar weight controlling the influence of input from other channels. By applying the forward Euler discretization, the continuous form of the differential equation can be approximated as: $\mathbf{h}_i(t + \Delta) = \Delta \cdot \mathbf{h}_i'(t) + \mathbf{h}_i(t)$. Accordingly, the multi-channel state transition equation can be derived as:

$$h_i(t+\Delta) = \underbrace{(\Delta\mathbf{A}+\mathbf{I})h_i(t)+\Delta\mathbf{B}x_i(t)}_{\text{Current channel state transition}} + \underbrace{\Delta\sum_{j\neq i}^{c}\mathbf{A_j^c}h_j(t)+\Delta\sum_{j\neq i}^{c}\alpha_j x_j(t)}_{\text{Spatial influence from other channels}}. \tag{3}$$

In the discretized state transition equation, the channel state at each time step is spatiotemporally coupled with other channels, leading to collaborative state transitions. Following the assumptions in Li et al. (2024) and Weng et al. (2026), we introduce two key simplifications to make the model tractable: (1) **Temporal sequential dependency**, where the state of each channel is influenced by the previous states of other channels rather than their current inputs, thereby modeling cross-channel interactions as delayed state dependencies and avoiding redundant instantaneous coupling; (2) **shared spatial transition mapping**, where a shared transformation is used to model cross-channel interactions, reducing model complexity while preserving essential coupling structures. This simplification substantially reduces computational complexity. Furthermore, by adopting a shared parameter matrix for spatial state transitions, the model can flexibly accommodate heterogeneous channel configurations, enabling more unified and scalable modeling across diverse EEG datasets. The simplified state transition equation is given as follows:

$$h_i(t+\Delta) \overset{\text{simplify}}{=} \underbrace{(\Delta\mathbf{A}+\mathbf{I})h_i(t)+\Delta\mathbf{B}x_i(t)}_{\text{Current channel state transfer}} + \underbrace{\Delta\mathbf{A^c}\sum_{j=1,j\neq i}^{j=c}h_j(t)}_{\text{Spatial transfer}}. \tag{4}$$

The State Mamba block adopts a PreNorm residual architecture, where RMSNorm is applied before both the State Mamba module and the feed-forward network (FFN), with residual connections preserving stable gradient flow. State Mamba extracts high-level semantics that are consistent with the underlying neural dynamics of EEG signals.

2.2.2 SPATIOTEMPORAL TRANSFORMER

Why we still need spatiotemporal Transformers in BrainMamba ?: State Mamba excels at modeling continuous spatiotemporal dynamics through structured state transitions, but suffers from limited expressiveness in capturing high-order cross-channel interactions due to its implicit low-rank state representation. In the deep EEG foundation model, deep stacking of State Mamba can cause information to be compressed into a limited dynamical subspace. We expand the recursive formulation of Mamba:$h_t = \sum_{\tau=1}^{t} A^{t-\tau}Bx_\tau, y_t = \sum_{\tau=1}^{t} CA^{t-\tau}Bx_\tau$. Thus, the influence of a historical EEG token x_τ on the spatiotemporal representation y_t is mediated by: $K_{t,\tau} = CA^{t-\tau}B$. The matrix rank constraints can be calculated:

$$\text{rank}(K_{t,\tau}) = \text{rank}(CA^{t-\tau}B) \leq d_s \tag{5}$$

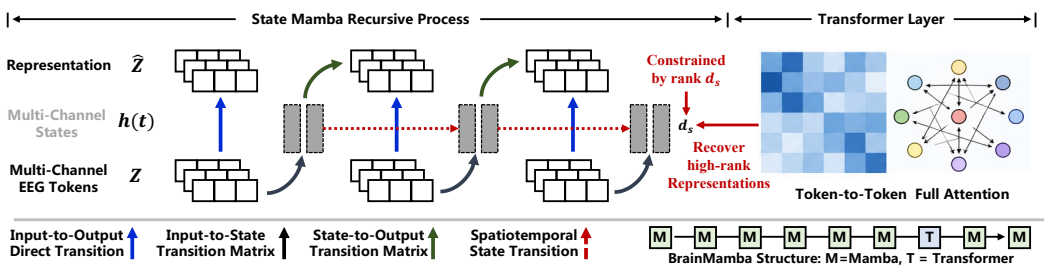


Figure 2: Low-rank compression and high-rank reconstruction in BrainMamba. To restore expressive capacity, spatiotemporal Transformer modules are periodically inserted to reconstruct high-rank representations through global interactions.

where d_s is the feature dimension of h. Due to matrix rank constraints, all token-to-token interactions are restricted to a subspace of dimension at most d_s. Stacking all historical inputs $z_{1:t} \in \mathbb{R}^{t \cdot d}$, the hidden states can be computed in parallel:$h_t = \mathcal{P}_t z_{1:t}$, where $\mathcal{P}_t = [A^{t-1}B, A^{t-2}B, ..., B] \in \mathbb{R}^{d_s \times (t \cdot d)}$. Since $\text{rank}(z_1! \to !h_t) \leq d_s$, regardless of the length of the historical EEG token sequence, the resulting state is confined to a d_s-dimensional subspace and thus retains at most d_s linearly independent directions. Consequently, there exist numerous non-zero perturbations Δz such that $h_t(z + \Delta z) = h_t(z)$ and $y_t(z + \Delta z) = y_t(z)$, indicating strictly irreversible information compression. This phenomenon becomes even more pronounced in multi-channel, multi-temporal EEG token sequences.

To address this limitation, we interleave lightweight spatiotemporal Transformer layers to periodically recompose global dependencies and enhance the model's capacity for non-local interaction modeling. The spatiotemporal Transformer consists of spatial and temporal Transformer modules. The spatial Transformer focuses on capturing cross-channel interactions and structural relationships among EEG patches, while the temporal Transformer models temporal dependencies across EEG tokens. The interaction rank of the Transformer is not constrained by d_s. Self-attention establishes explicit token-to-token interactions across all spatiotemporal positions, enabling the modeling of high-rank token relationships that cannot be captured by the state-space pathways in Mamba. Mamba cannot distinguish between $\mathbf{z}$ and $\mathbf{z} + \Delta\mathbf{z}$ when both are mapped to the same subspace representation. In contrast, under the self-attention mechanism, if a perturbation $\Delta\mathbf{z}$ modifies token-to-token relationships, then $(z + \Delta z)W_Q W_K^\top (z + \Delta z)^\top \neq z W_Q W_K^\top z^\top$. Because $\mathcal{T}(z + \Delta z) \neq \mathcal{T}(z)$, spatiotemporal Transformer can distinguish such inputs and recover high-order non-local dependencies and discriminative structures that are lost in the state compression process.

By periodically interleaving spatiotemporal Transformer layers, the model performs global dependency recomposition, effectively restoring high-rank interaction structures beyond the finite-state subspace and alleviating the intrinsic low-rank bottleneck of deep State Mamba stacks.

2.2.3 Mixture of Experts

To further enhance the model capacity for handling heterogeneous EEG distributions, we introduce a Mixture-of-Experts (MoE) module to enable adaptive subspace specialization. The proposed balanced MoE layer replaces the conventional feed-forward network by introducing a routing mechanism that dynamically dispatches spatiotemporal EEG tokens to specialized experts under capacity constraints. Each expert is parameterized as:

$$f_e(\mathbf{z}) = \mathbf{W}_2^{(e)} \sigma(\mathbf{W}_1^{(e)} \mathbf{z}) \tag{6}$$

where $e \in \{1, 2, 3..., E\}$ is the index of different experts. For each EEG tokens, a Top-k gating mechanism is applied to select the k most relevant experts according to the learned gating scores: $\mathcal{E}_i = \text{TopK}(\mathbf{g}_i)$. To ensure efficient and balanced computation, the MoE layer in BrainMamba adopts a capacity constraint to limit the number of tokens processed by each expert, preventing overloading and promoting balanced expert utilization. Through load-balanced gating and multi-expert aggregation, the model enables conditional computation over heterogeneous neural patterns,

significantly enhancing representation diversity and generalization across multi-subject, multi-task EEG distributions.

2.2.4 Overall Foundation Model Structure

The overall BrainMamba architecture hierarchically integrates State Mamba, MoE, and spatiotemporal Transformer modules, as illustrated in Figure 1. Specifically, six State Mamba layers are periodically followed by one MoE layer and one spatiotemporal Transformer layer. State Mamba efficiently models coupled spatiotemporal dynamics, MoE adaptively captures heterogeneous EEG patterns, and the spatiotemporal Transformer complements state-space modeling with global cross-channel and temporal interactions. Together, these modules provide a unified backbone for scalable and expressive EEG representation learning. The learned representations are used for self-supervised pretraining and subsequently adapted to downstream tasks through task-specific prediction heads.

2.3 Pretraining and Downstream Adaptation

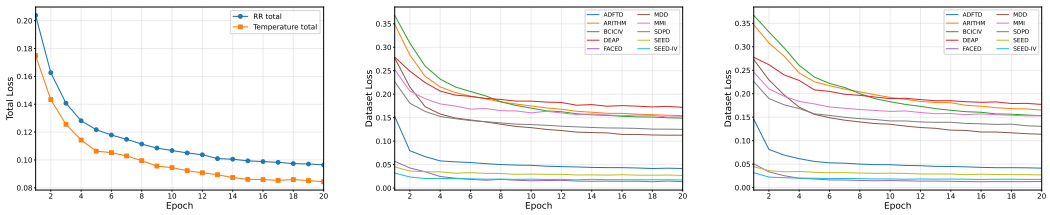
To fully exploit large-scale heterogeneous EEG data, we adopt a unified pretraining and downstream adaptation framework. In the pretraining stage, the goal of large scale self-supervised learning is to capture structural semantics that encompass intrinsic spatiotemporal dynamics and cross-channel interactions in diverse EEG signals. Due to the approximately stationary nature and strong local continuity of EEG signals, adopting the next-token prediction task Wang et al. (2024b) which is commonly used in language foundation models may lead to a degenerate solution. In such cases, the model can simply copy features from the previous time step without capturing meaningful frequency structures, cross-channel relationships, or cognitive states, resulting in training collapse. Moreover, EEG signals exhibit bidirectional dependencies, such as event-related potentials (ERP) and delayed synchronization across brain regions Quiroga et al. (2002), which cannot be effectively modeled through purely forward temporal prediction. Therefore, we adopt a masked reconstruction task Devlin et al. (2019), replacing the "**temporal extrapolation paradigm**" with the "**information completion paradigm**". This formulation encourages the model to learn structural semantics by reconstructing missing information, serving as a more challenging pretext task that promotes the modeling of intrinsic spatiotemporal dynamics and cross-channel interactions. The composite loss consists of the reconstruction loss, the expert balancing loss, and the regularization loss:

$$\mathcal{L}_{\text{pt}} = \underbrace{\frac{1}{|\mathcal{M}|}\sum_{(i,j)\in\mathcal{M}} \|\hat{\mathbf{z}}_{i,j} - \mathbf{z}_{i,j}\|_2^2}_{\text{Masked Reconstruction Loss}} + \underbrace{\lambda(E\sum_{e=1}^{E} P_e^2 + E\sum_{e=1}^{E} f_e P_e)}_{\text{MoE Load Balancing Loss}} + \underbrace{\|\Theta\|_2^2}_{\text{Weight Decay}}, \tag{7}$$

$$\text{where } P_e = \frac{1}{N}\sum_{i=1}^{N} g_{i,e},\ f_e = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[\arg\max_j g_{i,j} = e\right].$$

The first term is the masked reconstruction loss, which measures the discrepancy between the reconstructed tokens $\hat{\mathbf{z}}_{i,j}$ and the masked tokens $\mathbf{z}_{i,j}$ over the masked set $\mathcal{M}$; the second term is the load balancing regularization for the MoE layer, where P_e is the average routing probability assigned to specific expert e, f_e is the fraction of tokens actually dispatched to expert e, and $g_{i,e}$ is the routing probability calculated by the Top$k$ routing. The importance term $E\sum_{e=1}^{E} P_e^2$ penalizes highly skewed routing distributions, while the load term $E\sum_{e=1}^{E} f_e P_e$ constrains the consistency between the routing probabilities and the actual token assignments, preventing excessive concentration of tokens on a small subset of experts. Together, these two terms promote balanced expert utilization and effectively mitigate the expert collapse problem. The final term corresponds to standard weight decay, which regularizes the model parameters and improves generalization. The coefficient λ controls the trade-off between the pretraining task and the MoE regularization.

BrainMamba is jointly pretrained on multiple EEG datasets with heterogeneous distributions and channel configurations. To ensure balanced learning across diverse data sources, we adopt Round-Robin Sampling (RR) and Temperature Sampling (TS) to effectively handle heterogeneous EEG datasets, with implementation details provided in the Appendix and Figure 6.



(a) The overall Pretraining loss curve (RR vs TS)

(b) The pretraining loss curve for different datasets (RR)

(c) The pretraining loss curve for different datasets (TS)

Figure 3: Full pretraining loss curves under both round-robin (RR) and temperature sampling (TS) are presented. Specifically, both the overall loss and dataset-wise loss trends during pretraining are visualized, providing insights into the optimization dynamics and convergence behavior across different datasets.

Table 1: Full generalization evaluation results on SEED, SEED-IV, FACED, DEAP, and MDD datasets, comparing BrainMamba with state-of-the-art baselines.

Method	SEED, 3-class		SEED-IV, 4-class		FACED, 9-class		DEAP Arousal, 2-class		MDD, 2-class	
	Acc	Macro F1	Acc	Macro F1	Acc	Macro F1	Acc	Binary F1	Acc	Binary F1
BIOT NeurIPS'23	53.23 ± 1.32	52.97 ± 0.82	35.31 ± 0.40	34.92 ± 0.36	52.86 ± 2.67	53.71 ± 2.99	65.89 ± 4.06	66.19 ± 4.17	88.35 ± 0.96	89.25 ± 0.82
ContraWR JMIR	51.42 ± 0.54	52.35 ± 0.84	33.71 ± 0.92	32.10 ± 0.85	48.92 ± 2.39	48.79 ± 3.18	54.18 ± 4.79	55.80 ± 5.05	80.44 ± 0.82	81.35 ± 1.07
EEG2REP KDD'24	54.12 ± 0.92	54.92 ± 1.07	35.16 ± 0.37	35.24 ± 0.41	41.35 ± 5.21	43.28 ± 4.19	60.91 ± 5.52	61.82 ± 6.91	79.82 ± 1.45	80.43 ± 1.09
State Mamba AAAI'26	53.47 ± 1.15	52.66 ± 1.34	35.01 ± 0.35	36.12 ± 0.57	54.69 ± 5.34	53.89 ± 4.68	64.71 ± 4.67	63.99 ± 5.32	85.01 ± 1.85	86.32 ± 1.73
LaBraM-Base ICLR'24	55.43 ± 1.13	56.22 ± 1.08	36.71 ± 0.34	36.82 ± 0.30	53.19 ± 2.66	54.38 ± 2.47	64.75 ± 8.11	65.82 ± 6.87	87.95 ± 1.15	88.25 ± 1.06
CBraMod ICLR'25	56.17 ± 1.34	56.82 ± 1.60	36.92 ± 0.27	36.97 ± 0.36	56.53 ± 2.80	56.82 ± 2.54	64.28 ± 6.06	64.43 ± 4.35	91.27 ± 0.34	91.63 ± 0.42
EEGMamba NN'25	56.34 ± 1.28	56.77 ± 1.31	36.25 ± 0.48	35.87 ± 0.42	56.81 ± 2.82	57.42 ± 2.19	67.13 ± 4.16	68.04 ± 3.87	90.30 ± 0.81	90.14 ± 0.79
BrainMamba (RR)	58.41 ± 1.06	57.05 ± 0.97	38.61 ± 0.35	38.02 ± 0.41	65.89 ± 2.35	65.71 ± 2.18	**69.49 ± 4.32**	**70.06 ± 4.70**	91.75 ± 0.92	91.83 ± 0.73
BrainMamba (TS)	**58.94 ± 1.02**	**58.31 ± 1.14**	**39.14 ± 0.28**	**39.86 ± 0.34**	**67.38 ± 2.64**	**67.17 ± 2.82**	69.21 ± 4.18	69.70 ± 3.94	**93.15 ± 0.63**	**94.13 ± 0.51**

3 EXPERIMENTS

Innovatively, we propose a novel validation paradigm for EEG foundation models, which systematically evaluates their capabilities from multiple perspectives, including generalization, robustness, and transferability. The details of the evaluation paradigms are described in the Appendix.

