BrainPro: Process-conditioned Pretraining with Topographic Retrieval for Large EEG Models

Yi Ding[1,*] **Muyun Jiang**[1,*] **Weibang Jiang**[1,2] **Shuailei Zhang**[1] **Xinliang Zhou**[1]
Chenyu Liu[1] **Shanglin Li**[3] **Yong Li**[1,4] **Cuntai Guan**[1,†]
[1]Nanyang Technological University, [2]Shanghai Jiao Tong University
[3]Berlin Institute for the Foundations of Learning and Data, [4]Southeast University

Abstract

EEG foundation models are pretrained on heterogeneous corpora, but two challenges remain underexplored: (1) experimental protocols provide mental-process cues that are rarely used in masked pretraining, and (2) electrode montages vary across datasets, making spatial representations montage-dependent. Existing methods typically learn a single general-purpose representation and model spatial structure implicitly, limiting their ability to exploit process-related information under montage heterogeneity. We introduce **BrainPro**, a process-conditioned self-supervised pretraining framework that couples topographic retrieval with shared and process-associated representation learning. BrainPro maps dataset-specific montages to a universal channel-region template and retrieves channel- and region-level spatial filters to form a topographically aligned spatial basis. Over this basis, BrainPro learns a shared encoder for general EEG structure and additional affective, motor-related, and auxiliary residual branches for process-associated variation. Protocol-derived mental-process cues condition branch activation and region-weighted masked reconstruction, incorporating topographic spatial priors and process information. Across nine public BCI benchmarks, BrainPro achieves strong performance among evaluated baselines. Ablations, channel-drop analysis, encoder-configuration studies, and spatial-filter visualizations suggest that topographic retrieval and process-conditioned representation learning jointly improve EEG decoding. Our code is available at `https://github.com/yi-ding-cs/BrainPro`.

1 Introduction

Electroencephalography (EEG) measures scalp electrical activity and provides a non-invasive, indirect view of neural dynamics associated with cognition, affect, and behavior [1]. Different mental processes can involve both shared and process-associated neural substrates. For example, affective processing can engage motor-related cortical regions that overlap with motor processes [2], while also showing associations with frontal and temporal regions [3, 4]. Motor-related processes, in contrast, are strongly associated with sensorimotor rhythms and central scalp regions [2, 4]. Although EEG reflects these neural patterns indirectly and at coarse spatial resolution, it can capture coarse process-associated differences in spectral activity and scalp distributions [4]. These observations motivate EEG representations that preserve both shared dynamics and spatial-spectral variation associated with different mental processes.

Recent EEG foundation models (EFMs) improve transferability mainly through large-scale masked pretraining [5–9]. However, two challenges remain underexplored. First, existing EFMs typically learn a single general-purpose representation, although multi-dataset EEG corpora contain

*Equal contribution † Corresponding author (ctguan@ntu.edu.sg)

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

protocol-derived information about the recording context, such as affective stimulation [10] and motor imagery [11]. We define such information a *mental-process cue*: it indicates the dominant mental process types associated with an EEG segment. Ignoring these cues may mix general and process-associated variation in one representation space, limiting process-relevant features for downstream decoding.

Second, EEG datasets differ substantially in electrode montage, channel count, spacing, and regional coverage [6]. The same scalp region may correspond to different channel sets across datasets, and some regions may be missing in low-density montages. Thus, implicit spatial mechanisms such as channel embeddings or self-attention may learn montage-dependent spatial patterns [6, 8]. This is especially problematic for process-conditioned pretraining: without an aligned spatial basis, process-associated branches may capture dataset-specific channel patterns rather than reusable scalp-level structure. These challenges are coupled. Without process conditioning, large EEG models may underuse protocol-derived mental-process information that helps to learn representation useful for downstream decoding; without topographic alignment, process conditioning may amplify montage-specific spatial biases. This motivates our central question: **How can large EEG models learn process-conditioned representations over a topographically aligned spatial basis?**

We propose **BrainPro**, a process-conditioned self-supervised pretraining framework with topographic retrieval. BrainPro maps dataset-specific montages to a universal channel-region template and retrieves channel- and region-level spatial filters according to the available electrodes in the EEG sample. This allows EEG signals recorded from the same or nearby scalp regions to share the same spatial filter weights, forming a montage-consistent spatial basis. Using this basis, BrainPro learns a shared encoder for the general EEG structure and additional affective, motor-related, and auxiliary residual branches for process-associated variation. During masked pretraining, the mental-process cue determines the active process-associated branch and the regional reconstruction prior. The auxiliary residual branch handles heterogeneous non-affect/non-motor data. We evaluate whether this design improves downstream decoding on unseen datasets with related or overlapping EEG structure.

