

BRAINPRO: TOWARDS LARGE-SCALE BRAIN STATE-AWARE EEG REPRESENTATION LEARNING

Yi Ding^{1*}, Muyun Jiang^{1*}, Weibang Jiang^{1,2}, Shuailei Zhang¹, Xinliang Zhou¹,
Chenyu Liu¹, Shanglin Li³, Yong Li^{1,4}, & Cuntai Guan^{1,†}

¹Nanyang Technological University ²Shanghai Jiao Tong University

³Advanced Telecommunications Research Institute International ⁴Southeast University
ding.yi@ntu.edu.sg, ctguan@ntu.edu.sg

ABSTRACT

Electroencephalography (EEG) is a non-invasive technique for recording brain electrical activity, widely used in brain-computer interface (BCI) and healthcare. Recent EEG foundation models trained on large-scale datasets have shown improved performance and generalizability over traditional decoding methods, yet significant challenges remain. Existing models often fail to explicitly capture channel-to-channel and region-to-region interactions, which are critical sources of information inherently encoded in EEG signals. Due to varying channel configurations across datasets, they either approximate spatial structure with self-attention or restrict training to a limited set of common channels, sacrificing flexibility and effectiveness. Moreover, although EEG datasets reflect diverse brain states such as emotion, motor, and others, current models rarely learn state-aware representations during self-supervised pre-training. To address these gaps, we propose BrainPro, a large EEG model that introduces a retrieval-based spatial learning block to flexibly capture channel- and region-level interactions across varying electrode layouts, and a brain state-decoupling block that enables state-aware representation learning through parallel encoders with decoupling and region-aware reconstruction losses. This design allows BrainPro to adapt seamlessly to diverse tasks and hardware settings. Pre-trained on an extensive EEG corpus, BrainPro achieves state-of-the-art performance and robust generalization across nine public BCI datasets. Our codes and the pre-trained weights will be released.

1 INTRODUCTION

Electroencephalography (EEG) provides a non-invasive and cost-effective window into large-scale brain activity, supporting a wide range of applications in brain-computer interfaces (BCIs), cognitive neuroscience, and clinical neurotechnology. Despite its potential, EEG data are notoriously challenging to model due to low signal-to-noise ratio, non-stationarity, and variability across subjects and recording setups (Schalk et al., 2024; Wang et al., 2024b). Traditional EEG analysis methods relied on handcrafted features (e.g., spectral power, connectivity measures) tailored to specific tasks, but these approaches are labor-intensive and lack generalizability. With the advent of deep learning, supervised convolutional and recurrent neural networks have been applied to tasks such as motor imagery classification, sleep staging, and emotion recognition, but their reliance on large amounts of labeled data limits scalability (Jiang et al., 2024; Wang et al., 2025).

Inspired by the success of foundation models in language and vision (Devlin et al., 2018; He et al., 2022; Radford et al., 2021), researchers have recently proposed *EEG foundation models (EFMs)* trained on large-scale, unlabeled datasets using self-supervised learning (Zhou et al., 2025). These models aim to learn generalizable representations that transfer across tasks and datasets. For example, BIOT (Yang et al., 2023) pioneered cross-modal pretraining for biosignals, including EEG. LaBraM (Jiang et al., 2024) introduced discrete EEG tokens through vector quantization and trained encoders with masked signal modeling. CBraMod (Wang et al., 2025) incorporated spatial and temporal attention with dynamic positional encoding. EEGPT (Wang et al., 2024a) adopted Transformer-based architectures for large-scale EEG pretraining. Collectively, these approaches

demonstrate the promise of EFMs in enhancing EEG decoding performance and generalization beyond task-specific models. However, despite their progress, current EFMs face three fundamental limitations.

First, they underutilize *spatial interactions* between electrodes and brain regions. Most models either approximate spatial structure implicitly through self-attention (Jiang et al., 2024; Wang et al., 2025), or restrict to a fixed set of common channels (Yang et al., 2023), thereby sacrificing flexibility under heterogeneous montages. While CBraMod introduces spatial attention (Wang et al., 2025), explicit and flexible modeling of channel- and region-level dependencies remains limited.

Second, EEG signals inherently reflect diverse *brain states* in different brain processes, such as motor activity, emotion, and attention (Abiri et al., 2019), but existing EFMs pretrain a single encoder without explicitly disentangling brain state-related representations. This may restrict adaptability to downstream tasks probing distinct neurophysiological states.

Third, current EFMs employ a *single shared encoder*, which restricts downstream flexibility. In practice, many tasks involve overlapping or interacting processes (e.g., motor imagery engages both motor and emotional components) (Mane et al., 2020). A single encoder cannot selectively leverage such complementary signals. This limitation prevents EFMs from supporting flexible adaptation, where multiple specialized representations can be combined for richer downstream decoding.

To address these challenges, we propose **BrainPro**, a large EEG model designed to be **spatially adaptive, brain state-aware, and flexible in downstream usage**. BrainPro introduces (i) a *retrieval-based spatial learning block* that explicitly captures electrode- and region-wise interactions across arbitrary montages, and (ii) a *brain-state decoupling block* with parallel encoders, equipped with decoupling and region-aware reconstruction losses, to disentangle shared and process-specific representations. During downstream adaptation, BrainPro can activate the shared encoder together with one or more process-specific encoders, enabling flexible representation fusion. This design allows BrainPro to flexibly adapt to heterogeneous datasets, tasks, and hardware while capturing richer and more diverse representations than single-encoder EFMs.

Pre-trained on an extensive EEG corpus, BrainPro achieves **state-of-the-art performance** across nine public BCI datasets spanning six types of tasks. These results highlight its scalability, interpretability, and generalization ability, setting a new foundation for brain state-aware EEG representation learning.

Our contributions are summarized as follows:

- **Retrieval-based spatial learning.** We design a *retrieval-based spatial learning block* that explicitly models both channel-wise and region-wise interactions, enabling BrainPro to flexibly adapt across heterogeneous electrode montages and capture neurophysiological dependencies more effectively than prior EFMs, while also offering self-explainable spatial filters that provide neuroscientifically meaningful insights (seen in Appendix M).
- **Brain state-aware representation learning.** We introduce a *brain-state decoupling block* with parallel encoders, equipped with a decoupling loss and a region-aware reconstruction loss, to disentangle shared and process-specific representations, thereby enhancing adaptability across tasks reflecting diverse brain processes.
- **Flexible downstream adaptation.** Unlike single-encoder EFMs, BrainPro supports a mixture-of-experts style adaptation. During transfer, the shared encoder can be combined with one or more process-specific encoders, allowing richer and more diverse representations for complex downstream tasks that involve overlapping brain processes.
- **Extensive evaluation.** We pre-train BrainPro on a large-scale EEG corpus and evaluate it on nine public BCI datasets spanning motor, emotion, speech, stress, mental disease, and attention tasks. BrainPro consistently achieves state-of-the-art performance and demonstrates strong generalization across tasks, datasets, and hardware configurations.

2 METHOD

We propose **BrainPro**, a brain state-aware large EEG model with three key innovations: (i) it employs retrieval-based spatial learning to explicitly capture channel- and region-level dependencies,

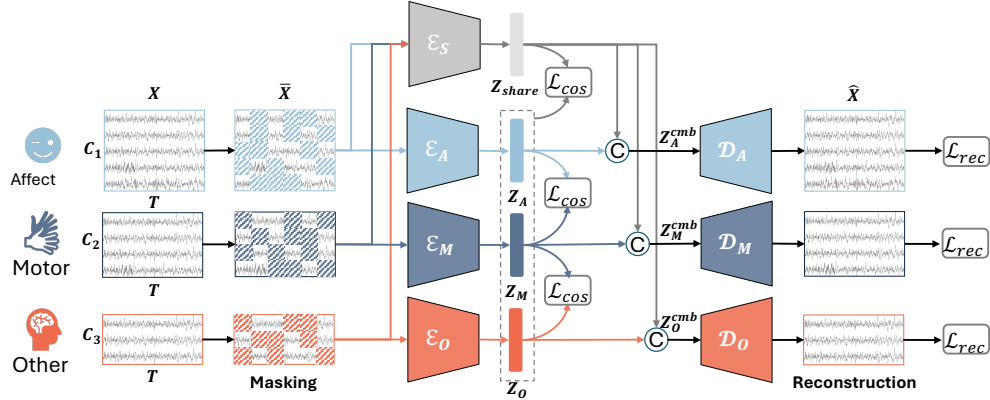


Figure 1: Overall BrainPro framework. BrainPro consists of a shared encoder, \mathcal{E}_S , for shared EEG representations and multiple brain state-specific encoders (\mathcal{E}_A for affect, \mathcal{E}_M for motor, and \mathcal{E}_O for others). Pre-training combines a region-aware masked reconstruction loss and a brain-state decoupling loss to learn disentangled and neurophysiology-guided representations.

(ii) it introduces parallel encoders with decoupling and region-aware reconstruction objectives to disentangle shared and brain-process-specific representations, and (iii) it supports flexible downstream adaptation by combining the shared encoder with one or more brain-state-specific encoders. The overall structure of BrainPro is shown in Figure 1. All encoders follow a common hierarchical design consisting of temporal encoding, retrieval-based spatial learning, patchification, and Transformer layers, detailed in Figure 2.

2.1 PROBLEM FORMULATION AND PIPELINE OVERVIEW

Let $\mathbf{X} \in \mathbb{R}^{C \times T}$ denote a raw EEG segment with C electrodes and T temporal samples. BrainPro aims to learn both shared and brain-state-specific representations

$$\mathbf{Z} = \mathbf{Z}_{\text{shared}} + \mathbf{Z}_{\text{state}}, \quad (1)$$

that are robust to variations in channel montage, brain region coverage, and task/state diversity.

As shown in Figure 1, a raw EEG signal is processed by (i) a *shared* encoder that learns universal features across tasks, and (ii) multiple *brain-state-specific* encoders that capture domain-relevant information (e.g., affect or motor). During pre-training, we combine a region-aware masked reconstruction loss with a cognition decoupling loss to enforce disentanglement between shared and brain-state-specific representations.

2.2 HIERARCHICAL ENCODER DESIGN

(1) Temporal Encoder: Local Dynamics Extraction. A temporal CNN extracts local, multi-scale dynamics along the time axis, independently for each channel. Each convolutional block consists of a 1D convolution followed by Group Normalization (GroupNorm), which normalizes activations within groups of channels to stabilize training with small batch sizes, and the Gaussian Error Linear Unit (GELU), a smooth nonlinear activation that weights inputs by their Gaussian cumulative distribution for improved expressivity. It can be formulated as

$$\mathbf{Z}^{(0)} = \mathbf{X}, \quad \mathbf{Z}^{(\ell)} = \text{GELU}\left(\text{GroupNorm}\left(\text{Conv1D}^{(\ell)}(\mathbf{Z}^{(\ell-1)})\right)\right), \quad \ell = 1, \dots, L_T. \quad (2)$$

where each $\text{Conv1D}^{(\ell)}$ applies K_T filters (kernel size k_ℓ , stride s_ℓ) along time. Padding is applied to keep the output size the same. The output $\mathbf{H}_{\text{temp}} = \mathbf{Z}^{(L_T)} \in \mathbb{R}^{K_T \times C \times T}$ preserves temporal resolution (T) while expanding the per-channel feature depth to K_T .