3.1 FULL GENERALIZATION EVALUATION

The full generalization evaluation assesses the general-purpose capability of EEG foundation models under large-scale heterogeneous pretraining. Specifically, the model is pretrained on all available unlabeled EEG datasets with diverse distributions, channel configurations, and task domains, and subsequently evaluated across multiple downstream tasks. This setting aims to measure the model's ability to learn unified structural semantic representations and generalize across datasets and tasks without domain-specific constraints.

Figure 3 presents the full pretraining loss curves under both round-robin (RR) and temperature sampling (TS), including both the overall loss and dataset-wise loss trends. From Figure 3a, both RR and TS exhibit stable convergence behavior, while TS consistently achieves lower overall loss throughout training. This indicates that temperature sampling provides a more effective optimization trajectory by balancing the contribution of heterogeneous datasets. As shown in Fig. 3b and Fig. 3c, the convergence rates across different datasets exhibit noticeable imbalance. In particular, larger-scale or distributionally distinct datasets, such as SEED and SEED-IV, tend to dominate the training process, thereby influencing the overall optimization trajectory. Despite this, under the proposed collaborative pretraining framework over heterogeneous datasets, the reconstruction loss consistently decreases across all datasets. This indicates that the model is capable of effectively learning cross-dataset structural semantics within a unified representation space, while achieving stable and coherent joint optimization across multiple datasets.

The fine-tuning results of the full generalization evaluation on different datasets are shown in Tables 1 and 2. Overall, BrainMamba consistently outperforms existing self-supervised and foundation model baselines across most datasets. Compared with existing EEG foundation models (LaBraM Base, CBraMod, and EEGMamba), our method achieves superior performance in both Accuracy and

Table 2: Full generalization evaluation results on BCICIV, EEGMMIDB, ADFTD, UCSD, and Arithmetic datasets, comparing BrainMamba with state-of-the-art baselines.

Method	BCICIV, 4-class		EEGMMIDB, 4-class		ADFTD, 3-class		UCSD, 2-class		Arithmetic, 2-class	
	Acc	Macro F1	Acc	Macro F1	Acc	Macro F1	Acc	Binary F1	Acc	Binary F1
BIOT NeurIPS'23	48.32 ± 0.54	48.49 ± 0.77	58.43 ± 1.52	58.77 ± 1.54	64.07 ± 0.77	63.12 ± 0.81	53.41 ± 2.04	58.38 ± 1.99	64.39 ± 1.81	65.71 ± 1.46
ContraWR JMIR	44.27 ± 1.23	45.76 ± 1.19	55.44 ± 0.97	54.98 ± 1.05	59.83 ± 1.34	61.08 ± 1.02	51.72 ± 1.39	54.81 ± 2.00	62.74 ± 1.40	60.38 ± 1.25
EEG2REP KDD'24	45.37 ± 1.42	45.81 ± 1.25	56.72 ± 1.74	54.23 ± 1.38	60.47 ± 0.85	62.10 ± 0.93	51.39 ± 1.65	53.77 ± 2.04	63.41 ± 1.03	63.20 ± 1.04
State Mamba AAAI'26	48.30 ± 1.29	49.11 ± 1.06	55.27 ± 1.41	56.89 ± 1.70	62.20 ± 1.12	60.85 ± 1.06	52.18 ± 1.71	56.74 ± 2.33	63.86 ± 1.37	60.82 ± 1.55
LaBraM-Base ICLR'24	48.14 ± 0.95	47.52 ± 1.14	58.74 ± 1.49	58.62 ± 1.50	62.86 ± 1.04	64.57 ± 1.35	52.70 ± 1.43	54.87 ± 2.01	66.53 ± 1.80	66.89 ± 1.95
CBraMod ICLR'25	51.03 ± 0.85	51.73 ± 1.14	**63.85 ± 1.36**	**63.82 ± 1.79**	65.38 ± 0.99	65.42 ± 0.93	54.10 ± 2.06	57.83 ± 2.31	70.14 ± 1.17	69.47 ± 1.99
EEGMamba NN'25	50.73 ± 0.98	51.30 ± 1.22	60.73 ± 1.51	58.46 ± 1.62	64.51 ± 0.83	64.97 ± 1.15	53.11 ± 2.74	55.72 ± 2.19	68.41 ± 1.42	67.39 ± 1.25
BrainMamba (RR)	52.24 ± 0.93	54.60 ± 0.82	61.34 ± 1.16	60.57 ± 1.32	**67.52 ± 0.84**	64.77 ± 1.06	54.35 ± 1.49	58.43 ± 2.31	67.52 ± 1.37	64.77 ± 1.98
BrainMamba (TS)	**53.82 ± 0.93**	**54.61 ± 0.78**	61.39 ± 1.42	60.43 ± 1.29	67.36 ± 1.05	**65.84 ± 1.27**	**57.18 ± 2.17**	**62.04 ± 1.98**	**73.96 ± 1.92**	**70.62 ± 2.05**

Macro-F1 (Binary F1 for binary classification tasks), demonstrating strong cross-dataset generalization capability. Notably, the improvements are more pronounced on complex multi-class datasets, such as SEED-IV and FACED, indicating that BrainMamba is capable of effectively capturing high-order spatiotemporal dependencies and cross-channel interactions. Besides, BrainMamba achieves relatively lower performance on the EEGMMIDB dataset. This can be attributed to several factors. First, motor imagery EEG signals typically exhibit a lower signal-to-noise ratio and larger inter-subject variability, making representation transfer more challenging. Second, discriminative patterns in EEGMMIDB are primarily concentrated in specific frequency bands (e.g., μ and β rhythms), which may not fully align with the global spatiotemporal representations learned during large-scale pretraining. These factors collectively lead to relatively poorer performance of BrainMamba on this dataset.

Besides, we compare two multi-dataset pretraining strategies, round-robin sampling (RR) and temperature-controlled sampling (TS). Overall, TS performs slightly better on most datasets, especially on larger or more complex datasets such as FACED and MDD, likely because it better balances dataset scale and sampling diversity. However, the gap between RR and TS remains small on several datasets, suggesting that both strategies provide sufficient representation coverage once the model is adequately trained. For relatively simple tasks, such as binary classification, the choice of sampling strategy has even less impact. Interestingly, RR achieves comparable or slightly better results on some smaller datasets, indicating that uniform sampling may offer more stable optimization under data imbalance.

4 Conclusion

In this paper, we propose BrainMamba, a scalable EEG foundation model that integrates structured state-space modeling, spatiotemporal Transformers, and MoE blocks for large-scale EEG multi-dataset pretraining. By explicitly modeling coupled spatiotemporal dynamics and cross-channel interactions, BrainMamba captures structural semantics in EEG signals and demonstrates strong performance across diverse downstream tasks. Extensive experiments demonstrate superior or competitive performance across a wide range of EEG applications, including emotion recognition, motor imagery, and clinical diagnosis, while maintaining robust generalization under heterogeneous settings. Furthermore, our analyses reveal important insights into multi-dataset EEG pretraining, such as the trade-offs between representation generality and task-specific specialization, as well as the impact of dataset scale, sampling strategies, and channel configurations on transfer performance. Despite these promising results, several limitations remain. First, although BrainMamba demonstrates strong generalization ability, its performance can still degrade under significant domain shifts, such as cross-task transfer (e.g., from emotion to motor imagery) or severe channel configuration mismatches. Second, the current framework relies on curated multi-dataset pretraining, and the scale and diversity of available EEG data are still limited compared to other modalities such as language and vision. Third, the computational cost of large-scale pretraining and the complexity of MoE-based architectures may pose challenges for deployment in resource-constrained scenarios. In future work, we plan to explore more scalable pretraining strategies, including larger and more diverse EEG corpora, more efficient model architectures, and improved sampling and alignment mechanisms to further enhance cross-domain generalization. We also aim to investigate unified multimodal modeling by integrating EEG with other physiological signals, which may provide a more comprehensive understanding of brain dynamics and improve real-world applicability.

AI USE STATEMENT

In this work, we used generative AI tools to assist with research execution, primarily for code implementation, debugging, and experimental workflow development. We also used generative AI tools to polish author-written text for grammar, clarity, and readability. Generative AI tools were not used to formulate the research ideas, hypotheses, methodology, or experimental design, nor were they used to generate synthetic datasets or formulate or prove mathematical claims. All AI-assisted code was reviewed, tested, and validated by the authors, and all reported results were independently verified against the original experimental outputs. All AI-assisted text was carefully checked and revised by the authors. The authors take full responsibility for the final content of this work, including all text, claims, experimental results, code, and artifacts produced with the assistance of generative AI.

REFERENCES

Khald Ali I Aboalayon, Miad Faezipour, Wafaa S Almuhammadi, and Saeid Moslehpour. Sleep stage classification using eeg signal analysis: a comprehensive survey and new investigation. *Entropy*, 18(9):272, 2016.

Salma Alhagry, Aly Aly Fahmy, and Reda A El-Khoribi. Emotion recognition based on eeg using lstm recurrent neural network. *International Journal of Advanced Computer Science and Applications*, 8(10), 2017.

Mashail Alsolamy and Anas Fattouh. Emotion estimation from eeg signals during listening to quran using psd features. In *2016 7th International Conference on computer science and information technology (CSIT)*, pp. 1–5. IEEE, 2016.

Claudio Babiloni, Roberta Lizio, Nicola Marzano, Paolo Capotosto, Andrea Soricelli, Antonio Ivano Triggiani, Susanna Cordone, Loreto Gesualdo, and Claudio Del Percio. Brain neural synchronization and functional coupling in alzheimer's disease as revealed by resting state eeg rhythms. *International Journal of Psychophysiology*, 103:88–102, 2016.

Clemens Brunner, Robert Leeb, Gernot Müller-Putz, Alois Schlögl, and Gert Pfurtscheller. Bci competition 2008–graz data set a. *Institute for knowledge discovery (laboratory of brain-computer interfaces), Graz University of Technology*, 16(1-6):1, 2008.

Michael Bruyns-Haylett, Jingjing Luo, Aneurin J Kennerley, Sam Harris, Luke Boorman, Elizabeth Milne, Nicolas Vautrelle, Yurie Hayashi, Benjamin J Whalley, Myles Jones, et al. The neurogenesis of p1 and n1: A concurrent eeg/lfp study. *Neuroimage*, 146:575–588, 2017.

Chi-Sheng Chen, Ying-Jung Chen, and Aidan Hung-Wen Tsai. Large cognition model: Towards pretrained eeg foundation model. *arXiv preprint arXiv:2502.17464*, 2025a.

Jingjing Chen, Xiaobin Wang, Chen Huang, Xin Hu, Xinke Shen, and Dan Zhang. A large finer-grained affective computing eeg dataset. *Scientific Data*, 10(1):740, 2023.

Junru Chen, Yang Yang, Tao Yu, Yingying Fan, Xiaolong Mo, and Carl Yang. Brainnet: Epileptic wave detection from seeg with hierarchical graph diffusion learning. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 2741–2751, 2022a.

Wensheng Chen, Cong Yu, Zhenhua Zhao, Nan Zheng, Han Li, and Yurong Li. Self-supervised eeg denoising via dual-branch consistency learning with masked reconstruction. *Knowledge-Based Systems*, pp. 114703, 2025b.

Xun Chen, Chang Li, Aiping Liu, Martin J McKeown, Ruobing Qian, and Z Jane Wang. Toward open-world electroencephalogram decoding via deep learning: A comprehensive survey. *IEEE Signal Processing Magazine*, 39(2):117–134, 2022b.

Yuqi Chen, Kan Ren, Kaitao Song, Yansen Wang, Yifan Wang, Dongsheng Li, and Lili Qiu. Eegformer: Towards transferable and interpretable large-scale eeg foundation model. *arXiv preprint arXiv:2401.10278*, 2024.

Hsiang-Yun Sherry Chien, Hanlin Goh, Christopher M Sandino, and Joseph Y Cheng. Maeeg: Masked auto-encoder for eeg representation learning. *arXiv preprint arXiv:2211.02625*, 2022.

Guanghai Dai, Jun Zhou, Jiahui Huang, and Ning Wang. Hs-cnn: a cnn with hybrid convolution scale for eeg motor imagery classification. *Journal of neural engineering*, 17(1):016025, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Arne Dietrich and Riam Kanso. A review of eeg, erp, and neuroimaging studies of creativity and insight. *Psychological bulletin*, 136(5):822, 2010.

Hauke Dose, Jakob S Møller, Helle K Iversen, and Sadasivan Puthusserypady. An end-to-end deep learning approach to mi-eeg signal classification for bcis. *Expert Systems with Applications*, 114: 532–542, 2018.

Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. Differential entropy feature for eeg-based emotion classification. In *2013 6th international IEEE/EMBS conference on neural engineering (NER)*, pp. 81–84. IEEE, 2013.

Jobi S George, Jon Strunk, Rachel Mak-McCully, Melissa Houser, Howard Poizner, and Adam R Aron. Dopaminergic therapy in parkinson's disease decreases cortical beta band coherence in the resting state and increases cortical beta band power during executive control. *NeuroImage: Clinical*, 3:261–270, 2013.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. *Advances in neural information processing systems*, 35:35971–35983, 2022.

Xiaotong Gu, Zehong Cao, Alireza Jolfaei, Peng Xu, Dongrui Wu, Tzyy-Ping Jung, and Chin-Teng Lin. Eeg-based brain-computer interfaces (bcis): A survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications. *IEEE/ACM transactions on computational biology and bioinformatics*, 18(5):1645–1666, 2021.

Yiyu Gui, MingZhi Chen, Yuqi Su, Guibo Luo, and Yuchao Yang. Eegmamba: Bidirectional state space model with mixture of experts for eeg multi-task classification. *arXiv preprint arXiv:2407.20254*, 2024.

Alice F Jackson and Donald J Bolger. The neurophysiological bases of eeg and eeg measurement: A review for the rest of us. *Psychophysiology*, 51(11):1061–1071, 2014.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

NN Kulkarni and VK Bairagi. Extracting salient features for eeg-based diagnosis of alzheimer's disease using support vector machine classifier. *IETE Journal of Research*, 63(1):11–22, 2017.

Junhong Lai, Jiyu Wei, Lin Yao, and Yueming Wang. A simple review of eeg foundation models: Datasets, advancements and future perspectives. *arXiv preprint arXiv:2504.20069*, 2025.

Wenbing Li, Hang Zhou, Junqing Yu, Zikai Song, and Wei Yang. Coupled mamba: Enhanced multimodal fusion with coupled state space model. *Advances in Neural Information Processing Systems*, 37:59808–59832, 2024.

Jianchao Lu, Yuzhe Tian, Yang Zhang, Quan Z. Sheng, and Xi Zheng. Lgl-bci: A motor-imagery-based brain-computer interface with geometric learning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 9(1), March 2025. doi: 10.1145/3699732. URL https://doi.org/10.1145/3699732.

Shenghuan Miao and Ling Chen. Wonderwall: A virtual-to-real foundation model for imu-based har. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 10(1), March 2026. doi: 10.1145/3789688. URL https://doi.org/10.1145/3789688.

Andreas Miltiadous, Katerina D. Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Nikolaos Grigoriadis, Dimitrios G. Tsalikakis, Pantelis Angelidis, Markos G. Tsipouras, Evripidis Glavas, Nikolaos Giannakeas, and Alexandros T. Tzallas. "a dataset of eeg recordings from: Alzheimer's disease, frontotemporal dementia and healthy subjects", 2024.

Navid Mohammadi Foumani, Geoffrey Mackellar, Soheila Ghane, Saad Irtza, Nam Nguyen, and Mahsa Salehi. Eeg2rep: enhancing self-supervised eeg representation through informative masked inputs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5544–5555, 2024.

Mostafa Neo Mohsenvand, Mohammad Rasool Izadi, and Pattie Maes. Contrastive representation learning for electroencephalogram classification. In *Machine learning for health*, pp. 238–253. PMLR, 2020.

Wajid Mumtaz, Likun Xia, Syed Saad Azhar Ali, Mohd Azhar Mohd Yasin, Muhammad Hussain, and Aamir Saeed Malik. Electroencephalogram (eeg)-based computer-aided technique to diagnose major depressive disorder (mdd). *Biomedical Signal Processing and Control*, 31:108–115, 2017.

Jennifer J Newson and Tara C Thiagarajan. Eeg frequency bands in psychiatric disorders: a review of resting state studies. *Frontiers in human neuroscience*, 12:521, 2019.

Minghui Qiu, Cekai Weng, Mingming Fan, and Kaishun Wu. Towards customizable foundation models for human activity recognition with wearable devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 9(3), September 2025. doi: 10.1145/3749479. URL https://doi.org/10.1145/3749479.