The main contributions of this work are:

- **Topographically aligned spatial retrieval.** We design a channel- and region-level retrieval mechanism that maps heterogeneous EEG montages to a universal channel-region template, enabling spatial filters to be shared according to scalp-topographic correspondence across datasets.
- **Process-conditioned self-supervised pretraining.** We propose a multi-branch pretraining architecture in which protocol-derived mental-process cues determine the active process-associated branch during masked pretraining, allowing BrainPro to learn shared, affective, motor-related, and auxiliary residual representations.
- **Process-guided region-weighted reconstruction.** We incorporate coarse neurophysiological priors into masked EEG reconstruction by emphasizing scalp regions associated with affective and motor-related processes, while using uniform weights for residual data.
- **Broad empirical evaluation.** We evaluate BrainPro on nine public BCI benchmarks spanning six task families. Ablations, channel-drop analysis, encoder-configuration studies, and spatial-filter visualizations suggest that topographic retrieval and process-conditioned self-supervised pretraining improve EEG decoding.

2 Preliminaries

Definition 1 (Mental-process Cue)**.** A *mental-process cue* denotes dataset- or protocol-derived information indicating the dominant mental process family associated with an EEG segment, such as affective or motor-related activity.

Remark 1 (Role of mental-process cues). In BrainPro, the mental-process cue conditions the active process-associated branch and the region-weighted reconstruction prior during masked pretraining.

Definition 2 (Overlapping Mental Processes)**.** Different mental processes may recruit partially overlapping cortical regions even when they arise from distinct behavioral or cognitive conditions. For example, affective processes can involve motor-related regions, while motor-related processes are strongly associated with sensorimotor areas [2, 3]. We say two mental processes overlap when their EEG patterns contain partially shared scalp-level spatial structure.

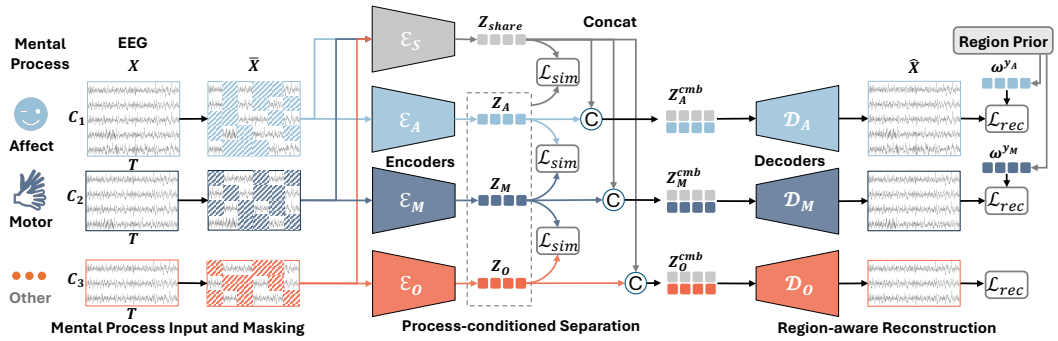


Figure 1: BrainPro framework. BrainPro consists of a shared encoder $\mathcal{E}_S$, affective and motor process-associated encoders ($\mathcal{E}_A$ and $\mathcal{E}_M$), and an auxiliary residual encoder $\mathcal{E}_O$ for heterogeneous non-affect/non-motor data. Pretraining combines process-guided region-weighted reconstruction and representation-level branch separation to learn shared and process-associated representations over a topographically aligned spatial basis.