(2) Retrieval-based Spatial Learner: Flexible Spatial Modeling. To address the variability of EEG montages across datasets, we adopt a retrieval-based spatial learning strategy. We first define a universal template with $C_{\text{pre}}=60$ electrodes (based on the SEED dataset (Zheng & Lu, 2015),

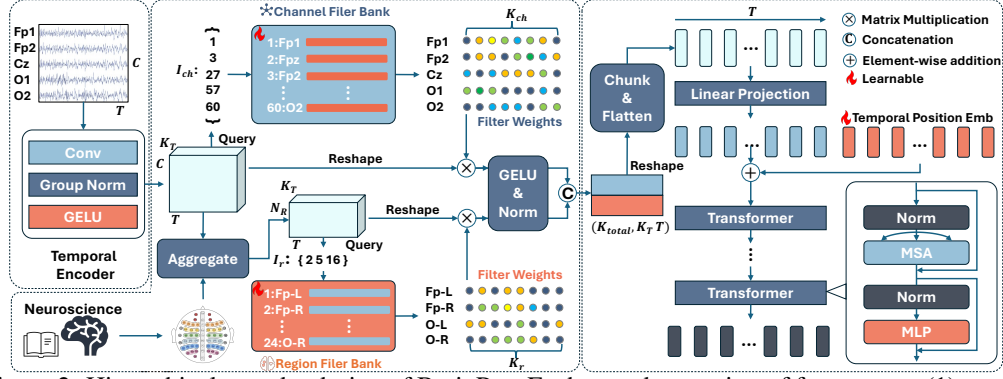


Figure 2: Hierarchical encoder design of BrainPro. Each encoder consists of four stages: (1) temporal encoder using channel-wise CNNs, (2) retrieval-based spatial learner with channel/region filter banks, (3) patchification and token embedding, and (4) Transformer layers for spatiotemporal modeling.

excluding reference channels) and $N_{\text{region}}=24$ functional regions following Ding et al. (2024). The detailed universal channel template and brain region definitions are provided in Appendix B. We maintain learnable filter banks aligned to the positions in the universal template for both channel- and region-level retrieval:

$$\mathbf{W}_C \in \mathbb{R}^{C_{\text{pre}} \times K_C}, \quad \mathbf{W}_R \in \mathbb{R}^{N_{\text{region}} \times K_R}, \quad (3)$$

where K_C and K_R denote the number of channel-wise and region-wise spatial filters, respectively. We reshape the \mathbf{H}_{temp} into $\mathbf{H}_{\text{temp}}^{\text{reshp}} \in \mathbb{R}^{C \times K_T T}$ for easy spatial learning.

Fine-grained channel features. Let $I_{\text{ch}} \subseteq \{1, \dots, C_{\text{pre}}\}$ be the set of indices in the universal template that are present in the current sample. This set is obtained by channel-name matching or, if unavailable, by nearest-neighbor mapping in standardized 3D head coordinates. We then retrieve the corresponding filters from the channel filter bank $\mathbf{W}_C[I_{\text{ch}}]$ and apply them to the temporal features:

$$\mathbf{H}_C = \sigma(\mathbf{W}_C[I_{\text{ch}}]^\top \mathbf{H}_{\text{temp}}^{\text{reshp}}) \in \mathbb{R}^{K_C \times K_T T}, \quad (4)$$

where $\sigma(\cdot)$ denotes normalization followed by activation (we use GELU+GroupNorm).

Coarse region features. We aggregate temporal features within each functional region present in the current sample and apply the region filter bank. Each channel has a corresponding brain region label from $I_r = \{r_1, \dots, r_C\}$ that specifies which functional region in the universal template it belongs to. Let $\mathcal{R}_{\text{uniq}}$ denote the set of unique functional regions covered in the sample, and $\mathcal{C}_j \subseteq I_r$ the subset of channels belonging to region j . For each region index $j \in \{1, \dots, |\mathcal{R}_{\text{uniq}}|\}$, we first average the temporal features of its constituent channels:

$$\mathbf{M}_{\text{region}}[j] = \frac{1}{|\mathcal{C}_j|} \sum_{c \in \mathcal{C}_j} \mathbf{H}_{\text{temp}}^{\text{reshp}}[c, :] \in \mathbb{R}^{1 \times K_T T}. \quad (5)$$

Then we retrieve the corresponding region spatial filters using $\mathcal{R}_{\text{uniq}}$. The region-wise representations are then processed by the region filter bank:

$$\mathbf{H}_R = \sigma(\mathbf{W}_R[\mathcal{R}_{\text{uniq}}]^\top \mathbf{M}_{\text{region}}) \in \mathbb{R}^{K_R \times K_T T}. \quad (6)$$

Finally, the fine- and coarse-level features are concatenated and reshaped to form the spatial representation:

$$\mathbf{H}_{\text{spatial}} = \text{Reshape}(\text{Concat}(\mathbf{H}_C, \mathbf{H}_R)) \in \mathbb{R}^{K_T K_{\text{total}} \times T}, \quad K_{\text{total}} = K_C + K_R. \quad (7)$$

(3) Patchification and Token Embedding. We partition $\mathbf{H}_{\text{spatial}}$ into N_p (possibly strided) temporal patches of length P :

$$\mathbf{Z}_{\text{patch}} \in \mathbb{R}^{N_p \times K_T K_{\text{total}} \times P}, \quad N_p = \left\lfloor \frac{T-P}{s_P} \right\rfloor + 1, \quad (8)$$

flatten each patch, and project to d -dimensional tokens with learnable embedding $\mathbf{E} \in \mathbb{R}^{d \times (K_T K_{\text{total}} P)}$ and positional embedding $\mathbf{p}_i \in \mathbb{R}^d$:

$$\mathbf{z}_i = \mathbf{E} \text{Flatten}(\mathbf{Z}_{\text{patch}}[i]) + \mathbf{p}_i, \quad i = 1, \dots, N_p. \quad (9)$$

Hence, we can get the tokens $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^{N_p} \in \mathbb{R}^{N_p \times d}$ for the Transformer layers to learn temporal contextual information.

(4) Transformer Encoder. To capture long-range spatiotemporal dependencies, each encoder applies L Transformer blocks. Each block follows a pre-norm structure consisting of multi-head self-attention (MSA) and a position-wise multilayer perceptron (MLP):

$$\mathbf{Z}'^{(\ell)} = \mathbf{Z}^{(\ell)} + \text{MSA}\left(\text{LN}(\mathbf{Z}^{(\ell)})\right), \quad (10)$$

$$\mathbf{Z}^{(\ell+1)} = \mathbf{Z}'^{(\ell)} + \text{MLP}\left(\text{LN}(\mathbf{Z}'^{(\ell)})\right), \quad \ell = 0, \dots, L-1, \quad (11)$$

where $\mathbf{Z}^{(0)} = \mathbf{Z}$ denotes the input patch embeddings and LN is layer normalization.

The MSA module models pairwise interactions among all patches by projecting $\mathbf{Z}^{(\ell)}$ into queries (Q), keys (K), and values (V), and computing

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V, \quad (12)$$

where d_k is the key dimension. Multiple attention heads are used in parallel and concatenated to form richer contextual representations. The MLP module consists of two fully connected layers with a hidden dimension d_{ff} and a nonlinear activation (GELU), applied independently to each token. Residual connections around both MSA and MLP stabilize training. After L blocks, we obtain the final token sequence $\mathbf{Z}^{(L)} \in \mathbb{R}^{N_p \times d}$.

2.3 PARALLEL ENCODERS FOR BRAIN-STATE DISENTANGLEMENT

A key innovation of BrainPro is the use of parallel encoders to disentangle shared and brain-state-specific representations. As shown in Figure 1, the model comprises one shared encoder $\mathcal{E}_{\text{shared}}$ and K brain-state-specific encoders $\{\mathcal{E}_k\}_{k=1}^K$ (e.g., affect, motor, etc.). In this work, we consider affect, motor and others three brain-states, $K=3$. Given an EEG segment \mathbf{X} , the encoders produce:

$$\mathbf{Z}_{\text{shared}} = \mathcal{E}_{\text{shared}}(\mathbf{X}), \quad \mathbf{Z}_k = \mathcal{E}_k(\mathbf{X}), \quad k = 1, \dots, K. \quad (13)$$

Selective gradient updates. During pre-training, each EEG sample is associated with a brain-state label y_{state} (e.g., affect). To enforce disentanglement, we update only the shared encoder \mathcal{E}_{S} and the corresponding brain-state-specific encoder $\mathcal{E}_{y_{\text{state}}}$. The outputs from all other encoders $\mathcal{E}_j, j \neq y_{\text{state}}$ are computed but detached from the computation graph, i.e.,

$$\mathbf{Z}_j = \text{stopgrad}(\mathcal{E}_j(\mathbf{X})), \quad j \neq y_{\text{state}}. \quad (14)$$

This ensures that only the relevant encoder pair receives gradient updates for each sample and we have the output from the other encoders for decoupling loss calculation.

Reconstruction with encoder fusion. To ensure that the masked reconstruction task leverages both shared and brain-state-specific information, we concatenate the outputs of the shared encoder and the active brain-state encoder and pass them jointly to a decoder:

$$\mathbf{Z}_{y_{\text{state}}}^{\text{comb}} = \text{Concat}(\mathbf{Z}_{y_{\text{state}}}, \mathbf{Z}_{\text{shared}}), \quad \hat{\mathbf{X}} = \mathcal{D}_{y_{\text{state}}}(\mathbf{Z}_{y_{\text{state}}}^{\text{comb}}), \quad (15)$$

where $\mathcal{D}_{y_{\text{state}}}$ is the decoder associated with brain state y_{state} . This design ensures that each brain-state encoder specializes in its domain, while the shared encoder captures shared brain-state patterns.

2.4 PRE-TRAINING OBJECTIVES

The pre-training objective of BrainPro is designed to (i) reconstruct masked EEG signals in a neurophysiology-aware manner, and (ii) disentangle shared from brain-state-specific representations through representation decoupling. The final loss combines these two components. The pre-training pipeline is illustrated in Algorithm 1.

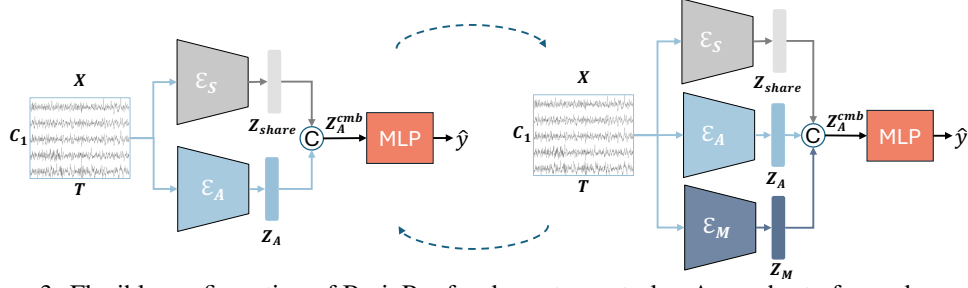


Figure 3: Flexible configuration of BrainPro for downstream tasks. Any subset of encoders can be selected and concatenated for task-specific adaptation. Emotion recognition is used for demonstration purpose.

Masking and reconstruction head. We randomly mask a subset of input patches with a mask ratio of $\rho=0.5$ (Jiang et al., 2024) and train the model to reconstruct the original EEG segment. A Transformer reconstruction head $\mathcal{D}_{y_{state}}(\cdot)$ takes the concatenated outputs of the shared encoder and the active brain-state encoder, and projects them back to the channel-time domain:

$$\hat{\mathbf{X}} = \mathcal{D}_{y_{state}}(\mathcal{E}_S(\bar{\mathbf{X}}) \parallel \mathcal{E}_{y_{state}}(\bar{\mathbf{X}})), \quad (16)$$

where masked EEG patches in $\bar{\mathbf{X}}$ are replaced with learned [MASK] embeddings.

Channel importance weights. To guide reconstruction toward neurophysiologically relevant areas, we incorporate brain-state-specific channel priors. For each brain state y_{state} , we define an importance vector $w^{(y_{state})} \in [0, 1]^{C_{pre}}$ over the universal template (e.g., frontal/temporal/central channels for affect; central/parietal channels for motor). For a dataset-specific montage, these priors are transferred to available channels. Instead of binary weighting, we apply a smooth weighting function:

$$\text{weights}(w) = 0.5 + \sigma(T \cdot (w - 0.5)), \quad (17)$$

where $\sigma(\cdot)$ is the logistic sigmoid and T is a sharpness parameter. We set T =current epoch to gradually increase the weighting. This smooth weighting ensures stable gradients during training.

Region-aware masked reconstruction loss. Let $M \subseteq \{1, \dots, C\} \times \{1, \dots, T\}$ denote the masked channel-time positions. The weighted MSE reconstruction loss is

$$\mathcal{L}_{\text{rec}} = \frac{1}{|M|} \sum_{(i,t) \in M} \text{weights}(w_i^{(y_{state})}) \cdot (X_{i,t} - \hat{X}_{i,t})^2. \quad (18)$$

This prioritizes accurate reconstruction of brain-state-relevant electrodes and regions.