Haohao Qu, Liangbo Ning, Rui An, Wenqi Fan, Tyler Derr, Hui Liu, Xin Xu, and Qing Li. A survey of mamba. *arXiv preprint arXiv:2408.01129*, 2024.

R Quian Quiroga, Thomas Kreuz, and Peter Grassberger. Event synchronization: a simple and fast method to measure synchronicity and time delay patterns. *Physical review E*, 66(4):041904, 2002.

Mohammad H Rafiei, Lynne V Gauthier, Hojjat Adeli, and Daniel Takabi. Self-supervised learning for electroencephalography. *IEEE Transactions on Neural Networks and Learning Systems*, 35 (2):1457–1471, 2022.

Mithun Saha, Maxwell A. Xu, Wanting Mao, Sameer Neupane, James M. Rehg, and Santosh Kumar. Pulse-ppg: An open-source field-trained ppg foundation model for wearable applications across lab and field settings. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 9(3), September 2025. doi: 10.1145/3749494. URL https://doi.org/10.1145/3749494.

Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Transactions on biomedical engineering*, 51(6):1034–1043, 2004.

Afshin Shoeibi, Delaram Sadeghi, Parisa Moridian, Navid Ghassemi, Jonathan Heras, Roohallah Alizadehsani, Ali Khadem, Yinan Kong, Saeid Nahavandi, Yu-Dong Zhang, et al. Automatic diagnosis of schizophrenia in eeg signals using cnn-lstm models. *Frontiers in neuroinformatics*, 15:777977, 2021.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022.

Nicole C Swann, Coralie De Hemptinne, Adam R Aron, Jill L Ostrem, Robert T Knight, and Philip A Starr. Elevated synchrony in parkinson disease detected with electroencephalography. *Annals of neurology*, 78(5):742–750, 2015.

Wei Tao, Chang Li, Rencheng Song, Juan Cheng, Yu Liu, Feng Wan, and Xun Chen. Eeg-based emotion recognition via channel-wise attention and self attention. *IEEE Transactions on Affective Computing*, 14(1):382–393, 2020.

Arjon Turnip and Keum-Shik Hong. Classifying mental activities from eeg-p300 signals using adaptive neural network. *Int. J. Innov. Comp. Inf. Control*, 8(7):5839–5850, 2012.

Ramesh Babu Vallabhaneni, Pankaj Sharma, Vinit Kumar, Vyom Kulshreshtha, Koya Jeevan Reddy, S Selva Kumar, V Sandeep Kumar, and Surendra Kumar Bitra. Deep learning algorithms in eeg signal decoding application: a review. *IEEE Access*, 9:125778–125786, 2021.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. *arXiv preprint arXiv:2412.07236*, 2024a.

Xiaoshuang Wang, Guanyu Wang, Tingting Wu, Ying Wang, Tommi Kärkkäinen, and Fengyu Cong. Eeg-based seizure onset detection of frontal and temporal lobe epilepsies using 1dcnn. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2025.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b.

Weining Weng, Yang Gu, Shuai Guo, Yuan Ma, Yuchen Liu, Yuan Jin, Fulong Xiao, and Yiqiang Chen. Wafformer: Wave-frequency cross-prediction and alignment for knowledge-enhanced sleep representation. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2673–2679. IEEE, 2024.

Weining Weng, Yang Gu, Shuai Guo, Yuan Ma, Zhaohua Yang, Yuchen Liu, and Yiqiang Chen. Self-supervised learning for electroencephalogram: A systematic survey. *ACM Computing Surveys*, 57(12):1–38, 2025.

Weining Weng, Yang Gu, Yuan Ma, Yuchen Liu, Yingwei Zhang, and Yiqiang Chen. State mamba: Spatiotemporal eeg state-space model with dynamic brain alignment for cross-subject representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17850–17858, 2026.

Jin Xie, Jie Zhang, Jiayao Sun, Zheng Ma, Liuni Qin, Guanglin Li, Huihui Zhou, and Yang Zhan. A transformer-based approach combining deep learning network and spatial-temporal information for raw eeg classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:2126–2136, 2022.

Wei Xiong, Junming Lin, Jiangtong Li, Jie Li, and Changjun Jiang. Alfee: Adaptive large foundation model for eeg representation. *arXiv preprint arXiv:2505.06291*, 2025.

Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36:78240–78260, 2023a.

Chaoqi Yang, Cao Xiao, M Brandon Westover, and Jimeng Sun. Self-supervised electroencephalogram representation learning for automatic sleep staging: model development and evaluation study. *JMIR AI*, 2(1):e46769, 2023b.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2):1–124, 2023.

Wei-Long Zheng and Bao-Liang Lu. 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, 2015.

Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE transactions on cybernetics*, 49 (3):1110–1122, 2018.

Mengni Zhou, Cheng Tian, Rui Cao, Bin Wang, Yan Niu, Ting Hu, Hao Guo, and Jie Xiang. Epileptic seizure detection based on eeg signals and cnn. *Frontiers in neuroinformatics*, 12:95, 2018.

Xinliang Zhou, Yuzhe Han, Zhisheng Chen, Chenyu Liu, Yi Ding, Ziyu Jia, and Yang Liu. Bit-mamsleep: Bidirectional temporal mamba for eeg sleep staging. *arXiv preprint arXiv:2411.01589*, 2024.

Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. Csbrain: A cross-scale spatiotemporal brain foundation model for eeg decoding. *arXiv preprint arXiv:2506.23075*, 2025.

Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1):14, 2019.

A Related Works

A.1 EEG-Based Tasks

With the development of neuroscience, EEG analysis has evolved from early studies of resting-state rhythms Babiloni et al. (2016) to a key technique for decoding a wide range of cognitive, perceptual, and behavioral states Chen et al. (2022b). Extensive studies in neuroscience have demonstrated that EEG signals contain a variety of critical time–frequency features, including oscillatory rhythms across frequency bands (e.g., alpha, beta, and theta waves) Newson & Thiagarajan (2019), power spectral density (PSD) Alsolamy & Fattouh (2016), differential entropy (DE) Duan et al. (2013), and representative event-related potentials Dietrich & Kanso (2010). These features exhibit stable and transferable relationships with various BCI tasks, including motor imagery, emotion recognition, cognitive load estimation, sleep staging, and epilepsy detection, highlighting the broad applicability of EEG in intelligent sensing and healthcare. For instance, Duan et al. (2013) demonstrated that differential entropy features and their hemispheric asymmetry in EEG signals serve as effective representations for emotion states, significantly improving the performance of emotion recognition; Aboalayon et al. (2016) provided a comprehensive review of EEG-based sleep analysis, showing that both time-domain and frequency-domain features play a critical role in improving sleep stage classification performance. Kulkarni & Bairagi (2017) demonstrated that spectral, wavelet-based, and complexity-derived EEG features capture discriminative patterns that enable effective classification of Alzheimer's disease. These handcrafted features have significantly driven the development of EEG signal analysis and decoding methods.

Besides, the development of deep learning methods has enabled end-to-end modeling for EEG signal decoding. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and Transformer-based architectures have been extensively used for EEG signal modeling, jointly capturing spatial dependencies across channels and temporal dynamics, leading to improved accuracy and robustness in various downstream EEG tasks Vallabhaneni et al. (2021). Specifically, convolutional neural networks model spatial dependencies across channels Shoeibi et al. (2021) and local temporal patterns Wang et al. (2025), whereas recurrent neural networks and attention mechanisms Tao et al. (2020) capture long-range temporal dependencies and global contextual information

at the sequence level. Compared to handcrafted feature-based approaches, end-to-end models autonomously learn hierarchical and discriminative representations from raw EEG signals Dose et al. (2018), significantly reducing reliance on prior feature engineering and enabling more flexible adaptation and stronger generalization across diverse tasks. However, due to substantial distributional discrepancies in EEG signals across tasks and subjects, these variations are further exacerbated in real-world applications by device heterogeneity, varying acquisition conditions, and inconsistent channel configurations, which significantly increases the difficulty of generalizable modeling. Therefore, developing a unified EEG modeling framework with strong generalization capability and robustness remains a central challenge Weng et al. (2025).

A.2 EEG-Based Foundation Models

Inspired by the success of large-scale pretrained language models (LLMs) and vision–language models (VLMs), as well as recent advances in foundation models for other physiological signals such as IMU Qiu et al. (2025); Miao & Chen (2026) and PPG Saha et al. (2025), developing foundation models for generalizable and robust representation learning in brain signals has emerged as a central research direction Lai et al. (2025). From early self-supervised pioneers to recent general-purpose large-scale models, the evolution of EEG foundation models clearly reflects a continuous improvement in both generalization capability and representational power Rafiei et al. (2022). In early self-supervised frameworks, pretext tasks such as masked signal reconstruction Chien et al. (2022); Chen et al. (2025b), temporal prediction Chen et al. (2022a), and contrastive learning Weng et al. (2024); Mohsenvand et al. (2020) were employed to learn generalizable representations from unlabeled EEG data. Kostas et al. proposed BENDR Kostas et al. (2021), inspired by self-supervised learning in speech processing, which leverages masked prediction pretraining on large-scale unlabeled EEG data, enabling effective generalization across heterogeneous hardware, subjects, and tasks. Yang et al. introduced ContraWR Yang et al. (2023b), a self-supervised framework that combines wavelet-based multi-scale time–frequency analysis with contrastive learning, capturing global statistical characteristics of EEG signals and achieving strong performance and robustness in large-scale real-world sleep staging tasks. Furthermore, Yang et al. proposed BIOT Yang et al. (2023a), a physiological signal architecture based on channel-wise tokenization that maps heterogeneous multi-channel signals into a unified token sequence. This design effectively addresses challenges such as mismatched channel configurations, varying sampling rates, and missing data, thereby enabling large-scale in-the-wild pretraining and robust cross-task transfer. These self-supervised approaches provide a strong foundation for large-scale EEG representation learning, significantly improving cross-task and cross-dataset generalization and advancing the development of robust EEG foundation models.

Based on the self-supervised method, recent studies have explored large-scale pretraining paradigms and unified architectures to further advance EEG foundation modeling. Jiang et al. (2024) proposed Labram, the first truly "large-scale brain model," which performs large-scale pretraining via EEG channel patching and vector-quantized neural spectral prediction, demonstrating strong generalization across subjects and tasks. Wang et al. introduced the Cbramod framework Wang et al. (2024a), which adopts an innovative Criss-Cross Transformer architecture with two parallel attention mechanisms to model the spatiotemporal dependencies of EEG signals, achieving state-of-the-art performance across up to ten downstream BCI tasks spanning twelve public datasets. Ma et al. proposed the CSBrain framework Zhou et al. (2025), which focuses on disentangled spatiotemporal representation learning for EEG signals. By constructing separate temporal and spatial encoders and incorporating a cross-view contrastive learning mechanism, the model is encouraged to capture consistent brain-state representations under diverse perturbations, effectively enhancing its ability to model non-stationary EEG dynamics. Xiong et al. proposed ALFEE Xiong et al. (2025), an adaptive EEG foundation model that employs a hybrid attention to jointly capture channel-wise aggregation and temporal dependencies. It introduces multi-dimensional pretraining objectives, including time–frequency dual-domain reconstruction, temporal forecasting, and spectral alignment, together with an adaptive channel encoder to address complex spatiotemporal heterogeneity in large-scale clinical data. While these approaches have achieved promising progress, they remain limited in their ability to fully unify spatiotemporal modeling and effectively handle the substantial distributional discrepancies inherent in real-world EEG scenarios.

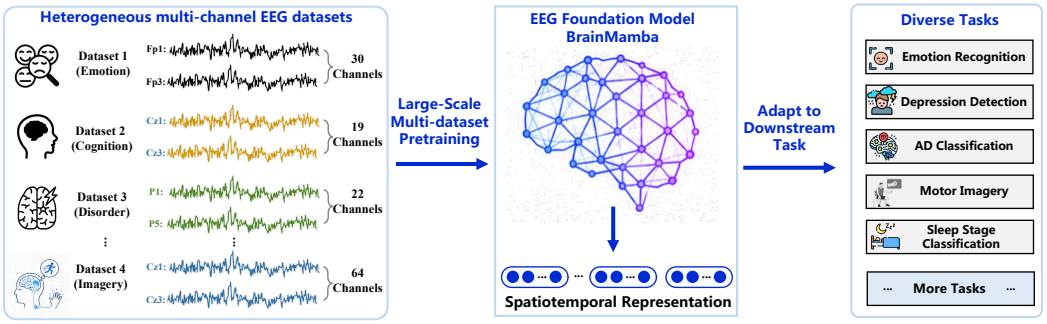


Figure 4: Overview of the BrainMamba EEG foundation model. Heterogeneous multi-channel EEG datasets from diverse sources are jointly leveraged through large-scale multi-dataset pretraining to learn unified spatiotemporal representations. The pretrained model can be effectively adapted to a wide range of downstream tasks, demonstrating strong generalization across diverse scenarios.

A.3 Preliminary

A.3.1 EEG Foundation Model Paradigm

A general modeling paradigm for EEG foundation models consists of input embedding, backbone architecture, pretraining strategies, and downstream task adaptation, enabling unified representation learning for diverse EEG tasks and scenarios.

(1) **Multi-Channel Signal Embedding**: Given a multi-channel EEG signal $\mathbf{X} \in \mathbb{R}^{C\times T}$, where C denotes the number of channels and T the temporal length. The signal is naturally segmented along the temporal dimension into a sequence of non-overlapping local segments. Each segment is denoted as $\mathbf{X}_i \in \mathbb{R}^{C\times L}$, where L is the segment length. To enable unified modeling, each segment is mapped into a latent representation space via an encoding function $\phi(\cdot)$:

$$\mathbf{z_i} = \phi(\mathbf{X_i}), \mathbf{z_i} \in \mathbb{R}^d \tag{8}$$

where d denotes the embedding dimension. The EEG sample is thus decomposed into a sequence of tokens $\{\mathbf{z}i\}_{i=1}^N$.

(2) **EEG Representation Backbone**. The tokenized EEG sequence is then processed by the deep backbone model $f_B(\cdot)$ to capture spatiotemporal dependencies and signal semantics:

$$\{\mathbf{s_i}\}_{i=1}^N = f_B(\{\mathbf{z}i\}_{i=1}^N). \tag{9}$$

The resulting representations $\{\mathbf{s}_i\}_{i=1}^N$ capture high-level spatiotemporal semantic features for downstream tasks.

(3) **Pretraining Paradigm and Downstream Adaptation**. In EEG foundation models, large-scale unlabeled pretraining provides essential semantic understanding of the signals, while downstream adaptation maps these general representations to specific tasks. During pretraining, unlabeled samples provide self-supervision through pseudo-labels, guiding the model to learn intrinsic structures and discriminative semantics of EEG signals. During downstream adaptation, the learned token-level representations are aggregated to form a global representation, which captures holistic spatiotemporal semantics and patterns for downstream tasks. This unified paradigm enables the model to transfer knowledge learned from large-scale unlabeled data to diverse downstream applications.

A.3.2 Mamba-based State Space Models

State-space models (SSMs) is a classical framework in control theory for describing dynamical systems Gu et al. (2022). They characterize the evolution of latent states via state and observation equations, where difference equations govern the interactions among inputs, states, and outputs. Owing to their ability to model continuous-time dynamics and long-range dependencies with a principled and efficient formulation, SSMs have recently emerged as a powerful approach for sequence modeling, including language and time-series analysis Qu et al. (2024).

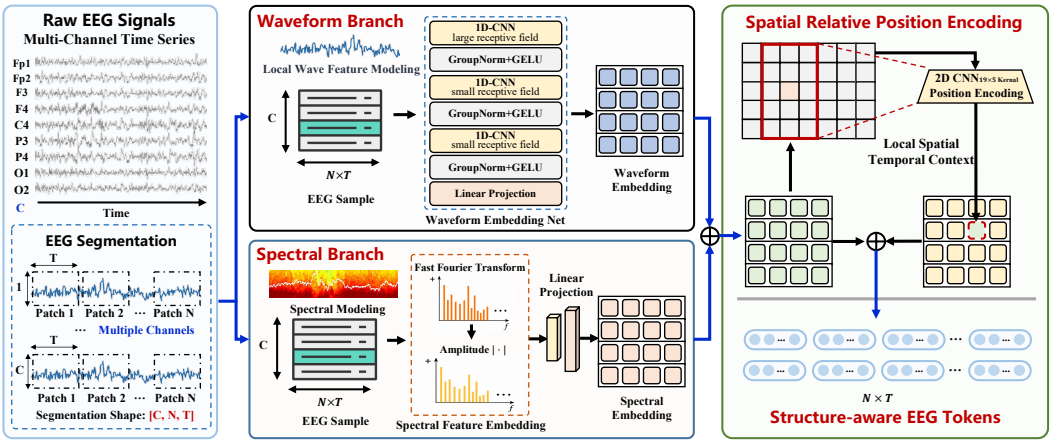


Figure 5: Structure-aware time-frequency embedding for EEG foundation modeling. Given raw multi-channel EEG signals, we construct unified token representations through a dual-branch encoding framework. The waveform branch captures temporal dynamics via multi-scale convolutional projections, while the spectral branch encodes frequency-domain information using FFT-based features. The spatial relative positional encoding module is incorporated to model local spatial–temporal relationships.