Remark 2 (Shared vs. Process-associated Representations). Because EEG signals contain both common neural dynamics and process-related variation, it is useful to distinguish between two types of representations: (i) *shared representations*, which capture signal structure that appears across datasets and mental processes, and (ii) *process-associated representations*, which capture variation more closely related to affective or motor-related activity. Modeling both aspects can provide complementary information for downstream EEG decoding.

Definition 3 (Explicit Spatial Encoding)**.** *Explicit spatial encoding* refers to learning spatial filters whose parameters are tied to scalp channel or region locations, instead of dataset-specific channel indices. Channels from different datasets are mapped to the same or nearby locations on a universal scalp-electrode template, retrieving shared spatial-filter parameters. This encourages spatial representations that are more consistent across heterogeneous electrode montages and provides an aligned basis for mental-process-conditioned representation learning. Further justification is provided in Appendix D.

Existing EFM Pretraining Most EFMs rely on masked reconstruction or related self-supervised objectives to learn general-purpose representations. At a high level, an encoder f_θ maps an input segment $\mathbf{X}$ to a latent representation $\mathbf{Z}$:

$$\mathbf{Z} = f_\theta(\mathbf{X}).$$

These approaches capture broad EEG signal structure, but typically learn a single general-purpose representation without using protocol-derived mental-process cues. Spatial information is also modeled implicitly, e.g., through channel embeddings or attention, which may not preserve topographic consistency across heterogeneous montages. Consequently, process-associated variation and montage-specific factors may become entangled in one latent space.

BrainPro Pretraining BrainPro extends this formulation by combining topographic retrieval with process-conditioned representation learning. First, topographic retrieval maps dataset-specific montages to a universal channel-region template, producing spatial features that are comparable across electrode layouts. Second, BrainPro represents each EEG segment through a shared encoder and a process-conditioned pathway, where the mental-process cue determines which additional process-associated branch contributes to the pretraining objective:

$$\mathbf{Z} = f_{\text{shared}}(\mathbf{X}) \parallel f_{y_{\text{proc}}}(\mathbf{X}),$$

where $y_{\text{proc}} \in \{\text{A}, \text{M}, \text{O}\}$ denotes the process-conditioned branch used for that segment. Here, A and M correspond to affective and motor-related process cues, respectively, while O denotes an auxiliary residual branch for heterogeneous non-affect/non-motor data. This formulation preserves a general EEG representation while learning additional process-associated pathways over an topographically aligned spatial basis.

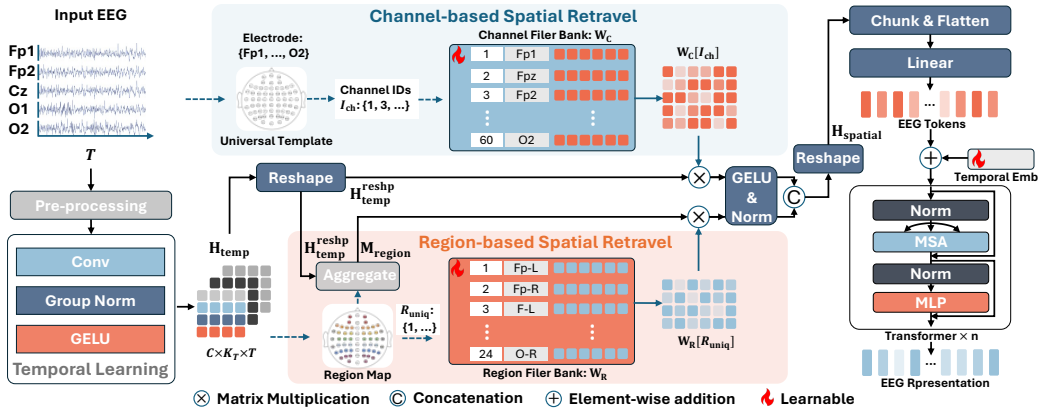


Figure 2: EEG encoder of BrainPro. Each encoder consists of three stages: (1) a temporal learning encoder based on temporal CNNs; (2) a topographic retrieval encoder that performs channel- and region-based spatial retrieval; and (3) patchification and token embedding, followed by Transformer encoders for spatiotemporal modeling.

3 Method

3.1 Overview of BrainPro Framework

The overall BrainPro framework is illustrated in Figure 1. It integrates topographic retrieval, process-conditioned branch separation, and process-guided region-weighted reconstruction into a unified pretraining pipeline (Algorithm 1). It consists of three main steps.