Brain-state decoupling loss. To enforce representation disentanglement, we apply a margin-based cosine loss. Let $\mathbf{Z}_{\text{shared}}$ be the pooled shared representation, $\mathbf{Z}_{y_{state}}$ the active brain-state representation, and \mathbf{Z}_k the inactive ones. The loss penalizes high similarity between shared and state-specific features, and between active and inactive state encoders:

$$\mathcal{L}_{\text{dec}} = \max(\cos(\mathbf{Z}_{\text{shared}}, \mathbf{Z}_{y_{state}}) - m, 0) + \sum_{k \neq y_{state}} \max(\cos(\mathbf{Z}_{y_{state}}, \mathbf{Z}_k) - m, 0), \quad (19)$$

where m is a margin hyperparameter and we set it as 0.1.

Total loss. The final pre-training objective is a sum of reconstruction and decoupling losses:

$$\mathcal{L}_{\text{BrainPro}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{dec}}. \quad (20)$$

This objective simultaneously encourages neurophysiology-aware signal modeling and brain-state disentanglement.

Table 1: Overview of downstream BCI tasks and datasets.

BCI Task	Dataset	Sampling Rate	# Channels	Duration	# Samples	Label
I. Emotion Recognition	FACED	250Hz	32	10s	10,332	9-class
	SEED-V	1000Hz	62	1s	117,744	5-class
	SEED-VII	1000Hz	62	1s	281,679	7-class
II. Motor Imagery	BCI-IV-2A	250Hz	22	4s	5,088	4-class
	SHU-MI	250Hz	32	4s	11,988	2-class
III. Imagined Speech	BCIC2020-3	256Hz	64	4s	5,250	5-class
IV. Mental Disorder Diagnosis	Mumtaz2016	256Hz	19	4s	3,525	2-class
V. Mental Stress Detection	MentalArithmetic	500Hz	20	4s	1,080	2-class
VI. Mental Attention Detection	ATTEN	500Hz	28	4s	4,680	2-class

2.5 FLEXIBLE DOWNSTREAM ADAPTATION

For a downstream task, any subset \mathcal{S} of encoders can be activated and concatenated:

$$\mathbf{Z}_\star = \text{Concat}(\mathbf{Z}_{\text{shared}}, \{\mathbf{Z}_k\}_{k \in \mathcal{S}}), \quad \hat{y} = f_{\text{cls}}(\mathbf{Z}_\star). \quad (21)$$

Figure 3 illustrates flexible encoder selection and long-sequence handling. The experiments in Appendix K shows the effectiveness of such flexible configuration design. For downstream classification tasks, we construct these tokens into a segment-level vector $\bar{\mathbf{Z}} \in \mathbb{R}^d$ using either mean pooling, moving average (averaging every 5 tokens), or use all of them. A detailed comparison between different token merge mode can be find in Appendix J. We adopt the similar MLP head as Wang et al. (2025) with a hidden factor to adapt the hidden size in MLP.

3 EXPERIMENT SETUP OF PRE-TRAINING

3.1 PRE-TRAINING DATASETS AND PRE-PROCESSING

We construct a large-scale pre-training corpus by combining diverse EEG datasets spanning affective, motor, and clinical domains. To ensure that the data capture meaningful brain states, we follow a selection protocol similar to LaBraM (Jiang et al., 2024), while excluding certain datasets and incorporating additional motor imagery datasets. In total, the pre-training collection covers approximately **2,400 hours** of EEG recordings from a wide range of paradigms. The datasets are grouped into three categories for process-specific representation learning: (1) *affect*, (2) *motor*, and (3) *others*. A complete list of datasets and their categorization is provided in Appendix C.1.

All EEG recordings are segmented into 10-second clips and uniformly resampled to 200 Hz. We adopt a 60-channel montage based on the SEED series as the reference configuration, mapping dataset-specific channels according to scalp proximity. Unmapped or unreliable channels are discarded. Following Jiang et al. (2024), signals are scaled to 0.1 mV, resulting in values between -1 and 1 . To improve stability, noisy segments with absolute values greater than 10 are removed, yielding approximately **2,180 hours** of clean data. Samples are shuffled before training, and the samples in each batch are organized by dataset to reduce computational overhead in spatial mapping. Further implementation details are provided in Appendix C.2.

3.2 IMPLEMENTATION DETAILS

Pre-training is conducted end-to-end with a shared encoder and process-specific encoders/decoders. We train for 30 epochs using the AdamW optimizer with a cosine learning rate schedule, warmup, and gradient clipping. The model architecture consists of temporal and spatial encoders, patch makers, and transformer-based encoder-decoder modules. Hyperparameters are chosen to balance efficiency and representational capacity, with the full configuration provided in Appendix C.3 (Table 4).

4 EXPERIMENTAL SETUP OF DOWNSTREAM TASKS

4.1 DOWNSTREAM BCI TASKS AND DATASETS

To comprehensively evaluate our method, we consider six representative downstream BCI tasks using nine publicly available datasets: emotion recognition, motor imagery, imagined speech, mental disorder diagnosis, mental stress detection, and attention detection. The selected tasks and their corresponding datasets are summarized in Table 1, with additional dataset-specific details in Appendix D.2. For consistency with pre-training, all EEG signals are resampled to 200 Hz. Further preprocessing details are provided in Appendix D.2.

4.2 BASELINES AND EVALUATION METRICS

We compare BrainPro against both non-foundation and foundation model baselines. Among non-foundation approaches, we include **EEGNet** (Lawhern et al., 2018), a compact CNN tailored for EEG decoding, and **Conformer** (Song et al., 2023), which integrates CNNs and Transformers to capture local and global dependencies.

For foundation-model baselines, we consider: **BIOT** (Yang et al., 2023), which introduces domain-invariant attention and contrastive learning for cross-dataset generalization; **LaBraM** (Jiang et al., 2024), a large-scale pretrained transformer for universal EEG representations; **EEGPT** (Wang et al., 2024a), which leverages masked autoencoding with spatio-temporal alignment; and **CBraMod** (Wang et al., 2025), a transformer-based EEG foundation model with a criss-cross spatial-temporal architecture. Further baseline details are in Appendix D.3.

For evaluation metrics, we adopt the ones suitable for binary and multi-class tasks. For binary classification, we report **Balanced Accuracy (ACC-B)**, **AUC-PR**, and **AUROC**. For multi-class classification, following prior work (Jiang et al., 2024; Wang et al., 2025), we report **Balanced Accuracy (ACC-B)**, **Cohen’s Kappa**, and **Weighted F1 (F1-W)**. Together, these metrics provide a comprehensive evaluation across downstream tasks.

4.3 IMPLEMENTATION DETAILS

Each dataset is split into training, validation, and test subsets without overlap. Models are trained on the training set, with validation used for model selection and hyperparameter tuning. Final performance is reported on the test set after a single evaluation. To reduce randomness, we fix seeds $\{0, 1, 2, 3, 4\}$ and report the mean and standard deviation across five runs. All experiments are conducted on a cluster with 5 NVIDIA A800 GPUs (80 GB each). Detailed hyperparameter configurations are provided in Appendix D.1. After loading pre-trained weights, we re-initialize the temporal position embeddings with Xavier uniform initialization, similar to practices in NLP and vision, where task- or resolution-specific embeddings are re-initialized to improve adaptation. The corresponding ablation is reported in Appendix I.

5 RESULTS AND ANALYSIS

Table 2 compares methods across six EEG benchmarks. EEGNet and Conformer perform reasonably on simpler datasets but degrade on complex multi-class tasks, particularly in Kappa and F1. BIOT and EEGPT exhibit weaker transferability with unstable results, while LaBraM and CBraMod improve performance by leveraging larger-scale pre-training. BrainPro consistently outperforms all baselines, delivering the highest balanced accuracy and robust gains across metrics. For example, BrainPro achieves 0.5937/0.5418/0.6023 (ACC-B/Kappa/F1-W) on FACED and 0.8083/0.8980/0.8512 (ACC-B/AUC-PR/AUROC) on Mental Arithmetic. These results confirm that BrainPro not only improves accuracy but also yields more reliable and generalizable representations, validating its effectiveness as a large model for EEG decoding.

Table 2: Comparison results of different methods on downstream tasks.

Methods	FACED (9-Class)			SEED-V (5-Class)		
	ACC-B	Kappa	F1-W	ACC-B	Kappa	F1-W
EEGNet	0.3692 \pm 0.0103	0.2880 \pm 0.0113	0.3693 \pm 0.0111	0.2408 \pm 0.0031	0.0536 \pm 0.0042	0.1908 \pm 0.0088
Conformer	0.4566 \pm 0.0108	0.3849 \pm 0.0121	0.4555 \pm 0.0098	0.3087 \pm 0.0095	0.1294 \pm 0.0134	0.2837 \pm 0.0229
BIOT	0.2992 \pm 0.0119	0.2105 \pm 0.0147	0.2954 \pm 0.0149	0.3575 \pm 0.0052	0.1971 \pm 0.0081	0.3636 \pm 0.0078
EEGPT	0.2809 \pm 0.0116	0.1912 \pm 0.0129	0.2810 \pm 0.0120	0.2421 \pm 0.0048	0.0557 \pm 0.0077	0.2438 \pm 0.0055
LaBraM	0.5224 \pm 0.0116	0.4610 \pm 0.0126	0.5259 \pm 0.0108	0.3986 \pm 0.0200	0.2491 \pm 0.0256	0.4040 \pm 0.0197
CBraMod	0.5669 \pm 0.0094	0.5112 \pm 0.0110	0.5729 \pm 0.0105	0.3960 \pm 0.0033	0.2521 \pm 0.0048	0.4050 \pm 0.0052
BrainPro	0.5937 \pm 0.0087	0.5418 \pm 0.0092	0.6023 \pm 0.0061	0.4078 \pm 0.0075	0.2612 \pm 0.0089	0.4115 \pm 0.0071

Methods	BCI-IV-2a (4-Class)			SHU (2-Class)		
	ACC-B	Kappa	F1-W	ACC-B	AUC-PR	AUROC
EEGNet	0.5521 \pm 0.0183	0.4028 \pm 0.0244	0.5346 \pm 0.0228	0.5664 \pm 0.0522	0.6609 \pm 0.0099	0.6791 \pm 0.0085
Conformer	0.4879 \pm 0.0183	0.3171 \pm 0.0244	0.4561 \pm 0.0206	0.6167 \pm 0.0172	0.6763 \pm 0.0054	0.6927 \pm 0.0072
BIOT	0.2392 \pm 0.0280	-0.0144 \pm 0.0373	0.1063 \pm 0.0060	0.4981 \pm 0.0044	0.5077 \pm 0.0025	0.5069 \pm 0.0060
EEGPT	0.3849 \pm 0.0226	0.1799 \pm 0.0301	0.3272 \pm 0.0336	0.5481 \pm 0.0139	0.5631 \pm 0.0165	0.5673 \pm 0.0197
LaBraM	0.5255 \pm 0.0329	0.3674 \pm 0.0439	0.5095 \pm 0.0405	0.6175 \pm 0.0172	0.6821 \pm 0.0214	0.6720 \pm 0.0253
CBraMod	0.5148 \pm 0.0402	0.3530 \pm 0.0536	0.5032 \pm 0.0512	0.6108 \pm 0.0233	0.6632 \pm 0.0277	0.6668 \pm 0.0338
BrainPro	0.5674 \pm 0.0148	0.4232 \pm 0.0198	0.5653 \pm 0.0169	0.6319 \pm 0.0107	0.7102 \pm 0.0076	0.7105 \pm 0.0076

Methods	Mental Arithmetic (2-Class)			Attention (2-Class)		
	ACC-B	AUC-PR	AUROC	ACC-B	AUC-PR	AUROC
EEGNet	0.5533 \pm 0.0280	0.5702 \pm 0.0539	0.5717 \pm 0.0456	0.6004 \pm 0.0123	0.6294 \pm 0.0288	0.6647 \pm 0.0180
Conformer	0.6867 \pm 0.0492	0.7691 \pm 0.0286	0.7039 \pm 0.0177	0.7198 \pm 0.0158	0.7819 \pm 0.0218	0.7992 \pm 0.0283
BIOT	0.5583 \pm 0.0306	0.5512 \pm 0.0516	0.5662 \pm 0.0250	0.6111 \pm 0.0411	0.7273 \pm 0.0132	0.7367 \pm 0.0162
EEGPT	0.5650 \pm 0.0341	0.5860 \pm 0.0805	0.5937 \pm 0.0762	0.6674 \pm 0.0560	0.8015 \pm 0.0372	0.8103 \pm 0.0303
LaBraM	0.6688 \pm 0.0279	0.7504 \pm 0.0649	0.7168 \pm 0.0452	0.6785 \pm 0.0223	0.7838 \pm 0.0307	0.7994 \pm 0.0198
CBraMod	0.7354 \pm 0.0410	0.8237 \pm 0.0225	0.7654 \pm 0.0203	0.6478 \pm 0.0258	0.7417 \pm 0.0175	0.7468 \pm 0.0198
BrainPro	0.8083 \pm 0.0156	0.8980 \pm 0.0052	0.8512 \pm 0.0083	0.7222 \pm 0.0291	0.7975 \pm 0.0392	0.8064 \pm 0.0259

Note: **Bold** indicates the best performance. Cyan highlight marks our BrainPro.