The continuous state-space model describes the evolution of a latent state $\mathbf{h}(t)$ driven by an input signal $\mathbf{x}(t)$:

$$\mathbf{h}'(t) = \mathbf{A}\mathbf{h}(t) + \mathbf{B}\mathbf{x}(t), \mathbf{y}(t) = \mathbf{C}\mathbf{h}(t) \tag{10}$$

where $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are learnable parameter matrices controlling the state transition dynamics, input influence, and output projection, respectively. This formulation models the continuous evolution of latent states via a linear dynamical system.

To make the continuous SSM applicable to discrete modeling, the state transition equation is discretized by solving the general solution of Eq. (10):

$$h(t) = e^{A(t-t_0)}h(t_0) + \int_{t_0}^{t} e^{A(t-\tau)}Bu(\tau)d\tau \tag{11}$$

let $t_0 = t_k$, $t = t_{k+1}$, and $t_{k+1} - t_k = \Delta$, the equation can be written as:

$$h_{k+1} = e^{A\Delta}h_k + \int_{t_k}^{t_{k+1}} e^{A(t_{k+1}-\tau)}Bu(\tau)d\tau \tag{12}$$

Under the zero-order hold assumption, where the input is constant within each interval, i.e., $u(\tau) = u_k$ for $\tau \in [t_k, t_{k+1}]$, the discrete-time formulation becomes:

$$h_{k+1} = e^{A\Delta}h_k + \int_{t_k}^{t_{k+1}} e^{A(t_{k+1}-\tau)}Bd\tau\, u_k \tag{13}$$

Therefore, in the discretized form, the state transition and input projection matrices are given by:

$$\bar{\mathbf{A}} = e^{\mathbf{A}\boldsymbol{\Delta}} = \exp(\mathbf{A}\boldsymbol{\Delta}), \bar{\mathbf{B}} = \int_0^{\Delta} e^{\mathbf{A}\tau}d\tau\mathbf{B} = (\mathbf{A}\boldsymbol{\Delta})^{-1}(exp(\mathbf{A}\boldsymbol{\Delta}) - \mathbf{I})\mathbf{B}, \tag{14}$$

where, $\bar{\mathbf{A}}$ governs the transition and evolution of the latent state across time, while $\bar{\mathbf{B}}$ controls the influence of the current input. In practice, $\bar{\mathbf{B}}$ is often approximated as being proportional to the step size: $\bar{\mathbf{B}} \approx \Delta\mathbf{B}$. In the Mamba architecture, the parameters $\mathbf{B}$, Δ, and $\mathbf{C}$ are parameterized as input-dependent functions via linear projections, allowing the state-space model to dynamically adapt its transition and emission dynamics conditioned on the input. This input-aware modulation significantly enhances the model's capacity to capture complex and heterogeneous sequence patterns.

A.4 IMPLEMENTATION

In this section, we provide detailed implementation settings for the experimental evaluation of BrainMamba, including model configurations, dataset descriptions, and baseline methods.

A.4.1 DATASET DESCRIPTION

For both pretraining and downstream tasks, we utilize 10 publicly available multi-channel EEG benchmark datasets, spanning diverse application domains such as emotion recognition, motor imagery, depression, neurological disorder detection, and cognitive load estimation. The detailed descriptions are provided as follows:

Emotion EEG datasets: Four benchmark EEG datasets for emotion recognition are included.

SEED Zheng & Lu (2015): The SEED dataset is a widely used EEG benchmark for emotion recognition, consisting of recordings from 15 subjects across three sessions. EEG signals are collected using a 62-channel system at a sampling rate of 200 Hz while subjects watch emotion-eliciting film clips, with a total recording duration exceeding 152,730 seconds. The data are segmented into 30,375 samples using a 5-second window.

SEED-IV Zheng et al. (2018): The SEED-IV dataset is an extension of the SEED dataset for EEG-based emotion recognition, consisting of recordings from 15 subjects across three sessions. EEG signals are collected using a 62-channel system at a sampling rate of 200 Hz while subjects watch emotion-eliciting film clips, with a total recording duration of 732,750 seconds. The data are segmented into 30,045 samples using a 5-second window. The dataset includes four emotional categories: happy, sad, fear, and neutral, and the multi-session design enables evaluation of cross-session and cross-subject generalization.

FACED Chen et al. (2023): The FACED dataset is a large-scale EEG benchmark for emotion recognition, consisting of recordings from 123 subjects exposed to 9 emotional stimuli (joy, tenderness, inspiration, amusement, anger, disgust, fear, sadness, neutral). EEG signals are recorded using a 32-channel system at a sampling rate of 256 Hz, with a total recording duration of 103,320 seconds. The data are segmented into 13,776 samples using a 10-second window. The dataset is characterized by large-scale subject diversity and rich emotional annotations, making it suitable for evaluating cross-subject generalization and large-scale representation learning.

DEAP Koelstra et al. (2011): The DEAP dataset is a widely used EEG benchmark for emotion recognition, consisting of recordings from 32 subjects. EEG signals are recorded using a 32-channel system at a sampling rate of 128 Hz, with a total recording duration of 80,640 seconds. The data are segmented into 3,840 samples using a 10-second window. The dataset provides binary classification settings for both arousal and valence dimensions.

Motor imagery EEG datasets: two benchmark EEG datasets for motor imagery task are included.

EEGMMIDB (MMI) Goldberger et al. (2000); Schalk et al. (2004):The EEG Motor Movement/Imagery Dataset (EEGMMIDB) is a widely used benchmark for motor imagery classification, consisting of recordings from 109 subjects. EEG signals are recorded using a 64-channel system at a sampling rate of 160 Hz, with a total recording duration of 39,348 seconds. The data are segmented into 9,837 samples using a 4-second window. The dataset is formulated as a four-class motor imagery classification task.

BCICIV-2a Brunner et al. (2008): The BCI Competition IV-2a dataset is a widely used benchmark for motor imagery classification, consisting of recordings from 9 subjects. EEG signals are recorded using a 22-channel system at a sampling rate of 250 Hz, with a total recording duration of 21,000 seconds. The data are segmented into 5,088 samples using a 4-second window. The dataset is formulated as a four-class motor imagery classification task.

Disease EEG datasets: three benchmark EEG datasets for disease detection are included.

MDD Mumtaz et al. (2017): The Major Depressive Disorder dataset is a benchmark for depression classification, consisting of EEG recordings from 61 subjects, including patients with major depressive disorder and healthy controls. EEG signals are recorded using a 19-channel system at a sampling rate of 256 Hz, with a total recording duration of 35,415 seconds. The data are segmented into 7,083 samples. The dataset is formulated as a binary classification task.

Table 3: Summary of EEG datasets used in pretraining and downstream evaluation.

Category	Dataset	#Subjects	#Channels	Sampling Rate (Hz)	Duration (s)	#Samples
Emotion	SEED	15	62	200	152,730	30,375
	SEED-IV	15	62	200	732,750	30,045
	FACED	123	32	256	103,320	13,776
	DEAP	32	32	128	80,640	3,840
Motor Imagery	EEGMMIDB	109	64	160	39,348	9,837
	BCICIV-2a	9	22	250	21,000	5,088
Disease	MDD	61	19	256	35,415	7,083
	UCSD-PD	31	32+9	512	5,952	1,201
	ADFTD	88	19	500	69,595	13,919
Stress	Arithmetic	35	20	500	6,715	1,701

UCSD-PD Swann et al. (2015); George et al. (2013): The UCSD Parkinson's EEG dataset is a clinical benchmark collected at the University of California, San Diego, consisting of recordings from 31 subjects, including patients with Parkinson's disease (PD) and healthy controls (HC). EEG signals are recorded using a 32-channel system along with 9 additional channels at a sampling rate of 512 Hz, with a total recording duration of 5,952 seconds. The dataset contains 1,201 samples and is formulated as a binary classification task.

ADFTD Miltiadous et al. (2024): The ADFTD dataset is a clinical EEG benchmark for cognitive disorder classification, consisting of recordings from 88 subjects, including patients with Alzheimer's disease (AD), frontotemporal dementia (FTD), and healthy controls. EEG signals are recorded using a 19-channel system at a sampling rate of 500 Hz, with a total recording duration of 69,595 seconds. The dataset contains 13,919 samples and is formulated as a three-class classification task.

Stress EEG datasets: One benchmark EEG dataset for cognitive load estimation is included.

Arithmetic Zyma et al. (2019): The Arithmetic dataset is an EEG benchmark for cognitive load estimation based on mental arithmetic tasks, consisting of recordings from 35 subjects. EEG signals are recorded using a 20-channel system at a sampling rate of 500 Hz, with a total recording duration of 6,715 seconds. The data are segmented into 1,701 samples using a 5-second window. The dataset is formulated as a binary classification task, making it suitable for evaluating cognitive state decoding and cross-subject generalization.

A.4.2 Model Settings

Model Structure. We adopt a unified BrainMamba configuration across all experimental settings. All EEG signals are resampled to a uniform sampling rate of 200 Hz. The model is built upon a stacked State Mamba backbone with a hidden dimension of 200. Specifically, the input and output dimensions are both set to 200, and the backbone consists of $n_{\text{layer}} = 12$ State Mamba blocks. Each State Mamba layer is parameterized with a state dimension of $d_{\text{state}} = 16$, a convolution kernel size of $d_{\text{conv}} = 4$, and an expansion ratio of 2. The sequence length is fixed to 30. The feed-forward dimension is set to 512 with a dropout rate of 0.1. To enhance model capacity, a Mixture-of-Experts (MoE) module is incorporated with 8 experts and Top-2 routing. The MoE balancing loss coefficient is set to 0.01. The expert capacity factor is set to 2.0 during both training and evaluation. The input EEG signals are first transformed through a structure-aware embedding module consisting of three components: (1) a temporal convolutional projection, denoted as `proj_in`, to capture waveform features; (2) a spectral projection via real-valued FFT, denoted as `spectral_proj`, to encode frequency-domain information; and (3) a convolution-based positional encoding module, denoted as `positional_encoding`, to model spatiotemporal relationships. The model follows a periodic hybrid design. Specifically, with a period parameter `perTrans` $= 6$, a MoE block and a spatiotemporal attention module are inserted after every six State Mamba layers. As a result, for $n_{\text{layer}} = 12$, the effective architecture consists of 12 State Mamba layers, 2 MoE blocks, and 2 spatiotemporal attention modules.

Pretraining Settings. BrainMamba adopts a large-scale pretraining strategy on heterogeneous EEG datasets. The model is trained for 60 epochs with a batch size of 24. We use the AdamW optimizer with an initial learning rate of 2×10^{-4}, a minimum learning rate of 1×10^{-5} , and a weight decay

of 1×10^{-5}. The learning rate is scheduled using CosineAnnealingLR. A masked reconstruction pretext task is employed with a masking ratio of 0.5. The reconstruction loss is implemented using the Huber loss with $\beta = 1.0$. Each epoch consists of 1,500 training steps, and round-robin sampling is also performed with 1,500 steps. In addition, temperature-based sampling is applied to balance heterogeneous datasets. To ensure stable training, dataset-level standardization is applied with 16 normalization batches, a global scaling factor of 1.0, and input clipping at 5.0. The training process logs results every 50 iterations and saves checkpoints every 5 epochs. All pretraining experiments are conducted on a high-performance computing server equipped with 8 NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs, each with approximately 96 GB of memory. The system is powered by an AMD EPYC 9654 96-Core processor and is configured with 503 GB of system memory.

Downstream Settings. For downstream tasks, the pretrained backbone is fine-tuned using task-specific classifiers. The model is trained for 30 epochs with a batch size of 64 using the AdamW optimizer. The learning rate is set to 1×10^{-4} , with a minimum learning rate of 1×10^{-6} and a weight decay of 0.05. Gradient clipping is applied with a threshold of 1.0. The classification head adopts a single-layer design over aggregated patch representations. Label smoothing with a factor of 0.1 is applied to improve generalization. Dataset standardization is consistent with pretraining. For different downstream tasks, we adopt task-specific data splitting strategies. For FACED and DEAP datasets that contain multiple subjects, the leave-one-subject-out (LOSO) protocol is used to evaluate cross-subject generalization. For SEED and SEED-IV, we employ a more challenging cross-session setting, which involves both cross-subject and cross-session evaluation, introducing significant distribution shifts and thus providing a more rigorous assessment of model generalization. For the remaining datasets with limited subjects or scenarios where LOSO is not applicable (e.g., datasets with subject-level labels such as healthy vs. disease groups), we adopt strict cross-subject splits to ensure a fair and realistic evaluation. In the experiment, Accuracy (Acc) and macro-averaged F1 score (Macro-F1) are adopted as evaluation metrics. Accuracy reflects the overall prediction correctness and provides an intuitive comparison across different methods. Macro-F1 computes the average F1 score across all classes with equal weighting, ensuring that performance on minority classes is properly evaluated.

A.4.3 BASELINES

To evaluate the performance of BrainMamba, in the experiment, we compare BrainMamba with a range of state-of-the-art baseline methods, including both self-supervised learning approaches and recent EEG foundation models. The self-supervised baselines focus on learning general representations from unlabeled EEG data, typically through contrastive learning or reconstruction-based pretext tasks. The EEG foundation models are pretrained on large-scale EEG corpora and aim to learn transferable representations for cross-task and cross-domain generalization. These baselines provide comprehensive comparisons with both conventional representation learning paradigms and recent large-scale pretrained approaches.

BIOT Yang et al. (2023a) employs a Transformer-based biosignal encoder that converts heterogeneous biosignals into a unified token representation, enabling cross-dataset pretraining across different channel configurations, signal lengths, and acquisition settings.

ContraWR Yang et al. (2023b) adopts a self-supervised contrastive learning framework, where augmented EEG samples are contrasted against a global world representation constructed from dataset-level statistics to learn robust representations without relying on conventional negative sampling.

EEG2Rep Mohammadi Foumani et al. (2024) follows a masked self-supervised learning paradigm, learning contextual EEG representations by reconstructing informative masked inputs and encouraging the model to capture transferable latent structures from unlabeled EEG signals.

State Mamba Weng et al. (2026) develops a spatiotemporal state-space architecture that jointly models spatial and temporal brain-state transitions with Mamba, complemented by global-local attention and representation/pattern alignment objectives for cross-subject representation learning.

LaBraM Jiang et al. (2024) adopts a Transformer-based masked EEG modeling framework, representing EEG segments as discrete neural tokens and learning general representations through large-scale masked pretraining.

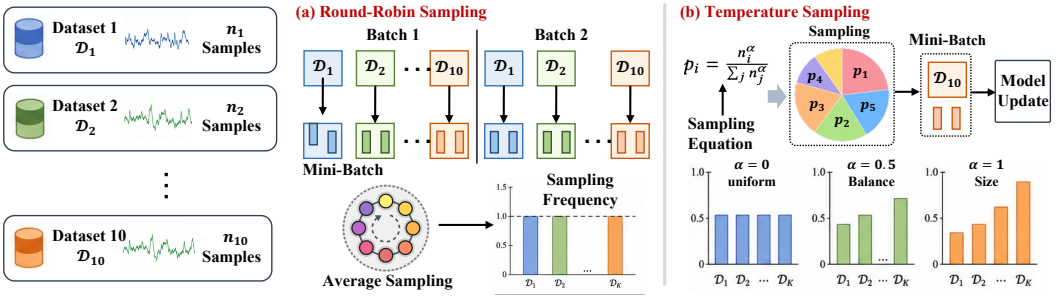


Figure 6: Two pretraining strategies: Round-Robin (RR) and Temperature-based Sampling (TS).

CBraMod Wang et al. (2024a) introduces a criss-cross Transformer architecture that separately models temporal and spatial dependencies in parallel, together with asymmetric conditional positional encoding to accommodate heterogeneous channel configurations.

EEGMamba Gui et al. (2024) replaces conventional Transformer backbones with a bidirectional Mamba encoder, leveraging state-space modeling for efficient long-range EEG sequence representation while maintaining scalability for large-scale pretraining.