Step-1. Given an EEG segment, the model first extracts temporal features and retrieves channel- and region-level spatial filters based on a universal template, enabling explicit and montage-robust spatial learning. These features are then processed by a Transformer encoder to obtain spatiotemporal representations (Section 3.2 and Figure 2).

Step-2. To learn complementary shared and process-associated representations, BrainPro uses one shared encoder, affective and motor process-associated encoders, and an auxiliary residual encoder. A margin-based cosine similarity loss discourages excessive similarity among the shared, active, and inactive branch representations.

Step-3. To encourage process-relevant spatial structure, reconstruction is performed using both shared and active branch representations. The reconstruction loss is weighted by coarse regional priors for affective and motor-related samples, while the auxiliary residual branch uses uniform weights because no coherent regional prior is assumed.

3.2 Topographic Retrieval

Motivation. EEG datasets use heterogeneous montages, making direct channel-index embeddings difficult to share across datasets. Process-conditioned pretraining is more meaningful when different datasets are first aligned to a common spatial basis.

Operation. We adopt a topographic retrieval strategy. We define a universal template with C_{pre}=60 electrodes and N_{region}=24 functional regions following [12]. More details about the template are in Appendix F. We maintain learnable spatial filter banks aligned to the positions in the universal template for both *channel-* and *region-level* retrieval:

$$\mathbf{W}_{\text{C}} \in \mathbb{R}^{C_{\text{pre}} \times K_{\text{C}}}, \quad \mathbf{W}_{\text{R}} \in \mathbb{R}^{N_{\text{region}} \times K_{\text{R}}}, \tag{1}$$

where K_{C} and K_{R} denote the number of channel-wise and region-wise spatial filters.

Given an input EEG, $X \in \mathbb{R}^{C \times T}$, a temporal encoder, $\mathcal{F}_{temp}$, learns temporal dynamics, $\mathbf{H}_{\text{temp}}$. More details about $\mathcal{F}_{temp}$ can be found in Appendix A. We reshape the $\mathbf{H}_{\text{temp}}$ into $\mathbf{H}_{\text{temp}}^{\text{reshp}} \in \mathbb{R}^{C \times K_T * T}$, and then extract channel- and region-level features in parallel.

For *channel*-level features, Let $I_{\text{ch}} \subseteq \{1, \ldots, C_{\text{pre}}\}$ be the set of indices in the universal template that are present in the current sample. This set is obtained by nearest-neighbor mapping in standardized 3D head coordinates. Unmatched channels are discarded (More details in Appendix Q.6). We then retrieve the corresponding filters from the channel filter bank and apply them to the temporal features:

$$\mathbf{H}_{\text{C}} = \sigma\Big(\mathbf{W}_{\text{C}}[I_{\text{ch}}]^{\top}\mathbf{H}_{\text{temp}}^{\text{reshp}}\Big) \in \mathbb{R}^{K_{\text{C}} \times K_T * T}, \tag{2}$$

where $\sigma(\cdot)$ denotes nonlinear operations (GELU and GroupNorm).

For *region*-level features, we aggregate temporal features within each functional region present in the current sample and apply the region filter bank. Let $\mathcal{R}_{\text{uniq}}$ be the set of functional regions covered by the current sample, and let C_j be the channels belonging to region j. We first aggregate temporal features within each region:

$$\mathbf{M}_{\text{region}}[j] = \frac{1}{|\mathcal{C}_j|} \sum_{c \in \mathcal{C}_j} \mathbf{H}_{\text{temp}}^{\text{reshp}}[c, :] \in \mathbb{R}^{1 \times K_T * T}. \tag{3}$$

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

$$\mathbf{H}_{\text{R}} = \sigma\big(\mathbf{W}_{\text{R}}[\mathcal{R}_{\text{uniq}}]^{\top}\mathbf{M}_{\text{region}}\big) \in \mathbb{R}^{K_{\text{R}} \times K_T * T}. \tag{4}$$

Finally, the channel- and region-level features are concatenated and reshaped to form the spatial representation:

$$\mathbf{H}_{\text{spatial}} = \text{Reshape}(\text{Concat}(\mathbf{H}_{\text{C}}, \mathbf{H}_{\text{R}})) \in \mathbb{R}^{K_T * K_{\text{total}} \times T}, \tag{5}$$

where $K_{\text{total}} = K_{\text{C}} + K_{\text{R}}$. This $\mathbf{H}_{\text{spatial}}$ is further processed by a Transformer encoder to obtain spatiotemporal representations for mental-process-conditioned separation and process-guided region-weighted reconstruction. Details on the tokenization and Transformer encoder are in Appendix B.