Table 3: Ablation studies of the main components on downstream tasks.

Methods	BCI-IV-2A (4-Class)			Mental Arithmetic (2-Class)		
	ACC-B	Kappa	F1-W	ACC-B	AUC-PR	AUROC
w/o masking	0.4314 \pm 0.1160	0.2419 \pm 0.1547	0.4084 \pm 0.1532	0.7483 \pm 0.0785	0.8930 \pm 0.0131	0.8556 \pm 0.0134
w/o reconstruction	0.4354 \pm 0.0329	0.2472 \pm 0.0439	0.3758 \pm 0.0548	0.7083 \pm 0.0903	0.9147 \pm 0.0209	0.8976 \pm 0.0273
w/o decoupling	0.4870 \pm 0.0184	0.3160 \pm 0.0246	0.4725 \pm 0.0219	0.7317 \pm 0.0696	0.9102 \pm 0.0356	0.8860 \pm 0.0408
w random retrieval	0.5125 \pm 0.0308	0.3500 \pm 0.0410	0.5052 \pm 0.0379	0.7733 \pm 0.0410	0.8879 \pm 0.0288	0.8606 \pm 0.0343
w/o pre-training	0.4479 \pm 0.1066	0.2639 \pm 0.1421	0.4224 \pm 0.1454	0.7117 \pm 0.0889	0.8926 \pm 0.0083	0.8647 \pm 0.0109
BrainPro	0.5674 \pm 0.0148	0.4232 \pm 0.0198	0.5653 \pm 0.0169	0.8083 \pm 0.0156	0.8980 \pm 0.0052	0.8512 \pm 0.0083

Note: **Bold** indicates the best performance. Cyan highlight marks our BrainPro.

6 ABLATION STUDIES

Table 3 presents ablation results on BCI-IV-2A and Mental Arithmetic. Removing masking, reconstruction, or decoupling degrades performance, demonstrating their critical role in robust representation learning. Interestingly, omitting reconstruction improves AUC-PR and AUROC on Mental Arithmetic but harms BCI-IV-2A, suggesting reconstruction mainly enhances generalizability in complex multi-class tasks. Random channel retrieval and removing pre-training also reduce performance, emphasizing the value of structured retrieval and pre-trained initialization. BrainPro achieves the strongest results, with 0.5674/0.4232/0.5653 (ACC-B/Kappa/F1-W) on BCI-IV-2A and 0.8083/0.8980/0.8512 (ACC-B/AUC-PR/AUROC) on Mental Arithmetic. These findings validate the complementary roles of masking, reconstruction, decoupling, and pre-training, and show that their integration is essential for maximizing BrainPro’s effectiveness across diverse EEG decoding tasks.

7 CONCLUSION

In this work, we introduced **BrainPro**, a large EEG model that is spatially adaptive, brain state-aware, and flexible for downstream applications. Unlike prior EFMs, BrainPro explicitly captures electrode- and region-level interactions through a retrieval-based spatial learning block, and disentangles shared and process-specific representations using a brain-state decoupling block. This design allows BrainPro to flexibly adapt across heterogeneous montages, diverse tasks, and overlapping brain processes.

Extensive experiments on nine public BCI datasets demonstrate that BrainPro consistently achieves state-of-the-art performance, outperforming both conventional architectures and existing EFMs. Ablation studies further validate the importance of masking, reconstruction, decoupling, and pre-training, highlighting the complementary roles of each component.

By enabling spatial adaptivity, state-aware representation learning, and flexible downstream adaptation, BrainPro establishes a new foundation for EEG modeling. Moreover, the retrieval-based spatial learning block offers self-explainable spatial filters that reveal neuroscientifically meaningful patterns (seen in Appendix M), providing interpretability alongside strong performance. BrainPro will serve as a powerful backbone for future EEG-based applications in BCI and healthcare, and provide a scalable path toward more general and interpretable brain decoding.

REFERENCES

- Reza Abiri, Soheil Borhani, Eric W Sellers, Yang Jiang, and Xiaopeng Zhao. A comprehensive review of eeg-based brain–computer interface paradigms. *Journal of Neural Engineering*, 16(1): 011001, jan 2019.
- Clemens Brunner, Robert Leeb, and Gernot Müller-Putz. Bci competition 2008–graz data set a, 2024. URL <https://dx.doi.org/10.21227/katb-zv89>.
- Jingjing Chen, Xiaobin Wang, Chen Huang, Xin Hu, Xinke Shen, and Dan Zhang. A large finer-grained affective computing EEG dataset. *Scientific Data*, 10:740, 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02650-w.
- Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A. Joshi, and Richard M. Leahy. Neuro-gpt: Towards a foundation model for eeg. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2024. doi: 10.1109/ISBI56570.2024.10635453.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Yi Ding, Neethu Robinson, Su Zhang, Qiuhaio Zeng, and Cuntai Guan. Tsception: Capturing temporal dynamics and spatial asymmetry from eeg for emotion recognition. *IEEE Transactions on Affective Computing*, 14(3):2238–2250, 2023.
- Yi Ding, Neethu Robinson, Chengxuan Tong, Qiuhaio Zeng, and Cuntai Guan. Lggnet: Learning from local-global-graph representations for brain–computer interface. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):9773–9786, 2024. doi: 10.1109/TNNLS.2023.3236635.
- Yi Ding, Yong Li, Hao Sun, Rui Liu, Chengxuan Tong, Chenyu Liu, Xinliang Zhou, and Cuntai Guan. Eeg-deformer: A dense convolutional transformer for brain-computer interfaces. *IEEE Journal of Biomedical and Health Informatics*, 29(3):1909–1918, 2025. doi: 10.1109/JBHI.2024.3504604.
- Yunyuan Gao, Zhen Cao, Jia Liu, and Jianhai Zhang. A novel dynamic brain network in arousal for brain states and emotion analysis. *Mathematical Biosciences and Engineering*, 18(6):7440–7463, 2021.
- Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, 2022.

- Dong Huang, Cuntai Guan, Kai Keng Ang, Haihong Zhang, and Yaozhang Pan. Asymmetric spatial pattern for EEG-based emotion detection. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, 2012.
- Ji-Hoon Jeong, Jeong-Hyun Cho, Young-Eun Lee, Seo-Hyun Lee, Gi-Hwan Shin, Young-Seok Kweon, José del R Millán, Klaus-Robert Müller, and Seong-Whan Lee. 2020 international brain–computer interface competition: A review. *Frontiers in human neuroscience*, 16:898300, 2022.
- Wei-Bang Jiang, Li-Ming Zhao, Ping Guo, and Bao-Liang Lu. Discriminating Surprise and Anger from EEG and Eye Movements with a Graph Network. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1353–1357, 2021. doi: 10.1109/BIBM52615.2021.9669637.
- Wei-Bang Jiang, Xuan-Hao Liu, Wei-Long Zheng, and Bao-Liang Lu. Multimodal Adaptive Emotion Transformer with Flexible Modality Inputs on A Novel Dataset with Continuous Labels. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM ’23, pp. 5975–5984, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085.
- Wei-Bang Jiang, Xuan-Hao Liu, Wei-Long Zheng, and Bao-Liang Lu. Seed-vii: A multimodal dataset of six basic emotions with continuous labels for emotion recognition. *IEEE Transactions on Affective Computing*, 16(2):969–985, 2025. doi: 10.1109/TAFFC.2024.3485057.
- Weibang Jiang, Liming 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*, 2024.
- Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces. *Journal of Neural Engineering*, 15(5):056013, Jul 2018.
- Min-Ho Lee, O-Yeon Kwon, Yong-Jeong Kim, Hong-Kyung Kim, Young-Eun Lee, John Williamson, Siamac Fazli, and Seong-Whan Lee. Eeg dataset and openbmi toolbox for three bci paradigms: an investigation into bci illiteracy. *GigaScience*, 8(5):giz002, 01 2019. ISSN 2047-217X.
- Hongli Li, Man Ding, Ronghua Zhang, and Chunbo Xiu. Motor imagery eeg classification algorithm based on cnn-lstm feature fusion network. *Biomedical Signal Processing and Control*, 72:103342, 2022. ISSN 1746-8094.
- Rui Li, Le-Dian Liu, and Bao-Liang Lu. Discrimination of Decision Confidence Levels from EEG Signals. In *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 946–949, 2021. doi: 10.1109/NER49283.2021.9441086.
- Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing Recognition Performance and Robustness of Multimodal Deep Learning Models for Multimodal Emotion Recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 2021.
- Wei Liu, Wei-Long Zheng, Ziyi Li, Si-Yuan Wu, Lu Gan, and Bao-Liang Lu. Identifying similarities and differences in emotion recognition with EEG and eye movements among Chinese, German, and French People. *Journal of Neural Engineering*, 19(2):026012, 2022.
- F Lotte, M Congedo, A Lécuyer, F Lamarche, and B Arnaldi. A review of classification algorithms for eeg-based brain–computer interfaces. *Journal of Neural Engineering*, 4(2):R1, jan 2007.
- Matthew D Luciw, Ewa Jarocka, and Benoni B Edin. Multi-channel EEG recordings during 3,936 grasp and lift trials with varying weight and friction. *Scientific Data*, 1(1):1–11, 2014.
- Shuai Luo, Yu-Ting Lan, Dan Peng, Ziyi Li, Wei-Long Zheng, and Bao-Liang Lu. Multimodal emotion recognition in response to oil paintings. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 4167–4170, 2022. doi: 10.1109/EMBC48229.2022.9871630.

- Jun Ma, Banghua Yang, Wenzheng Qiu, Yunzhe Li, Shouwei Gao, and Xinxing Xia. A large eeg dataset for studying cross-session variability in motor imagery brain-computer interface. *Scientific Data*, 9(1):531, Sep 2022.
- Ravikiran Mane, Tushar Chouhan, and Cuntai Guan. Bci for stroke rehabilitation: motor and beyond. *Journal of Neural Engineering*, 17(4):041001, aug 2020.
- Nicola Michielli, U. Rajendra Acharya, and Filippo Molinari. Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals. *Computers in Biology and Medicine*, 106:71–81, 2019. ISSN 0010-4825.
- Katherine R Mickley Steinmetz and Elizabeth A Kensinger. The effects of valence and arousal on the neural activity leading to subsequent memory. *Psychophysiology*, 46(6):1190–1199, 2009.
- Wajid Mumtaz. MDD Patients and Healthy Controls EEG Data (New). 11 2016. doi: 10.6084/m9.figshare.4244171.v2. URL https://figshare.com/articles/dataset/EEG_Data_New/4244171.
- G. Pfurtscheller, C. Brunner, A. Schlögl, and F.H. Lopes da Silva. Mu rhythm (de)synchronization and eeg single-trial classification of different motor imagery tasks. *NeuroImage*, 31(1):153–159, 2006. ISSN 1053-8119.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.
- Arman Savran, Koray Ciftci, Guillaume Chanel, Javier Cruz Mota, Luong Hong Viet, Bülent Sankur, Lale Akarun, Alice Caplier, and Michele Rombaut. Emotion detection in the loop from brain signals and facial images. In *eINTERFACE’06-SIMILAR NoE Summer Workshop on Multimodal Interfaces*, 2006.
- 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.
- Gerwin Schalk, Peter Brunner, Brendan Z. Allison, Surjo R. Soekadar, Cuntai Guan, Tim Denison, Jörn Rickert, and Kai J. Miller. Translation of neurotechnologies. *Nature Reviews Bioengineering*, 2(8):637–652, Aug 2024.
- Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017.
- Vinit Shah, Eva Von Weltin, Silvia Lopez, James Riley McHugh, Lillian Veloso, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. The temple university hospital seizure detection corpus. *Frontiers in Neuroinformatics*, 12:83, 2018.
- Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3): 532–541, 2020. doi: 10.1109/TAFFC.2018.2817622.
- 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, 2023. doi: 10.1109/TNSRE.2022.3230250.
- Le-Yan Tao and Bao-Liang Lu. Emotion Recognition under Sleep Deprivation Using a Multimodal Residual LSTM Network. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2020. doi: 10.1109/IJCNN48605.2020.9206957.