A.5 Round Robin and Temperature Sampling

(1) **Round-Robin sampling**. In the round-robin sampling strategy, training data are drawn from multiple EEG datasets in a cyclic manner. Specifically, at each iteration, the model sequentially samples mini-batches from each dataset in a fixed order, ensuring that all datasets are visited with equal frequency. This strategy prevents large datasets from dominating the training process and guarantees uniform exposure to all data sources, which is particularly important in heterogeneous EEG settings with varying dataset sizes and distributions. As a result, round-robin sampling provides a stable and balanced training process across multiple datasets.

(2) **Temperature Sampling**. To address the limitation of round-robin sampling in handling datasets with varying scales, we further adopt a temperature-based sampling strategy. Specifically, the probability of sampling from dataset i is defined as: $p_i = \frac{n_i^\alpha}{\sum_j n_j^\alpha}$, where n_i^α is the size of specific dataset i and $\alpha \in [0, 1]$ is the temperature parameter controlling the sampling distribution. When α is large, the sampling distribution tends to favor larger datasets; when α is small, it becomes more uniform, increasing the sampling probability of smaller datasets; when $\alpha = 0$, the sampling becomes uniform across datasets. This strategy provides a flexible mechanism to balance the trade-off between leveraging large-scale data and ensuring sufficient coverage of smaller datasets. As a result, temperature-based sampling improves the robustness and generalization of the model when trained on heterogeneous EEG datasets.

B Experiment

We conduct extensive experiments across diverse settings to comprehensively evaluate the proposed framework, focusing on the following research questions:

(1) **Full generalization evaluation** assesses the general-purpose capability of EEG foundation models. Specifically, the model is pretrained on large-scale unlabeled EEG data collected from multiple datasets with diverse distributions, and then evaluated across a range of downstream tasks to measure its generalization ability.

(2) **Leave-one-dataset-out generalization (LODO)** evaluates transferability to datasets that are entirely unseen during pretraining. For each target dataset, we exclude it from multi-dataset self-supervised pretraining, use only its training split for downstream adaptation, and evaluate on the held-out test split. This setting provides a stricter assessment of cross-dataset generalization because no unlabeled samples from the target dataset are observed during pretraining.

Table 4: Leave-one-dataset-out results on different downstream datasets, comparing BrainMamba with state-of-the-art baselines.

Method	SEED, 3-class		SEED-IV, 4-class		FACED, 9-class		DEAP Arousal, 2-class		MDD, 2-class	
	Acc	Macro F1	Acc	Macro F1	Acc	Macro F1	Acc	Binary F1	Acc	Binary F1
BIOT NeurIPS'23	50.27 ± 1.51	51.31 ± 1.04	33.19 ± 0.28	32.79 ± 0.25	51.77 ± 2.54	52.45 ± 2.37	58.41 ± 4.33	57.36 ± 4.50	86.54 ± 1.31	85.37 ± 1.28
LaBraM-Base ICLR'24	54.60 ± 1.07	53.78 ± 0.94	**34.88** ± 0.30	**34.70** ± 0.29	52.48 ± 2.30	54.02 ± 2.35	59.10 ± 6.28	61.38 ± 6.18	85.47 ± 1.03	85.34 ± 1.10
CBraMod ICLR'25	56.03 ± 1.20	56.37 ± 1.09	33.04 ± 0.22	33.28 ± 0.45	56.41 ± 2.73	56.38 ± 2.62	58.42 ± 5.99	59.35 ± 5.59	89.40 ± 0.29	89.73 ± 0.39
EEGMamba NN'25	54.54 ± 1.35	52.30 ± 1.44	32.19 ± 0.36	32.51 ± 0.34	55.48 ± 2.53	56.29 ± 2.35	59.24 ± 3.97	60.73 ± 4.11	88.91 ± 0.75	89.08 ± 0.82
BrainMamba (RR)	57.52 ± 1.27	57.10 ± 0.94	32.80 ± 0.27	32.76 ± 0.38	65.18 ± 2.06	65.56 ± 2.21	**60.00** ± 5.19	**61.73** ± 4.97	**90.01** ± 0.57	**90.45** ± 0.69
BrainMamba (TS)	**57.86** ± 1.29	**57.47** ± 1.17	33.89 ± 0.31	33.73 ± 0.37	**65.36** ± 2.37	**65.89** ± 2.41	57.90 ± 4.17	56.94 ± 4.21	89.74 ± 0.61	89.63 ± 0.59
Method	BCICIV, 4-class		EEGMMIDB, 4-class		ADFTD, 3-class		UCSD, 2-class		Arithmetic, 2-class	
	Acc	Macro F1	Acc	Macro F1	Acc	Macro F1	Acc	Binary F1	Acc	Binary F1
BIOT NeurIPS'23	44.17 ± 0.43	45.37 ± 0.58	55.24 ± 1.60	56.39 ± 1.73	60.38 ± 0.91	59.84 ± 1.43	52.94 ± 1.79	54.37 ± 1.84	63.05 ± 1.97	63.99 ± 1.24
LaBraM-Base ICLR'24	45.64 ± 0.89	45.33 ± 1.02	57.62 ± 1.37	57.49 ± 1.64	59.91 ± 1.37	60.11 ± 1.28	52.38 ± 1.51	52.69 ± 1.92	66.14 ± 1.76	66.84 ± 1.72
CBraMod ICLR'25	47.25 ± 0.77	48.30 ± 0.96	58.92 ± 1.04	58.99 ± 1.28	61.79 ± 0.83	62.13 ± 0.99	53.87 ± 1.84	53.29 ± 1.97	68.39 ± 1.45	68.82 ± 1.73
EEGMamba NN'25	47.05 ± 0.95	48.21 ± 0.98	58.06 ± 1.34	58.25 ± 1.57	60.32 ± 0.83	61.04 ± 1.01	53.04 ± 1.68	53.89 ± 1.75	66.50 ± 1.61	66.17 ± 1.39
BrainMamba (RR)	49.48 ± 0.94	49.64 ± 0.78	59.40 ± 1.07	**59.34** ± 1.23	**62.96** ± 0.70	63.11 ± 0.97	53.70 ± 1.33	55.74 ± 1.92	66.83 ± 1.29	64.48 ± 1.46
BrainMamba (TS)	**51.83** ± 0.87	**50.95** ± 0.74	**59.51** ± 1.33	59.33 ± 1.21	62.58 ± 0.95	**63.91** ± 1.40	**56.89** ± 2.06	**57.17** ± 1.43	**70.33** ± 1.76	**68.22** ± 1.91

(3) **Task-constrained evaluation** focuses on a task-specific pretraining setting, where only datasets relevant to the downstream task are used during pretraining. This setup aims to assess the model's ability to leverage task-related data and improve performance under limited domain alignment. For example, for emotion recognition tasks, the model is pretrained only on emotion-related EEG datasets and then evaluated on corresponding downstream benchmarks.

(4) **Channel-constrained evaluation** considers a channel-constrained pretraining setting, where only datasets with channel configurations consistent with the downstream task are used during pretraining. This setup aims to assess the model's ability to learn channel-specific representations under restricted spatial configurations. For example, the model is pretrained on datasets and evaluated on downstream tasks with similar channel configuration.

(5) **Self-evaluation** adopts a self-supervised learning setting, where the foundation model is first pretrained on a single dataset without labels and subsequently fine-tuned on downstream tasks from the same dataset. This protocol aims to assess the model's transfer ability from self-supervised pretraining to downstream supervised tasks, reflecting the quality and effectiveness of the learned representations.

(6) **Robustness to missing channel** evaluates model stability under incomplete channel observations. The model is trained with the complete montage and tested with progressively reduced subsets of electrodes without test-time adaptation. Multiple nested random montages are evaluated at each corruption level to measure both average performance degradation and sensitivity to the retained electrodes.

(7) **MoE routing specialization evaluation** analyzes whether different experts capture distinct EEG patterns across datasets, tasks, subjects, classes, and spectral characteristics. We record sample-level routing distributions from all MoE blocks and quantify expert specialization using expert-load entropy, Jensen–Shannon divergence, normalized mutual information, and correlations between expert utilization and EEG spectral features. This analysis reveals whether the MoE learns meaningful and interpretable specialization rather than merely distributing tokens across experts.

(8) **Mask reconstruction evaluation** examines whether the self-supervised objective recovers meaningful EEG structure beyond point-wise interpolation. We evaluate reconstruction only at masked positions, compare BrainMamba with masked-input and truncated-SVD baselines, and measure the recovery of waveform content, singular-value energy, and effective rank in the signal space.

(9) **Real-world application** assesses the performance of the model in real-world application scenarios. The pretrained model is applied to practical EEG datasets collected from real-world environments, which often exhibit complex noise, distribution shifts, and diverse recording conditions. This protocol aims to evaluate the robustness and generalization ability of the model under realistic and challenging settings.

B.1 Leave-one-dataset-out generalization

Leave-One-Dataset-Out (LODO) evaluation is designed to evaluate the cross-dataset transferability and out-of-distribution generalization capability of BrainMamba. For each target dataset, we re-

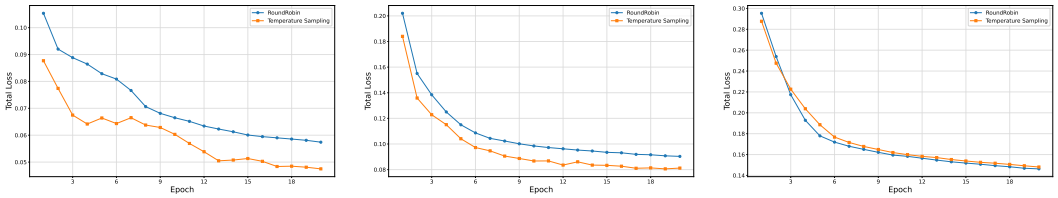
tain the original cross-subject generalization protocol within that dataset for downstream evaluation, while adopting a leave-one-dataset-out strategy across datasets. Specifically, the target dataset is completely excluded from the pretraining corpus, and BrainMamba is pretrained only on the remaining source datasets. The pretrained model is then adapted using the training subjects of the held-out target dataset and evaluated on unseen subjects from the same dataset. In this way, the model is required to transfer representations learned from heterogeneous source datasets to a completely unseen data distribution, while simultaneously generalizing across subjects within the target domain. Compared with standard cross-subject evaluation, this setting is substantially more challenging, as it introduces additional distribution shifts caused by differences in acquisition devices, channel configurations, experimental protocols, subject populations, and downstream tasks.

Table 4 presents the LODO results of BrainMamba compared with representative EEG foundation models and self-supervised baselines. It should be noted that several SSL baselines used in the full generalization setting cannot be directly adapted to heterogeneous channel configurations across different datasets. Therefore, in the LODO experiments, we only retain methods that can accommodate varying channel layouts across datasets, including BIOT, EEGFM, LaBraM, and CBraMod, to ensure a fair and consistent cross-dataset comparison. Consequently, a general performance degradation can be observed on several datasets. This degradation is particularly evident on SEED-IV, where BrainMamba exhibits a noticeable drop compared with its full-generalization performance and does not outperform the strongest baseline. This result suggests a relatively large distribution discrepancy between SEED-IV and the remaining pretraining datasets, such that representations learned from other datasets cannot be directly transferred to its target distribution. In addition, SEED-IV involves fine-grained four-class emotion recognition and exhibits substantial subject- and session-dependent variability, making its discriminative structures more dataset-specific. When SEED-IV is absent from pretraining, BrainMamba therefore lacks direct exposure to these distributional characteristics, leading to a larger transfer gap.

Among the compared baselines, CBraMod shows consistently strong LODO performance across multiple datasets. One possible reason is that its criss-cross Transformer explicitly alternates spatial and temporal interactions, providing a structured inductive bias for compressing and aggregating EEG dependencies along the channel and temporal dimensions. This modeling behavior is partially aligned with BrainMamba's motivation of progressively extracting structured spatiotemporal representations, although the two methods realize this process through different mechanisms: CBraMod relies on factorized attention-based interactions, whereas BrainMamba models coupled spatiotemporal evolution through state-space transitions and periodically restores high-order global interactions using spatiotemporal Transformers. Such structural inductive biases can be particularly beneficial under LODO, where the model needs to preserve transferable spatial-temporal patterns rather than relying on dataset-specific statistics.

Despite the increased difficulty of cross-dataset transfer, BrainMamba remains competitive or achieves the best performance on most held-out datasets. The improvements are especially pronounced on FACED, BCICIV, EEGMMIDB, UCSD, and Arithmetic, indicating that the learned representations retain substantial transferability even without observing the target dataset during pretraining. In particular, the strong results on datasets with different task types and channel configurations suggest that the proposed coupled state evolution and heterogeneous-data modeling mechanisms capture structural patterns that can be reused across domains. However, the variation in performance across datasets also reveals that cross-dataset transfer depends strongly on the similarity between source and target distributions. Datasets whose discriminative patterns are more consistent with the shared spatiotemporal structures learned from source datasets benefit more from large-scale pretraining, whereas datasets with stronger dataset-specific temporal, spectral, or subject-dependent characteristics, such as SEED-IV, remain considerably more challenging.

We also observe that neither RR nor TS consistently dominates under LODO. While temperature sampling achieves better results on several datasets, RR performs better on others such as DEAP and MDD. This differs from the relatively consistent advantage of TS under full generalization, suggesting that once the target distribution is completely unavailable during pretraining, the effectiveness of the sampling strategy is determined not only by dataset scale but also by how well the mixture of source datasets covers the latent structure of the unseen target domain.



(a) The pretraining loss curve of emotion recognition group (RR vs TS).

(b) The pretraining loss curve of disease diagnosis group (RR vs TS).

(c) The pretraining loss curve of motor imagery group (RR vs TS).

Figure 7: The pretraining loss curves under the task-constrained setting across three task groups: emotion recognition, disease diagnosis, and motor imagery.

Table 5: Task-constrained evaluation results on emotion recognition group.

Method	SEED, 3-class		SEED-IV, 4-class		FACED, 9-class		DEAP Arousal, 2-class	
	Acc	Macro F1	Acc	Macro F1	Acc	Macro F1	Acc	Binary F1
SSL Baseline	54.12 ± 0.92	54.92 ± 1.07	35.31 ± 0.37	36.12 ± 0.57	54.69 ± 5.34	53.89 ± 4.68	64.71 ± 4.67	63.99 ± 5.32
BIOT NeurIPS'23	53.41 ± 1.19	53.05 ± 1.25	33.18 ± 0.42	34.87 ± 0.29	54.06 ± 1.76	55.62 ± 1.48	66.71 ± 4.32	66.24 ± 4.38
LaBraM-Base ICLR'24	53.17 ± 1.85	55.94 ± 1.54	36.01 ± 0.19	36.46 ± 0.25	55.06 ± 2.48	56.48 ± 1.97	66.33 ± 5.62	67.82 ± 5.53
CBraMod ICLR'25	53.70 ± 2.50	55.78 ± 1.88	37.15 ± 0.23	36.21 ± 0.14	57.10 ± 1.35	59.94 ± 2.17	66.19 ± 5.72	67.30 ± 4.96
EEGMamba NN'25	55.01 ± 1.35	56.41 ± 2.01	36.74 ± 0.41	36.16 ± 0.35	56.13 ± 2.01	57.14 ± 2.41	68.55 ± 4.29	68.42 ± 3.91
BrainMamba (RR)	57.29 ± 1.52	57.43 ± 1.09	38.47 ± 0.39	38.79 ± 0.46	67.88 ± 2.07	66.94 ± 2.54	**70.15 ± 4.75**	**71.36 ± 4.62**
BrainMamba (TS)	**58.27 ± 0.92**	**58.20 ± 0.27**	**38.74 ± 0.23**	**39.54 ± 0.24**	**69.85 ± 4.02**	**69.72 ± 3.74**	69.38 ± 3.47	70.06 ± 4.02

B.2 Task-Constrained Evaluation

The objective of task-constrained evaluation is to explicitly enforce semantic and statistical alignment between the pretraining datasets and the target downstream task in multi-source heterogeneous EEG pretraining and transfer. This design mitigates overfitting to task-irrelevant biases by encouraging the model to focus on task-relevant discriminative features within a shared representation space. In the Task-Constrained Evaluation, tasks are categorized into three groups: **emotion recognition**, **motor imagery**, and **disease diagnosis**. The emotion recognition group includes four datasets, FACED, DEAP, SEED, and SEED-IV. The disease diagnosis group covers Alzheimer's disease detection (ADFTD), Parkinson's disease identification (UCSD), and depression recognition (MDD). The motor imagery group consists of two datasets: EEGMMIDB and BCICIV. Under this paradigm, pretraining is restricted to datasets within each group, and the model is subsequently fine-tuned on downstream tasks within the same group. In addition, because some self-supervised baselines (e.g., State Mamba, EEG2Rep, and ContraWR) cannot be directly adapted to multi-dataset pretraining, we adopt the best-performing self-supervised baseline (denoted as "SSL baseline") for each dataset as the reference, allowing for a fair comparison across different settings.