- Mastaneh Torkamani-Azar, Sumeyra Demir Kanik, Serap Aydin, and Mujdat Cetin. Prediction of reaction time and vigilance variability from spatio-spectral features of resting-state EEG in a long sustained attention task. *IEEE Journal of Biomedical and Health Informatics*, 24(9):2550–2558, 2020.
- Logan Trujillo. Raw EEG Data. 2020. doi: 10.18738/T8/SS2NHB. URL <https://doi.org/10.18738/T8/SS2NHB>.
- Logan T Trujillo, Candice T Stanfield, and Ruben D Vela. The effect of electroencephalogram (EEG) reference choice on information-theoretic measures of the complexity and integration of EEG signals. *Frontiers in Neuroscience*, 11:425, 2017.
- L Veloso, J McHugh, E von Weltin, S Lopez, I Obeid, and J Picone. Big data resources for EEGs: Enabling deep learning research. In *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–3. IEEE, 2017.
- Eva von Weltin, Tameem Ahsan, Vinit Shah, Dawer Jamshed, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. Electroencephalographic slowing: A primary source of error in automatic seizure detection. In *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–5. IEEE, 2017.
- Guagnyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 39249–39280. Curran Associates, Inc., 2024a.
- 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. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Ping Wang, Aimin Jiang, Xiaofeng Liu, Jing Shang, and Li Zhang. Lstm-based eeg classification in motor imagery tasks. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(11):2086–2095, 2018. doi: 10.1109/TNSRE.2018.2876129.
- Yiming Wang, Bin Zhang, and Lamei Di. Research progress of eeg-based emotion recognition: A survey. *ACM Comput. Surv.*, 56(11), July 2024b. ISSN 0360-0300.
- Chaoqi Yang, M Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 78240–78260. Curran Associates, Inc., 2023.
- Daoze Zhang, Zhizhang Yuan, YANG YANG, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 26304–26321. Curran Associates, Inc., 2023.
- W. Zheng, W. Liu, Y. Lu, B. Lu, and A. Cichocki. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics*, pp. 1–13, 2018. ISSN 2168-2267. doi: 10.1109/TCYB.2018.2797176.
- 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. doi: 10.1109/TAMD.2015.2431497.
- Peixiang Zhong, Di Wang, and Chunyan Miao. Eeg-based emotion recognition using regularized graph neural networks. *IEEE Transactions on Affective Computing*, 13(3):1290–1301, 2022. doi: 10.1109/TAFFC.2020.2994159.
- Xinliang Zhou, Chenyu Liu, Zhisheng Chen, Kun Wang, Yi Ding, Ziyu Jia, and Qingsong Wen. Brain foundation models: A survey on advancements in neural signal processing and brain discovery, 2025.
- Igor Zyma, Sergii Tukaev, Ivan Seleznev, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1), 2019. ISSN 2306-5729.

A RELATED WORK

A.1 PREVIOUS METHODS FOR EEG DECODING

Traditional EEG decoding relied on handcrafted features and classical classifiers such as common spatial patterns (CSP), linear discriminant analysis (LDA), and support vector machines (SVMs), which required extensive domain expertise and were often task-specific (Lotte et al., 2007). With the rise of deep learning, neural architectures emerged to learn spatiotemporal EEG representations directly from raw signals (Schirrneister et al., 2017). CNN-based models captured spatial correlations between electrodes (Lawhern et al., 2018; Ding et al., 2023), while recurrent networks such as LSTMs modeled temporal dynamics (Wang et al., 2018; Michielli et al., 2019). Hybrid CNN-LSTM frameworks (Li et al., 2022) and attention-based architectures (Song et al., 2023; Ding et al., 2025) further advanced decoding performance, and graph neural networks were introduced to exploit functional brain connectivity (Song et al., 2020; Zhong et al., 2022; Ding et al., 2024). These approaches achieved strong results in tasks including motor imagery classification, seizure detection, emotion recognition, and sleep staging. However, they typically relied on supervised learning tailored to specific datasets, limiting scalability in the face of EEG’s heterogeneous channel configurations, variable sampling rates, and low signal-to-noise ratios.

A.2 EEG FOUNDATION MODELS

Inspired by the success of large language and vision models, recent studies have proposed *EEG foundation models* (EEG-FMs) pre-trained on large unlabeled corpora with self-supervised objectives (Zhou et al., 2025). These models aim to learn universal neural representations that can be adapted to diverse downstream BCI tasks. Brant (Zhang et al., 2023) and NeuroGPT (Cui et al., 2024), explored masked modeling and contrastive learning strategies to enhance cross-task generalization. BIOT (Yang et al., 2023) introduced a unified encoder that tokenizes heterogeneous biosignals into a sentence-like representation to support cross-dataset pre-training. LaBraM (Jiang et al., 2024) introduced a neural tokenizer and masked EEG modeling trained on 2,500 hours of data, achieving state-of-the-art results in abnormal detection, event classification, and emotion recognition. EEGPT (Wang et al., 2024a) developed a 10-million-parameter pretrained transformer with a dual mask-based self-supervised framework and spatio-temporal representation alignment to improve robustness under low SNR. CBraMod (Wang et al., 2025) employed a criss-cross transformer with asymmetric positional encoding to separately model spatial and temporal dependencies, achieving strong generalizability across diverse BCI datasets.

Despite recent progress, EEG-FMs still face key limitations: they approximate spatial interactions only implicitly or with fixed channels, lack disentanglement of diverse brain processes (e.g., motor, emotion), and rely on a single shared encoder that limits flexibility for tasks involving overlapping processes. These constraints reduce adaptability and highlight the need for models that can explicitly capture spatial dependencies, process-specific representations, and mixture-of-experts style flexibility.

B MORE DETAILS FOR THE BRAIN REGION DEFINITION

EEG captures activity from multiple functional areas of the brain. Treating electrodes as isolated nodes or relying only on global connections overlooks localized cooperative activity. LGGNet (Ding et al., 2024) addresses this by defining local brain regions and modeling both within-area activity (local graphs) and between-area interactions (global graphs), thereby capturing the brain’s hierarchical organization. The hemisphere LGG further extends this structure by introducing symmetric subgraphs across hemispheres, motivated by evidence of inter-hemispheric asymmetries in cognitive and emotional processes. This enables more effective modeling of bilateral patterns and asymmetries, which is beneficial for tasks such as emotion recognition and preference prediction. Following LGGNet, BrainPro adopts the hemisphere-based LGG region definition with minor adjustments: Fp1/Fp2 are separated from the AF regions, O1/O2 from the PO regions, and the FT, T, and TP regions are split into distinct areas. The brain regions are shown in Figure 4.

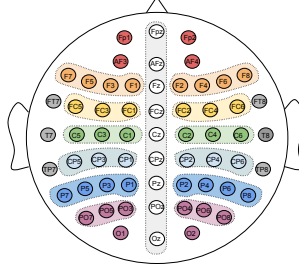


Figure 4: The brain region definition used in spatial learning block.

Algorithm 1 BrainPro Pre-training Pipeline

-
- 1: **Input:** EEG segment $\mathbf{X} \in \mathbb{R}^{C \times T}$, montage mapping, brain-state label y_{state}
 - 2: **Params:** Shared encoder \mathcal{E}_S , brain-state encoders $\{\mathcal{E}_k\}_{k=1}^K$, brain-state decoders $\{\mathcal{D}_k\}_{k=1}^K$, filter banks $\mathbf{W}_C, \mathbf{W}_R$, channel priors $w^{(y_{\text{state}})}$, mask ratio ρ
 - 3: **Step 1: Mask input**
 - 4: Randomly mask ρ fraction of patches in \mathbf{X} to obtain $\tilde{\mathbf{X}}$
 - 5: **Step 2: Parallel encoding**
 - 6: **for** each encoder $\mathcal{E}_j \in \{\mathcal{E}_S, \mathcal{E}_1, \dots, \mathcal{E}_K\}$ **do**
 - 7: Temporal encoding: $\mathbf{H}_{\text{temp}} \leftarrow \mathcal{F}_{\text{temp}}(\tilde{\mathbf{X}})$
 - 8: Spatial retrieval: $\mathbf{H}_{\text{spatial}} \leftarrow \mathcal{F}_{\text{spatial}}(\mathbf{H}_{\text{temp}}, \mathbf{W}_C, \mathbf{W}_R)$
 - 9: Patchify & embed: $\mathbf{Z} \leftarrow \text{PatchifyEmbed}(\mathbf{H}_{\text{spatial}})$
 - 10: Transformer encoding: $\mathbf{Z}_j \leftarrow \text{Transformer}(\mathbf{Z})$
 - 11: **if** $j \neq y_{\text{state}}$ and $j \neq \text{shared}$ **then**
 - 12: Detach \mathbf{Z}_j (inactive encoders, used only for decoupling loss)
 - 13: **end if**
 - 14: **end for**
 - 15: **Step 3: Reconstruction**
 - 16: Concatenate outputs: $\mathbf{Z}_{y_{\text{state}}}^{\text{comb}} \leftarrow \text{Concat}(\mathbf{Z}_{y_{\text{state}}}, \mathbf{Z}_{\text{shared}})$
 - 17: Reconstruct masked input: $\hat{\mathbf{X}} \leftarrow \mathcal{D}_{y_{\text{state}}}(\mathbf{Z}_{y_{\text{state}}}^{\text{comb}})$
 - 18: **Step 4: Loss computation**
 - 19: Region-aware reconstruction loss: $\mathcal{L}_{\text{rec}} \leftarrow \sum w^{(y_{\text{state}})} \cdot (\mathbf{X} - \hat{\mathbf{X}})^2$
 - 20: Brain-state decoupling loss: $\mathcal{L}_{\text{dec}} \leftarrow \text{margin-based cosine loss}$
 - 21: Total loss: $\mathcal{L}_{\text{BrainPro}} \leftarrow \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{dec}}$
 - 22: **Step 5: Parameter update**
 - 23: Update \mathcal{E}_S and $\mathcal{E}_{y_{\text{state}}}$ (detach all other encoders)
-

C MORE DETAILS FOR PRE-TRAINING

C.1 PRE-TRAINING DATASETS

A detailed description of the pre-training datasets for BrainPro is provided here. Most of these datasets overlap with those used for LaBraM, but we excluded some to ensure that the pre-training data represent meaningful brain states. In addition, we incorporated two extra motor imagery datasets to provide sufficient data for training each brain process encoder. In total, the datasets comprise around **2400 hours**.

- **Emobrain** (Savran et al., 2006): Multimodal emotion dataset with EEG (64 channels, 1024 Hz) and fNIRS from 16 subjects, collected via Biosemi Active 2. Emotions elicited with a subset of IAPS stimuli. (4.94 h)
- **Grasp and Lift EEG Challenge** (Luciw et al., 2014): EEG from 12 subjects (32 channels, 500 Hz) performing grasp-and-lift trials with a BrainAmp EEG amplifier. (11.72 h)

Table 4: Hyperparameters for pre-training.