Figure 7 presents the pretraining loss curves for the three task groups. In the emotion recognition group, noticeable fluctuations can be observed in the early stages of pretraining. This can be attributed to the heterogeneity and scale differences across multiple datasets. Among the emotion datasets, substantial heterogeneity exists in channel configurations (e.g., 62 vs. 30 channels), subject scales (e.g., 15 vs. 123 subjects), and underlying data distributions. In contrast, the disease diagnosis group exhibits a smoother and more stable convergence behavior. This is likely due to relatively consistent data characteristics and channel distributions across clinical datasets, which enable the model to learn more aligned representations during pretraining. For the motor imagery group, only two datasets are involved, the advantage of temperature sampling diminishes and effectively degenerates toward uniform sampling. As a result, the performance gap between RR and TS becomes marginal, and both strategies show similar convergence behavior. Overall, group-wise pretraining still demonstrates stable convergence across all task groups, indicating that the proposed framework can effectively adapt to different task domains while maintaining robust optimization dynamics.

Table 5 presents the results on the emotion group under the task-constrained setting. We observe that, compared to pretraining on all datasets, restricting pretraining to emotion-related datasets leads to slight performance drops on SEED and SEED-IV. This can be attributed to the fact that although

Table 6: Task-constrained evaluation results on motor imagery and disease diagnosis groups.

Method	BCICIV, 4-class		EEGMMIDB, 4-class		ADFTD, 3-class		UCSD, 2-class		MDD, 2-class	
	Acc	Macro F1	Acc	Macro F1	Acc	Macro F1	Acc	Binary F1	Acc	Binary F1
SSL Baseline	48.30 ± 1.29	49.11 ± 1.06	56.72 ± 0.97	56.89 ± 1.70	62.20 ± 1.12	62.10 ± 0.93	52.18 ± 1.71	56.74 ± 2.33	88.35 ± 0.96	89.25 ± 0.82
BIOT NeurIPS'23	51.37 ± 0.64	51.33 ± 0.61	59.35 ± 1.72	57.32 ± 1.91	64.35 ± 0.52	64.70 ± 0.75	53.79 ± 2.62	55.47 ± 2.04	88.17 ± 0.53	88.02 ± 0.94
LaBraM-Base ICLR'24	49.82 ± 1.34	50.36 ± 2.38	60.31 ± 0.81	60.72 ± 1.02	62.86 ± 1.04	64.57 ± 1.35	52.66 ± 1.70	55.21 ± 1.90	89.05 ± 1.32	87.42 ± 0.91
CBraMod ICLR'25	52.89 ± 0.84	53.04 ± 1.02	**64.19 ± 1.45**	**64.20 ± 1.08**	65.71 ± 1.14	66.82 ± 0.79	53.97 ± 1.97	55.31 ± 2.42	89.51 ± 0.92	87.25 ± 0.99
EEGMamba NN'25	52.41 ± 0.71	52.84 ± 1.03	61.33 ± 1.31	60.82 ± 1.42	66.03 ± 0.93	67.42 ± 1.24	52.85 ± 2.61	54.79 ± 2.44	88.46 ± 0.92	88.51 ± 1.01
BrainMamba (RR)	54.95 ± 0.71	54.42 ± 0.77	61.40 ± 1.08	61.49 ± 1.19	69.97 ± 1.15	68.03 ± 1.02	56.82 ± 2.06	59.47 ± 2.12	**90.01 ± 0.45**	**89.99 ± 0.64**
BrainMamba (TS)	**54.69 ± 0.82**	**54.60 ± 0.47**	61.42 ± 1.23	61.37 ± 1.01	**73.41 ± 1.29**	**68.84 ± 1.43**	**57.43 ± 2.31**	**61.37 ± 2.37**	88.53 ± 0.91	88.48 ± 0.35

the task semantics are aligned, the full pretraining setting benefits from gradient averaging across diverse tasks, resulting in more stable but coarse-grained representations. In contrast, the task-constrained setting focuses on semantically similar datasets, where statistical discrepancies (e.g., channel configurations, label spaces, and subject distributions) become more pronounced, leading to stronger gradient conflicts during optimization. Meanwhile, performance improvements are observed on the FACED and DEAP datasets, suggesting that the dominance of large-scale datasets (123 and 32 subjects) is amplified in the task-constrained setting, leading the model to better capture their distributions and structural semantics in pretraining.

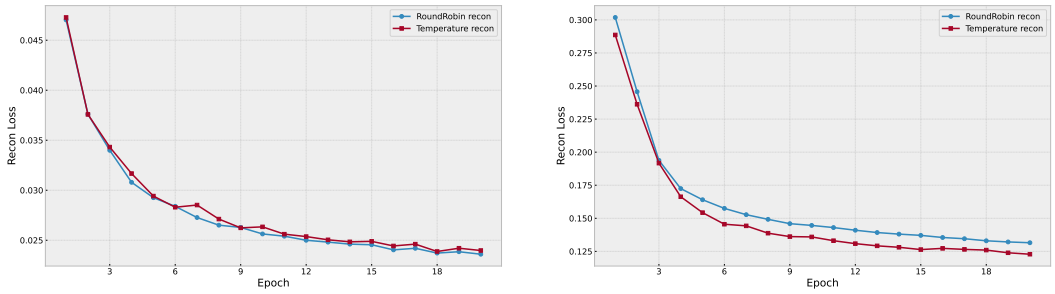
Table 6 presents the results on the motor imagery and disease diagnosis groups. For the motor imagery group, which contains only two datasets, RR and TS exhibit nearly identical performance, as the scheduling effect becomes negligible when the number of datasets is limited. In the disease diagnosis group, significant performance improvements are observed on the ADFTD dataset. This can be attributed to its relatively large number of subjects (88), which provides richer statistical diversity and more stable supervision signals during pretraining, thereby facilitating more effective representation learning. In contrast, performance on MDD decreases. This is likely due to the greater discrepancy between its task characteristics and those of other disease datasets. Under task-specific pretraining with fewer datasets, the generalization capability of the learned representations is inherently limited, making it more challenging for the model to learn robust and transferable representations.

Therefore, it can be concluded that restricting the pretraining of EEG foundation models from diverse general datasets to task-specific datasets also introduces inherent limitations: (1) The lack of cross-dataset diversity reduces representation coverage, thereby weakening generalization performance. (2) It amplifies dataset-specific biases, such as scale imbalance and distribution discrepancies, leading to more pronounced optimization dominance effects. Consequently, similar to large language models, the data composition and diversity are equally critical for EEG foundation models, as they directly influence the balance between generalization and task-specific performance.

B.3 Channel-Constrained Evaluation

To assess robustness under heterogeneous electrode configurations, we adopt a channel-constrained evaluation paradigm. This evaluation aims to systematically assess the transfer robustness of EEG foundation models under variations in channel configurations. In real-world scenarios, different datasets often employ varying numbers of electrodes and layouts (e.g., from 62/64 channels to 19/22 channels), which can lead to significant changes in spatial information density, cross-channel correlation structures, and noise characteristics, thereby affecting the reusability of pretrained representations in downstream tasks. To this end, in the channel-constrained evaluation, we divide the experiments into two settings: similar and different. In the similar setting, the channel configurations of the pretraining and downstream datasets are closely aligned, aiming to evaluate whether channel matching leads to more stable transfer. In contrast, the different setting introduces substantial discrepancies in channel configurations between pretraining and downstream tasks, allowing us to assess the model's cross-channel generalization capability and potential risks of negative transfer.

Specifically, we construct two groups of pretraining datasets with different channel configurations. The first group consists of SEED and SEED-IV, both of which share a 62-channel configuration, representing the **high-density channel configuration** with rich spatial information. The second group includes BCICIV-2a, MDD, and Arithmetic, with channel configurations of 22, 19, and 20, respectively, representing the **low-density channel configuration** with limited spatial resolution. We design two similar and two different downstream task settings to evaluate cross-channel transfer un-



(a) Pretraining reconstruction loss on SEED and SEED-IV datasets (high-density channel configuration).

(b) Pretraining reconstruction loss on Arithmetic, MDD, and BCICIV-2a datasets (low-density channel configuration).

Figure 8: Pretraining loss curves under the channel-constrained setting.

Table 7: Channel-constrained evaluation under different transfer settings.

Approach	Similar Transfer				Different Transfer			
	SEED+SEED-IV → MMI		BCIC+AR+MDD → ADFTD		SEED+SEED-IV → ADFTD		BCIC+AR+MDD → MMI	
	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1
BIOT NeurIPS'23	58.46 ± 0.84	58.31 ± 1.07	69.57 ± 1.72	68.42 ± 1.69	62.75 ± 1.51	60.33 ± 1.39	56.31 ± 1.17	57.06 ± 1.26
LaBraM-Base ICLR'24	57.36 ± 1.22	59.23 ± 0.94	68.62 ± 1.04	67.45 ± 1.10	59.38 ± 1.95	60.82 ± 1.40	55.46 ± 1.22	53.84 ± 1.25
CBraMod ICLR'25	**62.59 ± 1.54**	**62.33 ± 1.43**	71.82 ± 1.63	70.31 ± 1.39	63.82 ± 1.10	62.18 ± 1.71	**60.78 ± 1.60**	59.43 ± 1.79
EEGMamba NN'25	59.41 ± 0.89	59.99 ± 1.30	70.41 ± 1.77	70.04 ± 1.49	62.51 ± 1.52	60.37 ± 1.59	57.42 ± 1.21	56.47 ± 1.38
BrainMamba (RR)	59.46 ± 1.43	60.78 ± 1.24	**76.40 ± 1.26**	**72.42 ± 1.24**	**66.05 ± 1.21**	**63.92 ± 1.14**	59.79 ± 1.43	59.37 ± 1.60
BrainMamba (TS)	60.37 ± 0.81	61.43 ± 1.01	75.69 ± 1.17	71.87 ± 1.39	65.96 ± 1.32	63.82 ± 1.26	59.51 ± 1.52	**59.57 ± 1.01**

der varying channel configurations. For the similar settings, the channel scales between pretraining and downstream tasks are relatively aligned. Specifically, (1) SEED + SEED-IV → MMI (EEGMMIDB, 64 channels) represents a high-density-to-high-density transfer, and (2) BCICIV + ARITHM + MDD → ADFTD (19 channels) represents a low-density-to-low-density transfer. For the different settings, a clear mismatch in channel configurations is introduced. Specifically, (1) SEED + SEED-IV → ADFTD (19 channels) represents a high-density-to-low-density transfer, and (2) BCICIV + ARITHM + MDD → MMI (EEGMMIDB, 64 channels) represents a low-density-to-high-density transfer.

Figure 8 and Table 7 present the pretraining and downstream results of the channel-constrained evaluation, respectively. It can be observed that when transferring across mismatched channel configurations, significant performance degradation occurs on both EEGMMIDB and ADFTD tasks. This indicates that both high-density-to-low-density and low-density-to-high-density transfers introduce substantial information mismatch, due to discrepancies in spatial resolution, cross-channel dependency structures, and signal distributions, thereby hindering effective representation transfer. Under the channel-consistent (similar) setting, notable performance improvements are observed on the ADFTD dataset, suggesting that this dataset benefits from better-aligned spatial configurations and more consistent channel-wise representations during pretraining. In contrast, EEGMMIDB shows performance degradation across all models, which is likely due to semantic inconsistency between motor imagery tasks and the pretraining datasets, making transfer more challenging. Despite these challenges, BrainMamba demonstrates strong generalization capability across different channel configurations, indicating its ability to learn robust and transferable representations.

B.4 Self-Evaluation

We further introduce a self-evaluation paradigm to assess the intrinsic quality of the learned representations. Specifically, the foundation model is pretrained and fine-tuned on the same dataset, without cross-dataset transfer. The self-evaluation focuses on the model's capacity to capture dataset-specific structures and underlying signal semantics.

Figure 9 demonstrates the results of the self-evaluation paradigm. Overall, BrainMamba achieves strong performance across all datasets under the self-evaluation setting, demonstrating its ability

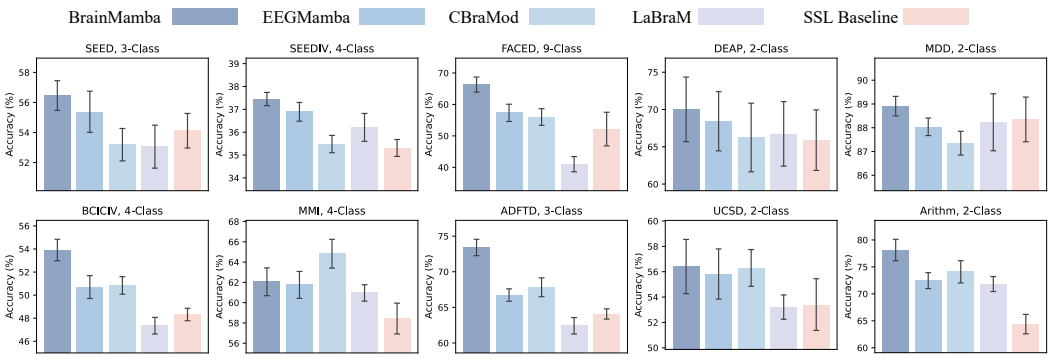


Figure 9: The results of the self-evaluation paradigm

to effectively capture dataset-specific structures and intrinsic signal semantics when pretrained and fine-tuned on the same dataset. Compared to full generalization and task-constrained evaluations, the performance gap between different methods is relatively reduced, indicating that the self-evaluation setting is less sensitive to representation transferability and more reflective of intrinsic modeling capacity. This dataset-isolated self-supervised evaluation demonstrates that BrainMamba is capable of learning both strong intrinsic representations and transferable features, achieving a favorable balance between dataset-specific adaptation and cross-dataset generalization.

B.5 Robustness to Missing Channels

To further evaluate the channel-wise scalability and generalization capability of BrainMamba, we conduct a controlled channel reduction experiment to assess its robustness under partially missing EEG electrodes. Starting from the original pretrained model, we progressively reduce the number of retained electrodes for each downstream dataset according to predefined ratios. For example, SEED is evaluated with 62, 48, 32, 16, and 8 retained channels, while the corresponding channel counts for other datasets are determined by mapping the same retention ratios to their original channel configurations. This setting simulates adapting a foundation model pretrained with full-channel EEG recordings to downstream scenarios with substantially fewer available electrodes, thereby testing whether the learned representations can remain effective under channel-distribution shifts and reduced spatial observations. At each channel-reduction level, we randomly generate three nested electrode subsets to reduce the potential bias introduced by any particular channel selection and to avoid attributing performance changes to a favorable or unfavorable electrode combination. Within each subset, smaller channel configurations are contained in larger ones, enabling controlled comparisons across different channel scales. The removed channels are zeroed out, while their original spatial positions and the overall input dimensionality are preserved, avoiding architectural changes across different channel configurations. For each reduced-channel setting, the model is initialized from the corresponding best-performing full-channel model and independently adapted and evaluated using a fixed retained-channel subset. We report the results aggregated over the three channel subsets, providing a more robust estimate of BrainMamba's ability to generalize from dense multi-channel pretraining to sparse electrode configurations.

Figure 10 demonstrates the performance of BrainMamba under progressively reduced electrode budgets across ten EEG datasets. For each channel budget, we report the mean Macro-F1 over three randomly generated nested electrode subsets, with the shaded regions reflecting the variation across different channel selections. Despite substantial reductions in the available spatial observations, BrainMamba maintains relatively stable performance across most datasets. SEED, MMI, BCICIV, ADFTD, and Arithmetic exhibit a generally smooth improvement as more electrodes become available, while SEED-IV, MDD, UCSD, and FACED show only moderate fluctuations across a wide range of channel budgets. Interestingly, on datasets such as MDD and UCSD, reducing the number of channels does not always lead to performance degradation and can occasionally yield slightly better results. One possible explanation is that full-channel EEG recordings may contain redundant, weakly informative, or noisy electrodes; removing a subset of such channels can suppress irrelevant spatial variation and act as an implicit regularizer during downstream adaptation. In addition, for

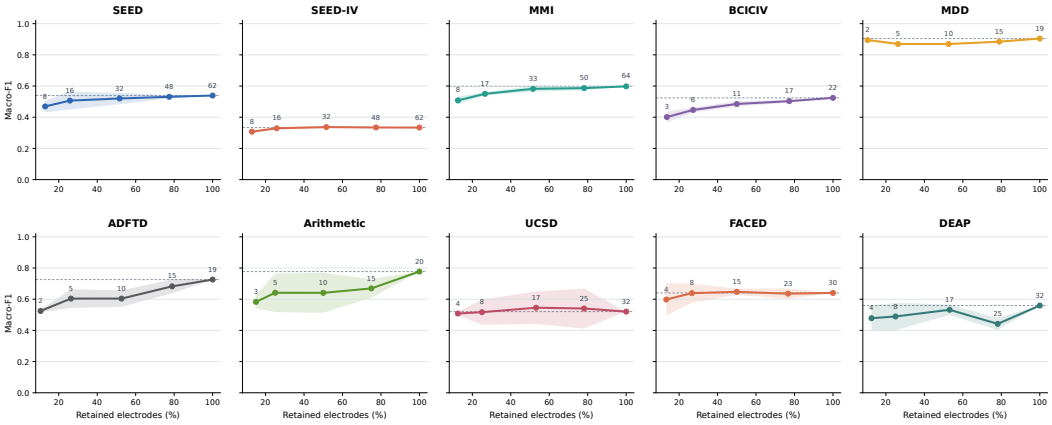


Figure 10: Robustness and Channel-wise Generalization of BrainMamba under Diverse Electrode Budgets and Reduced Channel Configurations.

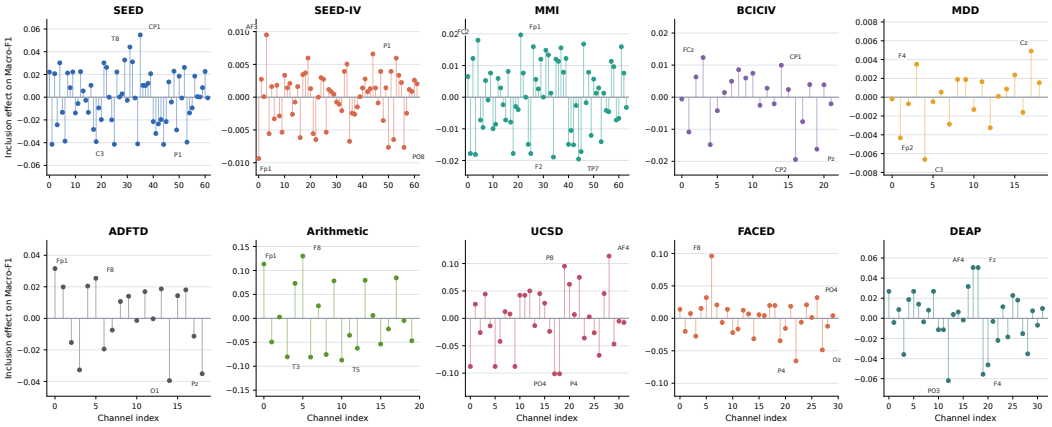
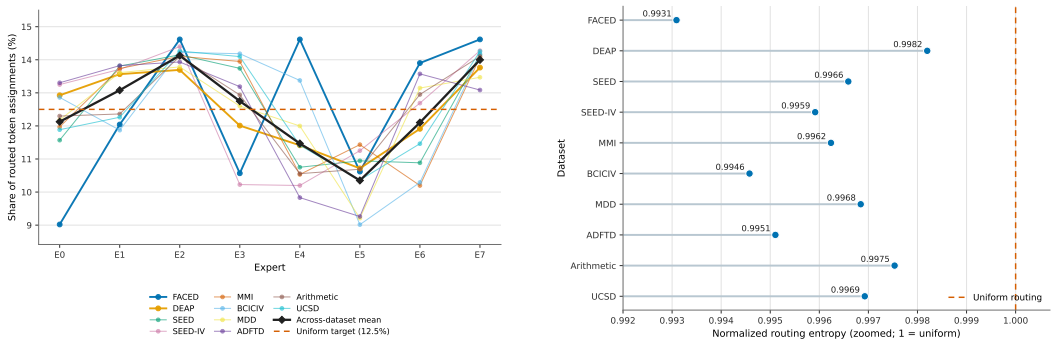


Figure 11: Fine-grained analysis of channel-specific contribution under varying electrode configurations.

tasks whose discriminative information is concentrated in a limited set of brain regions, a sparse subset that preserves these informative electrodes may retain most of the task-relevant signal while reducing interference from less relevant channels. We therefore do not interpret such local improvements as evidence that fewer channels are intrinsically preferable, but rather as an indication that BrainMamba can flexibly adapt to different spatial observation patterns.

The shaded regions further reveal an important property of channel reduction. As the number of retained electrodes decreases, the variance across different channel subsets generally becomes more pronounced on several datasets. This is expected because, under dense channel configurations, substantial spatial redundancy exists among neighboring or functionally correlated electrodes, such that removing a small subset can often be compensated by the remaining channels. In contrast, under sparse configurations, each retained electrode accounts for a larger fraction of the available spatial information, and the specific composition of the subset becomes increasingly important. Different subsets may preserve or remove different task-relevant cortical regions, resulting in larger variations in spatial coverage and discriminative information. Consequently, two configurations with the same number of retained electrodes can exhibit noticeably different performance depending on which electrodes are retained. The increasing spread under low-channel regimes therefore reflects the growing sensitivity to electrode configuration rather than instability of the model itself, and also motivates averaging over multiple randomly sampled subsets to reduce subset-specific bias.

To further understand why different reduced-channel configurations lead to different downstream performance, we perform a fine-grained channel influence analysis and quantify the contribution of



(a) Expert Load Distribution across Datasets.

(b) MoE Routing Balance and Expert Utilization.

Figure 12: The MoE expert load distribution and routing entropy.

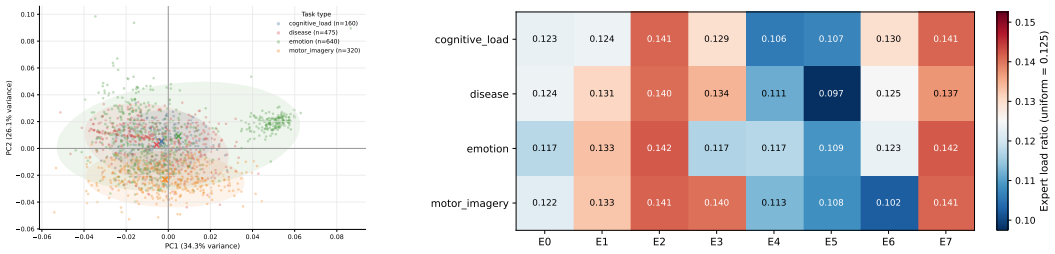
each electrode to Macro-F1. As shown in Figure 11, the inclusion effect varies substantially across electrodes within the same dataset, indicating that different channels contribute unequally to downstream prediction. Some electrodes yield consistently positive effects when included, while others exhibit near-zero or even negative contributions. For example, several channels in SEED, MMI, Arithmetic, UCSD, and FACED show relatively large positive inclusion effects, whereas other channels within the same datasets contribute only marginally or can slightly decrease performance. This heterogeneous contribution pattern provides a direct explanation for the larger variance observed under low-channel budgets: when only a small number of electrodes are retained, whether the subset preserves highly informative channels becomes increasingly important. The analysis also helps explain why reducing the number of channels can occasionally improve performance, as observed in datasets such as MDD and UCSD. Full-channel EEG does not necessarily provide uniformly useful information: some electrodes may capture weakly task-relevant activity, subject-specific variation, or measurement noise, and their inclusion can introduce redundant or distracting spatial information during downstream adaptation. Removing such channels may therefore suppress nuisance variation and effectively regularize the model. Conversely, if a reduced subset preserves electrodes with stronger task-discriminative contribution while excluding less informative channels, the resulting sparse configuration can match or even slightly outperform the full-channel setting. Importantly, this effect should not be interpreted as evidence that fewer channels are inherently better, but rather as evidence that the utility of additional channels depends on their information content and spatial relevance to the downstream task.

Overall, the results reveal that BrainMamba's channel robustness arises not from treating all electrodes as interchangeable, but from its ability to adaptively utilize the informative spatial evidence available under different channel configurations. This property is particularly important for real-world EEG deployment, where electrode layouts often vary across devices, acquisition protocols, and application scenarios.

B.6 MoE ROUTING SPECIALIZATION

To investigate expert specialization learned during pretraining, we analyze the routing behavior of the pretrained BrainMamba. For each input sample, we aggregate the token-level expert loads across all MoE blocks and normalize them into an eight-dimensional routing vector $r \in \mathbb{R}^8$, where each element denotes the relative utilization of one expert. We characterize expert specialization from two complementary perspectives: First, we measure inter-group routing separation using the pairwise Jensen–Shannon divergence between dataset-level routing distributions, with larger values indicating more distinct expert-allocation patterns. Second, we quantify the association between the Top-1 expert assignment and dataset or task using normalized mutual information (NMI), where higher NMI indicates stronger attribute-dependent routing specialization. We further report routing entropy and the number of active experts to characterize expert utilization diversity and identify potential expert collapse.

Figure 12 illustrates the load-balancing behavior of the MoE. The expert utilization remains well balanced across all pretraining datasets. As shown in the expert-load distribution, routed token



(a) Task-wise expert load distribution.

(b) The task-wise expert load distributions for MoE.

Figure 13: Task-aware Analysis of MoE Routing Specialization.

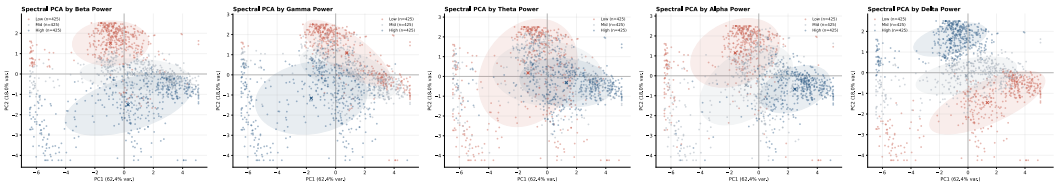


Figure 14: PCA-based analysis of routing distributions under band-specific spectral power

assignments are generally distributed around the uniform target of 12.5%, without any expert consistently dominating or becoming inactive. Although some experts, such as E2 and E7, receive slightly higher average loads while E4–E5 are used less frequently, the deviations remain moderate, indicating that the router avoids severe expert imbalance and collapse. Consistently, the normalized routing entropy ranges from 0.9931 to 0.9982 across datasets, remaining close to the maximum entropy of uniform routing. These results suggest that BrainMamba maintains diverse expert utilization while preserving mild routing preferences across heterogeneous EEG datasets, achieving a favorable trade-off between load balancing and expert specialization.

Figure 13 further examines whether MoE routing exhibits task-dependent specialization. As shown in Figure 13b, different EEG task types share a broadly consistent expert-utilization pattern, with E2 and E7 receiving relatively higher routing loads across tasks. Meanwhile, task-specific variations remain evident: for example, motor imagery assigns a larger fraction of tokens to E3 but less to E6, while disease samples show particularly low utilization of E5. These results suggest that the router does not allocate disjoint experts to individual tasks, but instead adapts the relative utilization of a shared expert pool according to task characteristics. This pattern is further reflected in the PCA visualization of sample-level routing profiles in Figure 13a. The first two principal components explain 34.3% and 26.1% of the routing variance, respectively. Although routing distributions from different tasks substantially overlap, their centers exhibit systematic shifts: motor-imagery samples are displaced toward lower PC2 values, whereas emotion samples tend to occupy relatively higher PC2 regions. The coexistence of substantial overlap and task-dependent shifts indicates that BrainMamba learns a shared routing space with moderate task-conditioned specialization, rather than rigid task-to-expert partitioning.

Figure 14 analyzes the spectral specialization of MoE routing. The PCA visualization shows that samples with different band-power levels exhibit structured separation in the routing space across multiple frequency bands, indicating that expert allocation is systematically related to EEG spectral characteristics. Consistently, individual experts exhibit distinct spectral correlation profiles shown in Figure 15. For instance, E0–E1 are negatively associated with beta-related features, whereas E2, E4, E5, and E7 show stronger positive beta associations; E6 instead favors delta activity while suppressing beta/gamma-related patterns. These results suggest that different experts capture complementary spectral characteristics, supporting meaningful functional specialization rather than redundant expert behavior.

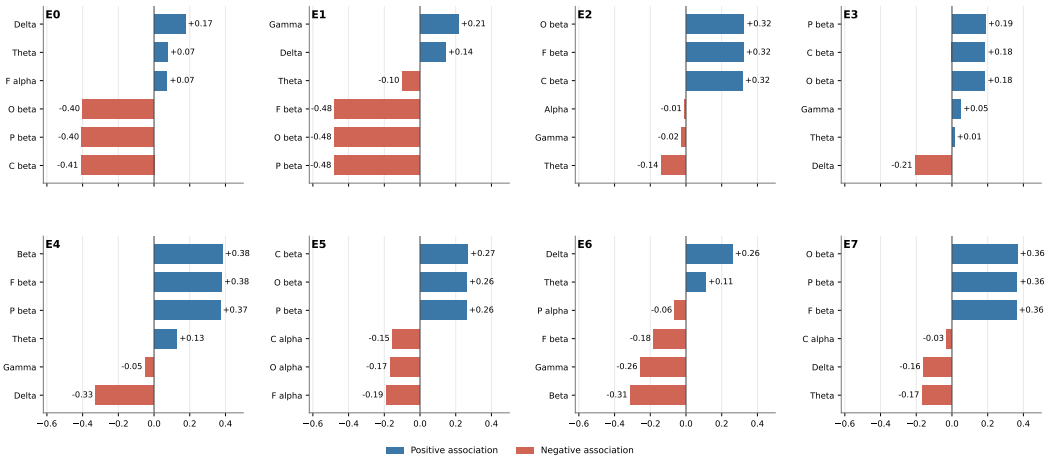


Figure 15: Expert-wise correlation between routing ratios and EEG spectral features

B.7 Mask reconstruction evaluation

In language models pretraining, next-token prediction can be directly evaluated using metrics such as perplexity and token-level accuracy, providing an interpretable measure of how well the pretraining objective is learned. Similarly, for BrainMamba, we aim to evaluate masked reconstruction beyond the overall reconstruction loss. We therefore analyze the recovered EEG signals from multiple perspectives, including waveform fidelity, spectral structure, and low-rank organization, to determine whether the model merely performs local point-wise interpolation or reconstructs the underlying structured dynamics of EEG signals.

Figure 16 provides a qualitative evaluation of masked EEG reconstruction from both temporal and spectral perspectives. In the left part, BrainMamba consistently reconstructs masked waveform segments across heterogeneous EEG datasets while preserving diverse dataset-specific temporal characteristics. The recovered signals retain high-frequency oscillatory activities, abrupt state transitions, plateau-like responses, slow-varying trends, and irregular waveform morphology, demonstrating robustness to substantially different signal dynamics. Notably, the reconstructed segments remain temporally coherent with the surrounding visible context and preserve the overall amplitude and transition patterns of the original signals, rather than degenerating into locally smoothed or linearly interpolated trajectories. In the right part, we further examine whether such recovery extends beyond individual channels to the multichannel structural level. The reconstructed representations closely reproduce the dominant cross-channel activation patterns and their temporal evolution, including localized activation bands, synchronized channel-wise variations, high-contrast transition regions, and large-scale temporal structures. These patterns remain visually consistent with the original signals across datasets with markedly different spatial organizations, suggesting that the model captures dependencies shared across channels rather than independently reconstructing each channel. Overall, the qualitative results indicate that BrainMamba recovers not only local waveform content but also the broader temporal and cross-channel organization of EEG signals. This provides evidence that masked reconstruction encourages the model to learn structured EEG dynamics instead of merely performing point-wise signal completion.