Type	Factor	Value
Temporal Encoder	Input channels	{1, 32, 32}
	Output channels	{32, 32, 32}
	Kernel size	{15, 3, 3}
	Stride	{1, 1, 1}
	Padding	{7, 1, 1}
Spatial Encoder	Channel-wise Filter Number	32
	Region-wise Filter Number	32
	Total EEG Channel	60
	Total Brain Region	24
Patch Maker	Patch Length	20
	Patch Stride	20
	MLP size	64
Transformer Encoder	Layers	4
	Hidden size	32
	MLP size	64
	Attention heads	32
Transformer Decoder	Layers	2
	Hidden size	32
	MLP size	64
	Attention heads	32
Training	Batch size	160 (32 per GPU)
	Peak learning rate	1e-4
	Minimal learning rate	1e-5
	Scheduler	Cosine
	Optimizer	AdamW
	Adam β	(0.9, 0.98)
	Weight decay	0.05
	Total epochs	30
	Warmup epochs	2
	Gradient clipping	3
	Mask ratio	0.5

- **EEG Motor Movement/Imagery** (Schalk et al., 2004): Recordings from 109 volunteers (64 channels, 160 Hz) performing baseline (eyes open/closed), movement, and imagery (both fists/feet) tasks using the BCI2000 system. (47.3 h)
- **KU** (Lee et al., 2019): Data from 54 subjects conducting motor imagery tasks. It was recorded with 62 channels, 1000 Hz. (\sim 24.00 h)
- **HGD** (Schirrmeyer et al., 2017): Data from over 14 subjects, recorded 128 channels, 500 Hz. (\sim 7.49 h)
- **Raw EEG Data** (Trujillo, 2020): EEG (64 channels, 256 Hz) collected during information-integration and multidimensional rule-based categorization tasks. (34.35 h)
- **Resting State EEG Data** (Trujillo et al., 2017): EEG from 22 subjects during 8-minute resting (4 minutes eyes closed, 4 minutes eyes open), recorded with 64 channels at 256 Hz using BioSemi active Ag/AgCl electrodes. (3.04 h)
- **SEED Series** (Zheng & Lu, 2015; Zheng et al., 2018; Liu et al., 2022): Emotional EEG datasets including SEED (15 subjects), SEED-IV (15 subjects), SEED-GER (8 subjects), and SEED-FRA (8 subjects). Signals recorded at 1000 Hz from 62 channels with the ESI NeuroScan System while subjects viewed videos. (166.75 h)
- **SPIS Resting State** (Torkamani-Azar et al., 2020): EEG from 10 subjects (64 channels, 2048 Hz) with 2.5-minute eyes-open/closed sessions before a 105-minute sustained attention task. (0.83 h)

- **TUEP** (Veloso et al., 2017): Subset of TUEG with 100 epilepsy and 100 control subjects, verified by a neurologist; EEG recorded with 19–23 channels at 256 Hz. (591.22 h)
- **TUSZ** (Shah et al., 2018): Seizure-annotated EEG corpus (19–23 channels, 256 Hz) including onset, offset, channel, and seizure type information. (1138.53 h)
- **TUSL** (von Weltin et al., 2017): TUEG subset with slowing event annotations (23 channels, 256 Hz), used in seizure detection error analysis. (20.59 h)
- **SHJT EEG Data** (Jiang et al., 2023; 2021; Luo et al., 2022; Li et al., 2021; Tao & Lu, 2020): Data from over 140 subjects, recorded with the ESI NeuroScan System (62 channels, 1000 Hz). (342.23 h)

Table 5: Categories of the pre-training datasets.

Brain Process Category	Datasets
Affect	Emobrain, SEED, SEED-IV, SEED-GER, SEED-FRA, and SHJT EEG Data
Motor	Grasp and Lift EEG Challenge, EEG Motor Movement/Imagery, KU, and HGD
Others	Raw EEG Data, Resting State EEG Data, SPIS Resting State, TUEP, TUSZ, TUSL, and TUSL

C.2 EXPERIMENT SETTINGS

We describe the experimental setting for pre-training as follows. All EEG recordings are segmented into 10-second clips and resampled to 200 Hz. We predefine a 60-channel montage based on the SEED-series datasets (excluding two reference channels) as the reference configuration. Channels from different pre-training datasets are mapped to this predefined montage according to their spatial distance on the scalp. Channels that are absent from the predefined set or cannot be reliably mapped are discarded. To enable process-specific representation learning, we categorize the datasets into three groups, *affect*, *motor* and *other*, for brain-process decoupling training (see Table 5 for details). To make the pre-training stable, we discard the noisy segments that have large absolute values resulting around 785,000 clear samples, around 2,180 hours. These samples from different datasets are shuffled before being fed into the model. To improve training efficiency, samples from the same dataset are grouped within each training batch, ensuring that the spatial configuration requires only one spatial filter retrieval. During training, the gradients of samples belonging to different brain processes are passed to their corresponding encoders and decoders, while the shared encoder processes all samples to capture universal EEG representations.

C.3 HYPER-PARAMETERS

The pre-training of BrainPro is in an end-to-end manner, and the hyper-parameter settings are shown in Table 4.

C.4 PRE-TRAINING RESULTS

The image in Figure 5 shows the training loss curve of BrainPro during pre-training. The curve starts above 1.0 and drops sharply within the first few thousand steps, indicating fast initial convergence. After this rapid decline, the loss continues to decrease smoothly, reaching below 0.1 by the end of training. Only minor fluctuations are observed around the mid-stage, but the overall downward trajectory remains stable. This behavior demonstrates that BrainPro not only converges quickly but also maintains stable optimization, enabling the learning of robust and informative EEG representations during pre-training. We saved checkpoints at epochs 10, 20, and 30, and selected the one from epoch 10 based on validation performance on the FACED dataset. This checkpoint was then used for all remaining downstream datasets.

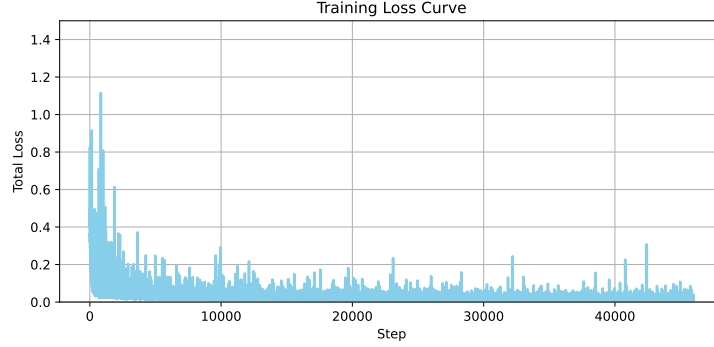


Figure 5: The pre-training loss of BrainPro.

D MORE DETAILS FOR DOWNSTREAM TASKS

D.1 HYPER-PARAMETERS

Table 6 lists the fine-tuned hyper-parameters for each downstream task, while the common training hyper-parameters are shown in Table 7.

Table 6: Tuned hyper-parameters for downstream tasks.

Dataset	Learning Rate	Hidden Factor	Dropout	Token Selection Mode
FACED	5e-4	8	0.1	All
SEED-V	3e-4	1	0.1	All
SEED-VII	3e-4	8	0.1	All
SHU	1e-4	1	0.1	Aggr
BCI-IV-2A	1e-3	1	0.2	Aggr
Mumtaz2016	1e-4	2	0.1	Aggr
Mental Arithmetic	1e-4	1	0.1	Mean
ATTEN	1e-3	4	0.1	Aggr
BCIC2020-3	5e-4	1	0.1	All

Table 7: Fixed hyper-parameters for fine-tuning.

Hyper-parameters	Settings
Epochs	50 (Others) / 30 (ATTEN)
Batch size	64
Optimizer	AdamW
Adam β	(0.9, 0.999)
Adam ϵ	1e-8
Weight decay	5e-2
Scheduler	CosineAnnealingLR
Cosine cycle epochs	50
Minimal learning rate	1e-6
Clipping gradient norm	1
Label smoothing (multi-class classification)	0.1

D.2 DOWNSTREAM DATASETS

We evaluate BrainPro on nine downstream datasets covering diverse BCI applications, including emotion recognition, motor imagery, imagined speech, mental disorder diagnosis, stress detection,

and attention decoding. These datasets differ substantially in subjects, channel configurations, sampling rates, trial durations, and class definitions, providing a comprehensive and challenging testbed for evaluating EEG foundation models. All EEG signals are resampled to 200 Hz, segmented into fixed-length windows according to task protocol, and the detailed training, validation, and test split is introduced in this section.

The **FACED** dataset (Chen et al., 2023) is used for the task of emotion recognition. It consists of 32-channel EEG recordings sampled at 250 Hz from 123 participants, each exposed to 28 video clips designed to elicit nine distinct emotional states: amusement, inspiration, joy, tenderness, anger, fear, disgust, sadness, and neutral. Each EEG trial lasts 10 seconds and is resampled to 200 Hz, yielding a total of 10,332 clean EEG segments. Following Wang et al. (2025), subjects 1–80 are used for training, 81–100 for validation, and 101–123 for testing, ensuring subject-independent evaluation.

The **SEED-V** dataset (Liu et al., 2021) is another benchmark for emotion recognition, covering five emotional categories: happy, sad, neutral, disgust, and fear. It contains 62-channel EEG recordings sampled at 1000 Hz from 16 subjects, each completing three sessions with 15 trials per session. The signals are segmented into 1-second windows and resampled to 200 Hz, resulting in 117,744 samples. Each session is evenly divided into training, validation, and testing subsets (5 trials each).

The **SEED-VII** dataset (Jiang et al., 2025) extends the SEED series to seven emotion categories: neutral, sad, fear, disgust, happy, surprise, and anger. It contains 62-channel EEG recordings at 1000 Hz from 20 subjects, each watching 80 film clips designed to elicit emotions in four sessions. Using the same pre-processing steps as SEED-V, EEG signals are segmented into 1-second windows and resampled to 200 Hz, producing 281,679 labeled samples. As there are 20 trials in each session, they are split at the trial level into training (10), validation (5), and test subsets (5).

The **BCI-IV-2A** dataset (Brunner et al., 2024) is a standard benchmark for motor imagery. It consists of EEG from 9 subjects recorded with 22 channels at 250 Hz while performing four classes of motor imagery: left hand, right hand, both feet, and tongue. Each trial lasts 4 seconds, and signals are resampled to 200 Hz. We applied a 0.1–70 Hz band-pass filter. In total, 5,088 trials are obtained. We follow subject-independent splits consistent with Wang et al. (2025), ensuring that train, validation, and test partitions are disjoint at the subject level.

The **SHU-MI** dataset (Ma et al., 2022) supports binary motor imagery classification. EEG was collected from 25 participants imagining either left- or right-hand movements, recorded with 32 channels at 250 Hz. The data are resampled to 200 Hz and segmented into 4-second non-overlapping windows, yielding 11,988 labeled samples. Following Wang et al. (2025), subjects 1–15 are used for training, 16–20 for validation, and 21–25 for testing.

The **BCIC2020-3** dataset (Jeong et al., 2022) focuses on imagined speech recognition. It includes EEG recordings from 64 channels at 256 Hz, where 15 participants imagine speaking one of five words. Each trial lasts 4 seconds and is resampled to 200 Hz, yielding 5,250 labeled samples. We use the official validation set as our test set, and the official training data are further divided into training and validation sets at a ratio of 9:1.

The **Mumtaz2016** dataset (Mumtaz, 2016) is designed for mental disorder diagnosis, specifically distinguishing patients with major depressive disorder (MDD) from healthy controls. EEG signals were recorded from 19 electrodes at 256 Hz, preprocessed with 0.3–75 Hz bandpass and 50 Hz notch filters, and then resampled to 200 Hz. Segments of 4 seconds are extracted, producing 3,525 labeled samples. We follow Wang et al. (2025) for subject-wise splits to ensure reliable evaluation.

The **Mental Arithmetic** dataset (Zyma et al., 2019) supports the task of mental stress detection. EEG was recorded from 36 subjects using 20 electrodes at 500 Hz, under two cognitive states: resting (“no stress”) and active mental arithmetic (“stress”). The data are resampled to 200 Hz, band-pass filtered (0.5–45 Hz), and segmented into 4-second windows, producing 1,080 labeled samples. We under-sample the class with more data samples to make the class balanced as Ding et al. (2024). Following Wang et al. (2025), subjects 1–28 are used for training, 29–32 for validation, and 33–36 for testing.

The **ATTEN** dataset is used for mental attention detection. It consists of EEG data from 26 subjects, recorded using 28 channels at 500 Hz during the Discrimination/Selection Response (DSR) task, which assesses cognitive attention. We adopt the pre-processing steps described in (Ding et al., 2025). The signals are resampled to 200 Hz and segmented into 4-second non-overlapping windows,

yielding 4,680 labeled samples. Subjects 1–20 are used as the training set, subjects 21–23 as the validation set, and the last 3 subjects as the test set.

D.3 BASELINES

EEGNet (Lawhern et al., 2018): It is a lightweight convolutional neural network tailored for EEG-based BCI applications. By employing depthwise and separable convolutions, it efficiently extracts discriminative EEG features, enabling effective and computationally efficient EEG decoding.

Conformer (Song et al., 2023): It integrates CNNs with Transformers to model the spatio-temporal characteristics of EEG signals. It employs 1-D CNNs to extract local features and leverages a self-attention mechanism to capture global temporal dependencies.

BIOT (Yang et al., 2023): It is a transformer-based model developed to address cross-dataset EEG classification challenges under domain shifts. It incorporates a domain-invariant attention mechanism alongside contrastive representation learning, which together improve generalization across diverse recording setups and subject groups.