To further evaluate whether masked reconstruction recovers the intrinsic structural organization of EEG signals, rather than merely minimizing point-wise reconstruction error, we conduct a low-rank structural consistency analysis across multiple pretraining datasets. Specifically, we compare three inputs: the masked input, which directly reflects the structural distortion introduced by masking; SVD-r8, a rank-8 truncated singular value decomposition baseline that reconstructs the signal using only the eight largest singular components; and the reconstruction produced by BrainMamba. The SVD-r8 baseline serves as a controlled low-rank reference, allowing us to distinguish whether BrainMamba simply imposes a generic low-rank prior or is able to adaptively recover the dataset-dependent intrinsic structure of EEG signals. We compare all reconstructed signals against the corresponding original targets using three complementary metrics: top-k spectrum error, cumulative-

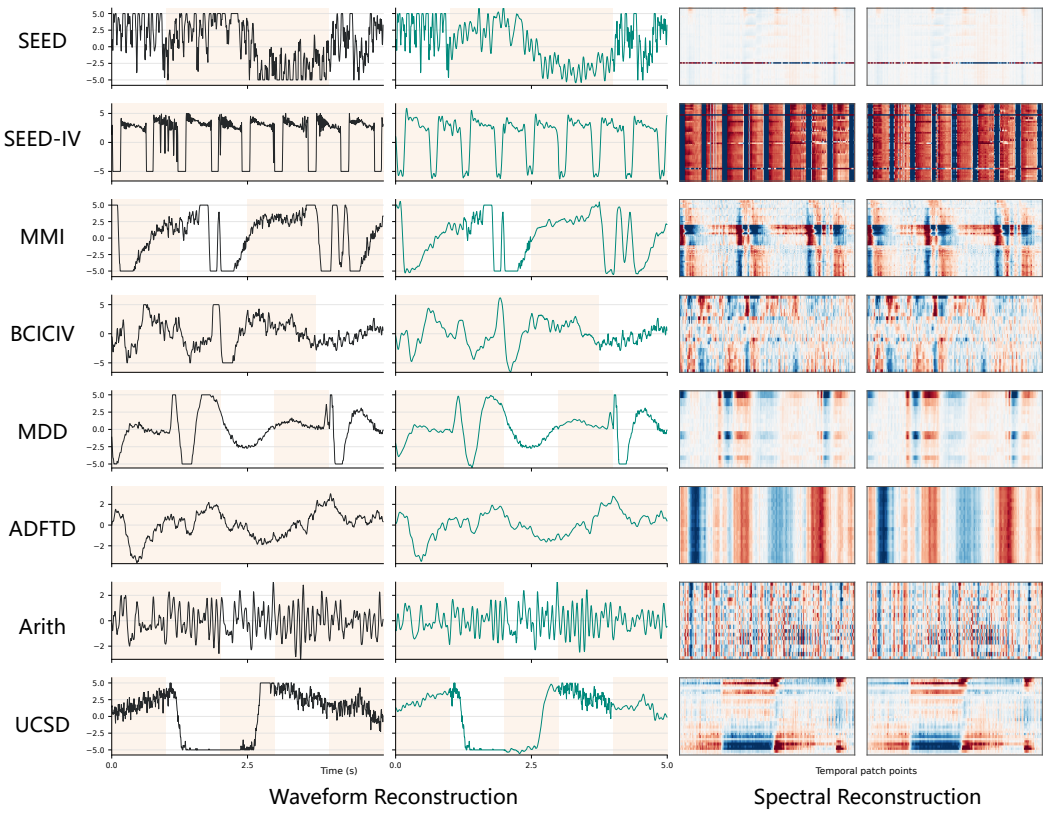


Figure 16: Qualitative visualization of masked EEG reconstruction across different datasets. The left two columns show waveform-level reconstruction, where the black curves denote the original signals, the shaded regions indicate the masked segments, and the green curves show the reconstructed signals. The right two columns visualize the corresponding multi-channel spectral representations before and after reconstruction. Across heterogeneous EEG datasets, BrainMamba consistently recovers both local waveform dynamics and global spectral structures, demonstrating that masked reconstruction preserves not only point-wise signal content but also the underlying temporal and cross-channel organization.

energy L1 distance, and effective rank. As shown in Figure 17, BrainMamba achieves consistently lower top-k spectrum error than both the masked input and SVD-r8 across nearly all datasets, indicating more faithful recovery of the dominant singular spectrum. The improvement is particularly pronounced on ADFTD, where the spectral deviation is substantially reduced, suggesting that the reconstruction preserves not only local signal values but also the dominant global structure encoded by leading singular components. The cumulative-energy L1 distance further shows that BrainMamba better matches the energy distribution of the original signals across singular modes. It achieves the lowest distance on almost all datasets, clearly outperforming the structurally distorted masked input and the fixed-rank SVD-r8 baseline. This indicates that BrainMamba more effectively preserves the overall spectral-energy organization of EEG signals. The effective rank provides a complementary view of intrinsic structural dimensionality. Masking generally inflates the effective rank, whereas SVD-r8 constrains it to a nearly fixed value of 7–8 and therefore cannot capture dataset-specific structural complexity. In contrast, BrainMamba closely follows the effective rank of the original targets across datasets, demonstrating that it adaptively recovers intrinsic dimensionality rather than simply imposing a generic low-rank prior. These results demonstrate that BrainMamba does not simply impose a generic low-rank prior, but adaptively reconstructs the dataset-dependent structural complexity of EEG signals.

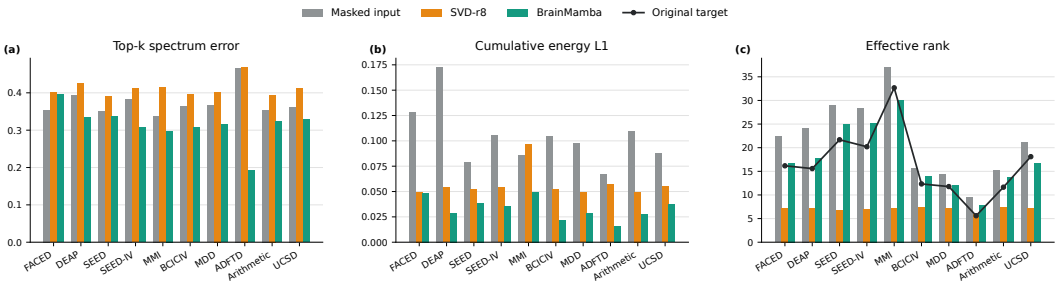


Figure 17: Low-rank structural consistency of masked EEG reconstruction across datasets. We compare masked input, SVD-r8, and BrainMamba with the original target using (a) top-k spectrum error, (b) cumulative-energy L1 distance, and (c) effective rank. BrainMamba better preserves the dominant spectrum, energy distribution, and intrinsic dimensionality of the original EEG signals.

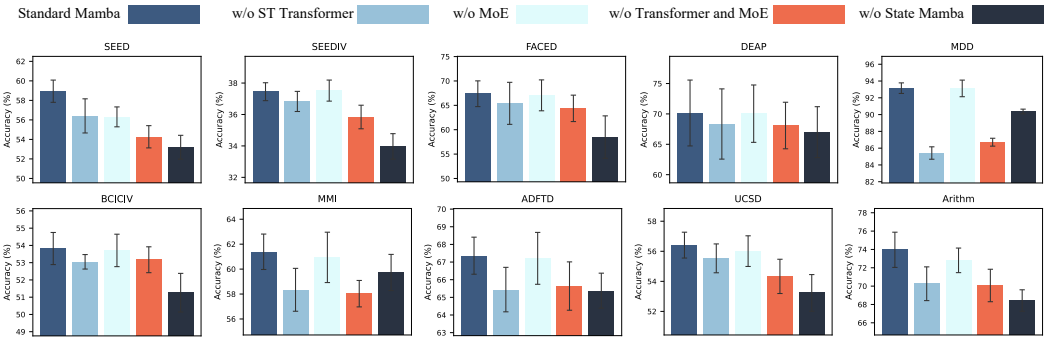


Figure 18: The results of the architectural ablation study of BrainMamba

B.8 Ablation Study

We conduct comprehensive ablation studies on both the architectural design and model depth to validate the effectiveness and rationality of BrainMamba.

Architectural Ablation. We investigate the contribution of each core component in BrainMamba. Specifically, *Standard BrainMamba* denotes the full model; *w/o ST Transformer* removes the spatiotemporal Transformer module; *w/o MoE* replaces the Mixture-of-Experts module with a standard FFN; *w/o State Mamba* replaces the State Mamba module with a spatiotemporal Transformer. All architectural ablations are conducted under the full pretraining setting on all datasets, followed by fine-tuning on each downstream task, ensuring fair and consistent comparison.

Depth Ablation. To evaluate the impact of model capacity, we vary the number of State Mamba layers and construct models with 6, 12, and 18 layers. For depth ablation, we adopt a supervised training setting on each dataset rather than full multi-dataset pretraining. This design isolates the effect of model depth on representation capacity, avoiding the confounding influence of multi-dataset pretraining dynamics and ensuring a fair comparison across different model sizes.

Figure 18 presents the results of the architectural ablation study of BrainMamba. It can be observed that removing the spatiotemporal Transformer (ST Transformer) leads to a significant performance degradation. In this case, the model degenerates into the purely State Mamba Weng et al. (2026), which constrains the model's representation capacity to a low-rank subspace. This demonstrates the critical role of the ST Transformer in modeling global spatiotemporal relationships. Conversely, when the State Mamba module is entirely replaced by the ST Transformer, the model cannot capture coupled spatiotemporal dynamics in multi-channel EEG signals, resulting in degraded performance. In addition, we find that removing the MoE module does not uniformly affect performance across all datasets. On complex datasets, the absence of MoE leads to significant performance drops, which can be attributed to increased data heterogeneity and distribution diversity, where a single shared representation is insufficient to capture diverse patterns. In contrast, on relatively simple binary

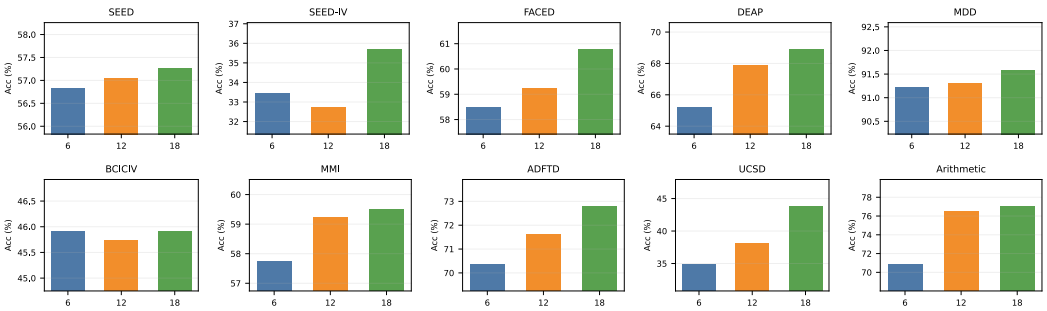


Figure 19: The results of the depth ablation study of BrainMamba

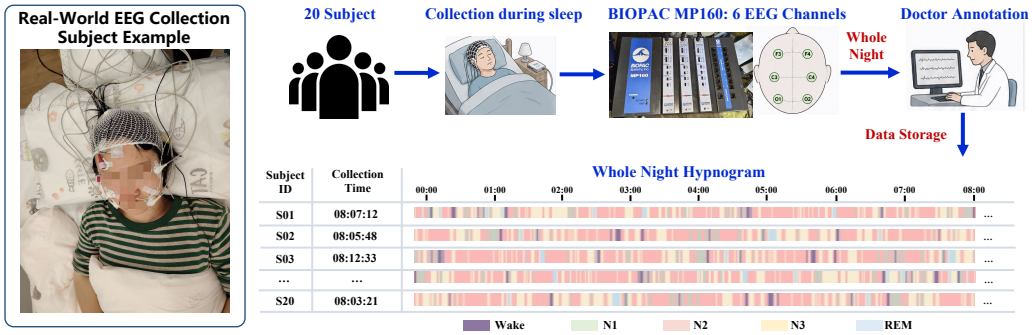


Figure 20: Overview of the real-world sleep EEG data collection system. Multi-channel EEG signals are collected from 20 subjects during overnight sleep using the BIOPAC MP160 system with six EEG channels. Each recording spans a full-night session and is manually annotated by a professional sleep clinician according to standard sleep staging criteria, including Wake, N1, N2, N3, and REM. The resulting hypnograms provide temporally aligned labels across the entire recording, forming a dataset for sleep stage analysis.

classification datasets (e.g., MDD and UCSD), the impact of removing MoE is less pronounced, suggesting that such tasks require less model capacity and adaptive specialization. These results validate the effectiveness of the BrainMamba architecture, demonstrating that the combination of structured state-space modeling, global spatiotemporal interaction, and adaptive expert routing is crucial for achieving strong performance across diverse EEG tasks.

Figure 19 presents the results of the depth ablation study, where the supervised performance on each dataset is illustrated using bar charts. It can be observed that, for most datasets, increasing model depth leads to stronger representation capacity and improved performance. This indicates that deeper BrainMamba architectures are more effective in capturing complex temporal dependencies and cross-channel interactions in EEG signals. However, for certain datasets (e.g., SEED-IV and BCICIV), the performance does not exhibit a clear monotonic relationship with depth. This may be attributed to factors such as limited data scale, increased optimization difficulty, or potential overfitting in deeper models, which can offset the benefits of increased model capacity. Overall, these results demonstrate that while increasing depth generally enhances representation capacity, its effectiveness depends on the characteristics of the dataset. This suggests that employing deeper BrainMamba architectures for pretraining on large-scale EEG datasets enables the model to learn more expressive, robust, and generalizable spatiotemporal representations, bringing it closer to general-purpose and highly transferable EEG foundation models, analogous to large language models.

B.9 Real-World Application

To further evaluate the robustness and real-world applicability of BrainMamba, we collect a real-world sleep EEG dataset to assess the model's generalization capability under challenging conditions. The sleep EEG collection system is illustrated in Figure 20. A total of 20 subjects are recorded

Table 8: Performance comparison on the sleep dataset under supervised training and finetuning settings.

Approach	Supervised		Foundation Finetune	
	Acc	Macro-F1	Acc	Macro-F1
LaBraM-Base	38.41 ± 4.17	35.35 ± 3.86	40.92 ± 3.89	38.81 ± 3.71
CBraMod	<u>40.91</u> ± 3.76	<u>38.72</u> ± 2.90	<u>45.99</u> ± 2.95	<u>42.31</u> ± 3.77
EEGMamba	39.87 ± 2.86	35.41 ± 4.10	42.27 ± 3.10	41.28 ± 2.85
BrainMamba	**42.31** ± **3.94**	**39.70** ± **3.76**	**51.22** ± **4.85**	**45.81** ± **4.21**

overnight using the BIOPAC MP160 system. For each subject, six EEG channels are collected, including F3, C3, O1, F4, C4, and O2. The raw signals are sampled at 1000 Hz, with a total recording duration of approximately 162.12 hours and an effective labeled duration of approximately 162.03 hours. The recording duration varies across subjects, with the shortest recording lasting 6.96 hours, the longest 9.05 hours, and an average of approximately 8.11 hours per subject. Sleep stages are annotated into five categories according to standard clinical protocols, including Wake (W), N1, N2, N3, and REM. The overall label distribution is highly imbalanced, with W: 2968 (15.74%), N1: 685 (3.63%), N2: 8314 (44.09%), N3: 2814 (14.92%), and REM: 4075 (21.61%). **The collected real-world dataset has been approved by the institutional ethics committee, and informed consent was obtained from all participants.** This dataset constitutes a highly challenging evaluation scenario for pretrained models, characterized by (1) unseen task settings (sleep staging), (2) cross-subject variability, (3) long-duration continuous recordings, (4) severely imbalanced label distributions, and (5) channel configurations that differ significantly from those used in pretraining datasets. Such conditions provide a rigorous testbed for evaluating the robustness, transferability, and real-world applicability of EEG foundation models.

In the experiments, two evaluation protocols are implemented. The first protocol involves pretraining on all public datasets followed by fine-tuning on the sleep dataset, aiming to evaluate the transferability and generalization capability of the pretrained model. The second protocol directly performs supervised training on the sleep dataset, aiming to assess the performance of task-specific learning without pretraining. The experiments follow the cross-subject evaluation mode, where specific subjects are held out for testing while the remaining subjects are used for training, ensuring that the test subjects are completely unseen during training.

Table 8 presents the results on the real-world sleep EEG dataset. The results clearly demonstrate that pretraining on large-scale public EEG datasets enables the model to learn generalizable structural semantics, which can be effectively transferred to the unseen sleep staging task, leading to consistent performance improvements over training from scratch. Notably, BrainMamba achieves the best performance among all compared methods, highlighting its superior capability in modeling complex spatiotemporal dynamics and capturing fine-grained cross-channel interactions. This advantage becomes particularly evident in long-duration sleep recordings, where temporal dependencies span across multiple time scales and channel relationships are highly structured. The results indicate that the combination of structured state-space modeling and global spatiotemporal interaction is crucial for learning robust EEG representations. Furthermore, under the challenging cross-subject setting, where test subjects are completely unseen during training, BrainMamba maintains the best performance, demonstrating its robustness to inter-subject variability and domain shifts. This is especially important in real-world clinical scenarios, where subject-specific differences are inevitable.