LaBraM (Jiang et al., 2024): It introduces a scalable transformer framework designed to learn universal EEG representations from large-scale brain signal datasets. Through pretraining on a broad range of EEG recordings, the model captures comprehensive temporal and spatial characteristics that can be effectively transferred to downstream BCI tasks. Its architecture integrates efficient self-attention operations with task-specific adapters, supporting flexible fine-tuning for varied applications.

EEGPT Wang et al. (2024a): It adopts a dual self-supervised learning paradigm that combines masked autoencoding with spatio-temporal representation alignment. By emphasizing high signal-to-noise ratio (SNR) features rather than raw inputs, it improves the quality of learned representations. The model employs a hierarchical design that decouples spatial and temporal processing, thereby enhancing both computational efficiency and adaptability across BCI scenarios.

CBraMod Wang et al. (2025): It is a transformer-based EEG foundation model tailored to capture the complex spatial and temporal dependencies in EEG data. It introduces a criss-cross transformer structure with parallel spatial and temporal attention modules, enabling simultaneous yet independent modeling of spatial and temporal dynamics.

D.4 EVALUATION METRICS

In this section, we describe the evaluation metrics adopted in this work. Following LaBraM (Jiang et al., 2024; Wang et al., 2025), we employ the following measures:

- **Balanced Accuracy:** Accounts for class imbalance by computing the average recall across all classes. It is applied in both binary and multi-class classification settings.
- **AUC-PR:** The area under the precision–recall (PR) curve, used to evaluate performance in binary classification tasks.
- **AUROC:** The area under the receiver operating characteristic (ROC) curve, a standard metric for assessing binary classification performance.
- **Cohen’s Kappa:** A statistical measure of inter-rater agreement, commonly applied to imbalanced multi-class classification problems to quantify consistency beyond chance.
- **Weighted F1:** The weighted average of class-wise F1-scores, where each class is weighted by its sample size. This provides a robust assessment of performance in multi-class classification tasks.

E MORE EXPERIMENTAL RESULTS

E.1 SEED-VII

On the SEED-VII dataset (Table 8), we observe lower overall performance compared to FACED or SEED-V, highlighting the challenge of distinguishing seven closely related emotions under high

inter-subject variability. Simpler models such as EEGNet and EEGPT perform poorly, while Conformer and BIOT provide moderate improvements. Foundation models achieve stronger results, with LaBraM performing best overall and CBraMod showing competitive accuracy. BrainPro reaches 0.3315/0.2223/0.3350 (ACC-B/Kappa/F1-W), closely matching CBraMod and LaBraM. Although LaBraM slightly outperforms BrainPro, the margin is small, and BrainPro maintains stable generalization, confirming its effectiveness even in this demanding multi-class setting. This suggests that BrainPro’s retrieval-based spatial modeling and state-decoupling contribute to consistent robustness across diverse emotion recognition tasks.

Table 8: Results on the SEED-VII dataset.

Methods	ACC-B	Kappa	F1-W
EEGNet	0.1892 \pm 0.0056	0.0562 \pm 0.0071	0.1472 \pm 0.0123
Conformer	0.2719 \pm 0.0032	0.1566 \pm 0.0039	0.2712 \pm 0.0033
BIOT	0.3005 \pm 0.0054	0.1859 \pm 0.0070	0.3055 \pm 0.0069
EEGPT	0.2213 \pm 0.0060	0.0954 \pm 0.0071	0.2147 \pm 0.0120
LaBraM	0.3346 \pm 0.0122	0.2256 \pm 0.0146	0.3391 \pm 0.0121
CBraMod	0.3318 \pm 0.0056	0.2236 \pm 0.0065	0.3384 \pm 0.0059
BrainPro	0.3315 \pm 0.0038	0.2223 \pm 0.0055	0.3350 \pm 0.0045

Note: **Bold** indicates the best performance. Cyan highlight marks our BrainPro.

E.2 MENTAL DISORDER DIAGNOSIS

On the MDD dataset, all methods achieve relatively high accuracy, suggesting that distinguishing patients from controls is more tractable than fine-grained emotion recognition. EEGNet and Conformer already reach strong performance, while BIOT and LaBraM offer further improvements. BrainPro achieves 0.9161/0.9831/0.9803 (ACC-B/AUC-PR/AUROC), outperforming all baselines across metrics. These results confirm that BrainPro provides robust and reliable representations for clinical EEG applications, where high sensitivity and specificity are crucial.

Table 9: Results on the Major Depressive Disorder dataset.

Methods	ACC-B	AUC-PR	AUROC
EEGNet	0.9113 \pm 0.0104	0.9632 \pm 0.0045	0.9512 \pm 0.0096
Conformer	0.8422 \pm 0.0344	0.9442 \pm 0.0502	0.9613 \pm 0.0064
BIOT	0.8789 \pm 0.0190	0.9744 \pm 0.0083	0.9664 \pm 0.0136
EEGPT	0.8475 \pm 0.0933	0.9695 \pm 0.0076	0.9669 \pm 0.0069
LaBraM	0.8986 \pm 0.0018	0.9791 \pm 0.0041	0.9754 \pm 0.0050
CBraMod	0.8909 \pm 0.0037	0.9769 \pm 0.0047	0.9726 \pm 0.0062
BrainPro	0.9161 \pm 0.0054	0.9831 \pm 0.0038	0.9803 \pm 0.0046

Note: **Bold** indicates the best performance. Cyan highlight marks our BrainPro.

E.3 SPEECH

On the imagined speech dataset, performance is generally modest across methods, reflecting the inherent difficulty of decoding internally generated speech from EEG. Simpler models such as EEGNet show very low balanced accuracy, while Conformer and EEGPT achieve only small gains. Foundation models perform better, with LaBraM and CBraMod leading the results. BrainPro achieves 0.5253/0.4067/0.5244 (ACC-B/Kappa/F1-W), the highest among all methods, indicating that retrieval-based spatial learning and state-decoupling effectively capture subtle neural dynamics underlying imagined speech.

Table 10: Results on the Speech dataset.

Methods	ACC-B	Kappa	F1-W
EEGNet	0.2755 \pm 0.0183	0.0943 \pm 0.0228	0.2737 \pm 0.0180
Conformer	0.3339 \pm 0.0104	0.1673 \pm 0.0130	0.3269 \pm 0.0153
BIOT	0.3016 \pm 0.0206	0.1270 \pm 0.0257	0.2999 \pm 0.0217
EEGPT	0.2787 \pm 0.0061	0.0983 \pm 0.0076	0.2758 \pm 0.0060
LaBraM	0.4880 \pm 0.0161	0.3600 \pm 0.0201	0.4871 \pm 0.0165
CBraMod	0.5061 \pm 0.0211	0.3827 \pm 0.0263	0.5050 \pm 0.0212
BrainPro	0.5253 \pm 0.0084	0.4067 \pm 0.0105	0.5244 \pm 0.0090

Note: **Bold** indicates the best performance. Cyan highlight marks our BrainPro.

F EFFECTS OF LEARNING RATE

Figure 6 shows how different learning rates affect BrainPro on BCI-IV-2a and Mental Arithmetic. On BCI-IV-2a, the highest discriminative performance is achieved at a relatively larger learning rate (1×10^{-3}), whereas very small values reduce accuracy and kappa, suggesting that sufficient step size is needed for stable convergence on motor-related tasks. In contrast, Mental Arithmetic benefits from smaller learning rates, with accuracy gradually improving as the rate decreases, while global metrics such as AUC-PR and AUROC remain largely stable. Although these results are based on only two datasets, they highlight that learning rate can influence fine-tuning in a dataset-dependent manner, and that adapting optimization strategies to task characteristics may further enhance BrainPro’s generalization ability.

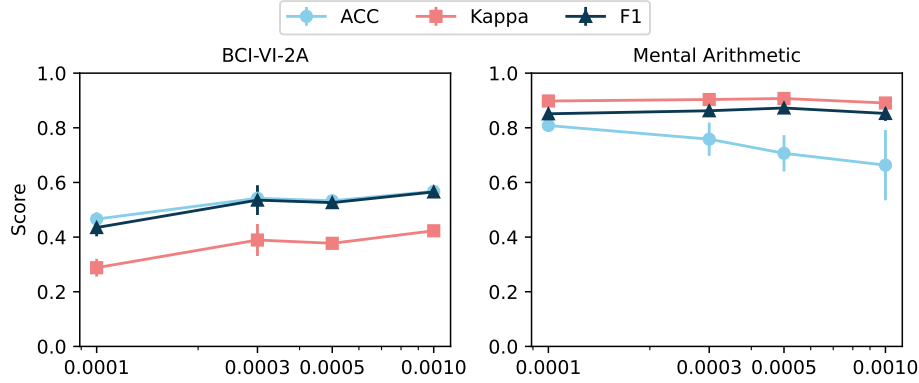


Figure 6: Effects of learning rate on BrainPro pre-training.

G EFFECTS OF MODEL SIZE

Figure 7 shown the effects on performance with varying model size using FACED dataset. We evaluated models of varying sizes ranging from 10.04M to 33.45M parameters. The results indicate a general trend of performance improvement with increasing model size up to around 28.25M parameters, where the highest scores are observed across ACC-B (0.5980), Kappa (0.5458), and F1-W (0.6034). Beyond this point, at 33.45M, the performance plateaus or slightly declines, suggesting diminishing returns from further scaling. Standard deviations are relatively low for larger models compared to the smallest configuration (10.04M), implying more stable training outcomes as the model grows. Overall, these findings suggest that moderately larger models (about 28M parameters) strike the best balance between performance and stability for the FACED dataset.

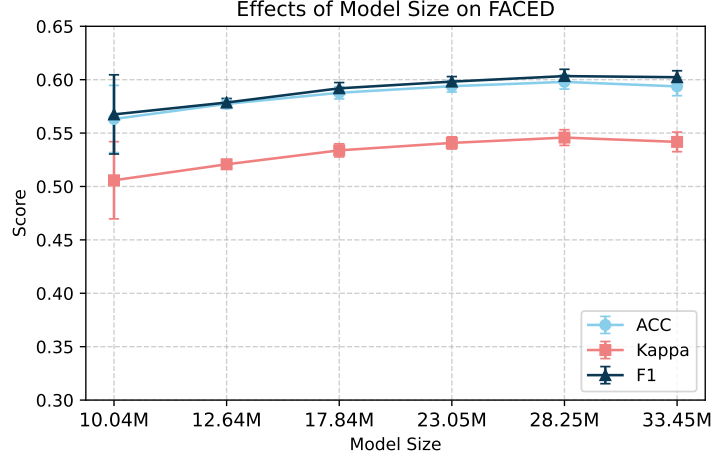


Figure 7: Effect of Model Size on Performance (FACED).

H EFFECTS OF CLEAR DATA PRE-TRAINING

Figure 8 illustrates the effect of noisy samples during pre-training. On the BCI-IV-2a dataset, models pre-trained with clear data outperform those trained with noisy data across all classification metrics, with substantial gains in accuracy, Cohen’s kappa, and F1-score. On the Mental Arithmetic dataset, the differences are less pronounced, with only moderate improvements in AUC-PR and AUROC, and in some cases noisy pre-training yields comparable results. These findings suggest that while global performance metrics are relatively robust to noise, discriminability is significantly compromised when noisy data are included in pre-training. This emphasizes that high-quality pre-training data is critical for enabling BrainPro to learn robust and generalizable EEG representations.

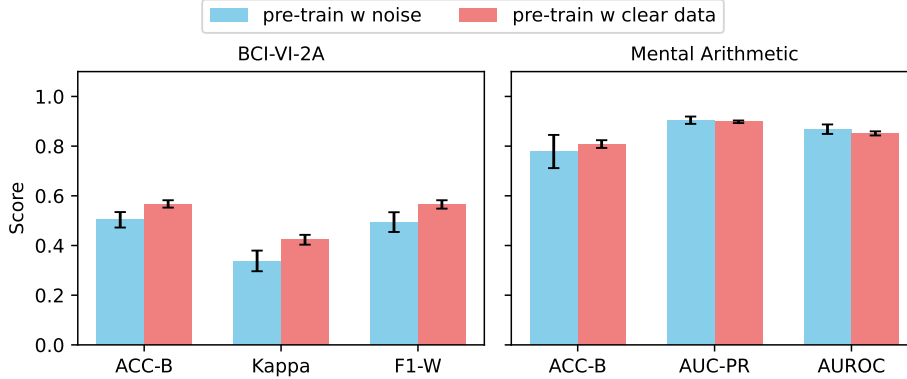


Figure 8: Effects of noisy samples on BrainPro pre-training.

I EFFECTS OF RESETTING THE TEMPORAL POSITION EMBEDDING

Table 11 reports the results of resetting vs. not resetting temporal position embeddings on both BCI-VI-2A and Mental Arithmetic datasets. The results show that resetting consistently improves performance across all metrics. On BCI-VI-2A, resetting yields notable gains, with ACC-B increasing from 0.5222 to 0.5674 and Kappa improving from 0.3630 to 0.4232, indicating better discriminative ability and robustness. Similarly, for the Mental Arithmetic dataset, resetting provides smaller but consistent improvements across ACC-B, AUC-PR, and AUROC.

The resetting strategy is applied after loading the pre-trained weights by re-initializing the temporal position embeddings with a Xavier uniform distribution. Importantly, after resetting, the embeddings remain learnable and are optimized during downstream fine-tuning. This ensures that the model retains the ability to encode task-specific temporal dynamics, while avoiding potential misalignment introduced by directly transferring positional priors from pre-training. The motivation behind this strategy lies in the observation that, although pre-training captures general brain state patterns, the temporal ordering of such states may vary substantially across unseen downstream tasks. Resetting the embeddings allows the model to relearn temporal structures tailored to the target dataset, which likely explains the consistent improvements observed

Table 11: Effect of resetting temporal position embedding on BCI-VI-2A and Mental Arithmetic datasets.

Setting	BCI-VI-2A			Mental Arithmetic		
	ACC-B	Kappa	F1-W	ACC-B	AUC-PR	AUROC
w/o reset	0.5222 \pm 0.0406	0.3630 \pm 0.0541	0.5109 \pm 0.0490	0.8017 \pm 0.0122	0.8955 \pm 0.0113	0.8471 \pm 0.0184
reset	0.5674 \pm 0.0148	0.4232 \pm 0.0198	0.5653 \pm 0.0169	0.8083 \pm 0.0156	0.8980 \pm 0.0052	0.8512 \pm 0.0083

Note: Bold values indicate superior performance compared to the non-reset baseline.

J EFFECTS OF DIFFERENT TOKEN SELECTION MODE

Table 12 compares different token selection strategies for constructing segment-level representations. On BCI-VI-2a, the aggregation-based strategy (“Aggr”) achieves the best performance across all three metrics, outperforming both mean pooling and using all tokens, which suggests that aggregation provides a more compact and discriminative summary of motor imagery signals. In contrast, on the Mental Arithmetic dataset, mean pooling yields the best results on all metrics, with substantial gains in AUC-PR and AUROC, highlighting that a simpler averaging strategy is more effective for cognitive-related signals. Using all tokens performs worse on both datasets, likely due to redundancy and noise accumulation. These results demonstrate that the optimal token merging strategy is dataset-dependent, and that BrainPro’s flexible design allows for effective adaptation across different EEG domains.

Table 12: Comparison of different token selection strategies on BCI-VI-2A and Mental Arithmetic datasets.

Token Merge	BCI-VI-2A			Mental Arithmetic		
	ACC-B	Kappa	F1-W	ACC-B	AUC-PR	AUROC
Aggr	0.56736 \pm 0.01482	0.42315 \pm 0.01976	0.56531 \pm 0.01691	0.73667 \pm 0.03151	0.83525 \pm 0.01750	0.78222 \pm 0.01410
Mean	0.40903 \pm 0.04703	0.21204 \pm 0.06272	0.36640 \pm 0.07028	0.80833 \pm 0.01559	0.89800 \pm 0.00520	0.85122 \pm 0.00827
All	0.53733 \pm 0.03154	0.38310 \pm 0.04205	0.52952 \pm 0.03350	0.58833 \pm 0.03466	0.67525 \pm 0.01305	0.62933 \pm 0.01600

Note: Bold indicates the best performance for each dataset.

K FLEXIBLE CONFIGURATION OF ENCODERS FOR DOWNSTREAM TASKS

As shown in Table 13, even with a single state-specific encoder (with \mathcal{E}_A or \mathcal{E}_M), BrainPro achieves state-of-the-art results on the FACED dataset and performance comparable to existing SOTA methods on BCI-IV-2a. This confirms the strength of our pre-training and demonstrates that BrainPro’s encoders are effective even in isolation. Incorporating the shared encoder \mathcal{E}_S yields consistent gains, particularly on BCI-IV-2a, highlighting its role in capturing global cross-state information. The full combination of affect-, motor-, and shared-encoders achieves the best results overall, validating our motivation from Section 1: unlike prior EFMs that rely on a single encoder, BrainPro’s brain state-decoupling block enables disentangled yet complementary representations, and its compositional architecture provides the flexibility to adaptively fuse them for different tasks. These findings show that BrainPro addresses the limitations of existing EFMs by being both state-aware and flexibly adaptable across heterogeneous EEG datasets.

Table 13: Effects of different encoder configurations on FACED and BCI-VI-2A datasets.

Methods	FACED			BCI-VI-2A		
	ACC-B	Kappa	F1-W	ACC-B	Kappa	F1-W
w \mathcal{E}_A (\mathcal{E}_M)	0.58186 \pm 0.01187	0.52677 \pm 0.01286	0.58348 \pm 0.01030	0.50590 \pm 0.02723	0.34121 \pm 0.03630	0.49866 \pm 0.02916
w \mathcal{E}_A (\mathcal{E}_M) + \mathcal{E}_S	0.59372 \pm 0.00873	0.54176 \pm 0.00917	0.60231 \pm 0.00605	0.56736 \pm 0.01482	0.42315 \pm 0.01976	0.56531 \pm 0.01691
w \mathcal{E}_A + \mathcal{E}_M + \mathcal{E}_S	0.59882 \pm 0.00569	0.54770 \pm 0.00575	0.60979 \pm 0.00425	0.57969 \pm 0.02291	0.43958 \pm 0.03054	0.57637 \pm 0.02454

Note: **Bold** indicates the best performance. Cyan highlight marks the BrainPro default configuration.

Table 14: Comparison of model complexity in terms of parameters and FLOPs.

Method	Parameters	FLOPs
EEGNet	0.005M	10.03M
Conformer	0.17M	50.09M
BIOT	3.28M	229.35M
EEGPT	51.66M	9.31G
LaBraM	9.43M	392.68M
CBraMod	8.45M	350.79M
BrainPro	7.69M	531.04M

L PARAMETERS SIZE AND FLOPS

Table 14 compares the parameter counts and FLOPs of different EEG models, using the BCI-IV-2A dataset as the example for evaluation. Lightweight baselines such as EEGNet and Conformer have very few parameters (0.005M and 0.17M, respectively) and correspondingly low FLOPs, making them efficient but limited in representational capacity. In contrast, LaBraM, BIOT, EEGPT, CBraMod, and our proposed BrainPro are representative large EEG models, with parameter counts ranging from 3M to over 50M and FLOPs spanning hundreds of millions to several billions. Among them, LaBraM is reported with all tokens enabled, since this configuration achieves stronger performance despite higher computational cost. For a fair comparison, the MLP structure and token selection strategy are kept consistent across all foundation models. Notably, BrainPro contains 7.69M parameters and requires 531.04M FLOPs, positioning it as a mid-sized large model that balances complexity and efficiency. Compared to EEGPT (51.66M parameters, 9.31G FLOPs), BrainPro achieves competitive model capacity at a fraction of the computational cost, demonstrating a favorable trade-off between scalability and practicality.

M VISUALIZATION OF LEARNED SPATIAL FILTERS IN PRE-TRAINING

The visualization of the learned spatial filters highlights the self-explainable nature of the retrieval-based spatial learning framework. As shown in Figure 9, the affect encoders predominantly emphasize frontal and temporal regions, which are consistent with neural substrates associated with emotional processing (Gao et al., 2021; Huang et al., 2012; Mickley Steinmetz & Kensinger, 2009), whereas the motor encoders focus on central and parietal regions, aligning with sensorimotor activity (Pfurtscheller et al., 2006). In contrast, the shared encoders exhibit more diverse and distributed spatial patterns, suggesting that they capture cross-domain representations that generalize across both affective and motor states. These results demonstrate that the proposed pre-training strategy learns spatial filters that not only support downstream performance but also provide neuroscientifically meaningful insights into the types of information encoded.

N DISCUSSION

New insights. This work demonstrates that explicitly incorporating brain state awareness into Large EEG models yields consistent gains across diverse EEG decoding tasks. Unlike prior single-encoder EFMs, BrainPro disentangles shared and brain state-specific representations through selective gradient updates and decoupling objectives. This design not only improves performance but also provides a flexible mixture-of-experts style adaptation mechanism that allows downstream models to combine relevant encoders as needed. The retrieval-based spatial learner further reveals that explicitly modeling channel- and region-level dependencies across heterogeneous montages can

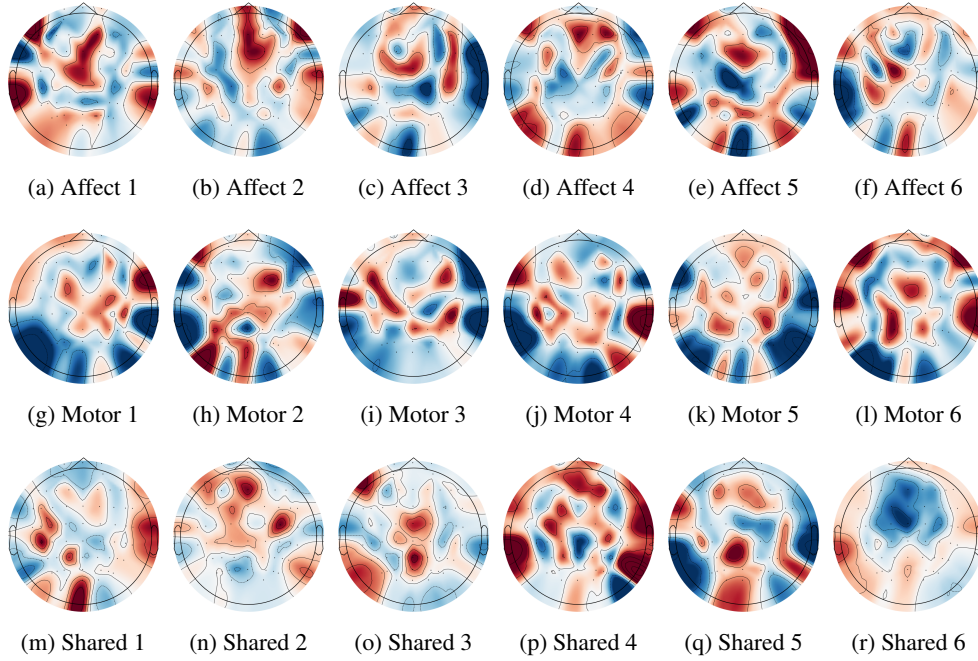


Figure 9: Visualization of learned filters for three groups of encoders after pre-training: Affect (top row), Motor (middle row), and Shared (bottom row).

be both feasible and beneficial. Together, these findings suggest that brain state-aware modeling and flexible spatial adaptation should be considered fundamental design principles for next-generation Large EEG models.

Limitations. Despite these advances, BrainPro faces several limitations. First, the current brain state taxonomy (affect, motor, others) is coarse-grained and does not capture finer distinctions such as attention, memory, or language processing, potentially restricting specialization. Second, our selective update strategy increases training complexity and may lead to imbalanced optimization when some states are underrepresented in the pre-training corpus. Finally, BrainPro remains computationally heavier than lightweight CNN-based models, which may limit deployment in real-time or resource-constrained BCI scenarios.

Future work. Several avenues remain open for future exploration. Extending BrainPro to a richer set of brain states, potentially through hierarchical or dynamically discovered state encoders, could improve generalizability and interpretability. Incorporating multimodal neurophysiological signals (e.g., fNIRS, MEG, or eye tracking) may further enrich representation learning and support cross-modal transfer. Improving efficiency, for example via parameter sharing, pruning, or distillation, will be critical for on-device BCI applications. Another promising direction is to integrate interpretable mechanisms (e.g., attention maps aligned with neurophysiological knowledge) to better understand how shared and state-specific representations contribute to decoding. Ultimately, BrainPro opens a path toward scalable and adaptive large EEG models that more faithfully reflect the diversity of human brain states and support practical BCI deployment.