Effect. This retrieval mechanism produces montage-consistent spatial representations and allows spatial parameters to be shared according to scalp location.

3.3 Process-conditioned Representation Separation

Motivation. EEG contains both shared signal structure and process-associated variation, but standard masked pretraining may mix them in one representation space. BrainPro uses protocol-derived process cues to separate shared and process-associated representations, with the latter potentially capturing complementary information beyond the shared representation.

Operation. As shown in Figure 1, BrainPro contains one shared encoder $\mathcal{E}_{\text{S}}$, two primary process-associated encoders $\mathcal{E}_{\text{A}}$ and $\mathcal{E}_{\text{M}}$ for affective and motor-related EEG as they have relatively stable scalp-level priors and are consistently available across large-scale public EEG datasets [13, 14], and an auxiliary residual encoder $\mathcal{E}_{\text{O}}$ for heterogeneous non-affect/non-motor data. Given an EEG segment $\mathbf{X}$, the encoders produce:

$$\mathbf{Z}_{\text{shared}} = \mathcal{E}_{\text{S}}(\mathbf{X}), \mathbf{Z}_k = \mathcal{E}_k(\mathbf{X}), \quad k \in \{\text{A}, \text{M}, \text{O}\}. \tag{6}$$

For an EEG sample with mental-process cue y_{proc}, BrainPro updates the shared encoder $\mathcal{E}_{\text{S}}$ and the active branch $\mathcal{E}_{y_{\text{proc}}}$ only. Inactive branch outputs are detached, so they can serve as separation targets without receiving gradients from samples outside their assigned process.

To enforce representation separation, we define a separation loss $\mathcal{L}_{\text{sep}}$ consisting of two margin-based cosine similarity losses, denoted as $\mathcal{L}_{\text{sim}}$. The similarity loss is:

$$\mathcal{L}_{\text{sim}}(\mathbf{a}, \mathbf{b}) = \max\big(\text{sim}_{\cos}(\mathbf{a}, \mathbf{b}) - m, 0\big), \tag{7}$$

where m=0.1 is a margin. The first similarity term penalizes similarity between shared and active mental-process representations, and the second penalizes similarity between the active process encoder and all inactive encoders:

$$\mathcal{L}_{\text{sep}} = \underbrace{\mathcal{L}_{\text{sim}}(\mathbf{Z}_{\text{shared}}, \mathbf{Z}_{y_{\text{proc}}})}_{\text{shared-active}} + \underbrace{\sum_{k \neq y_{\text{proc}}} \mathcal{L}_{\text{sim}}(\mathbf{Z}_{y_{\text{proc}}}, \mathbf{Z}_k)}_{\text{active-inactive}}. \tag{8}$$

Effect. The shared encoder captures general EEG patterns across datasets, while the active branch learns complementary process-associated variation.

3.4 Masking and Region-aware Reconstruction

Motivation. Masked reconstruction provides a self-supervised learning guidance, but treating all channels uniformly may underuse process-related spatial priors. BrainPro therefore conditions reconstruction on protocol-derived process cues and a shared topographic template.

Operation. We mask the input patches with a ratio of ρ=0.5 [6] and pass the masked signal $\bar{\mathbf{X}}$ through both the shared encoder and the active process-associated encoder. Their outputs are concatenated and decoded:

$$\hat{\mathbf{X}} = \mathcal{D}_{y_{\text{proc}}}\left(\mathcal{E}_{\text{S}}(\bar{\mathbf{X}}) \,\|\, \mathcal{E}_{y_{\text{proc}}}(\bar{\mathbf{X}})\right), \tag{9}$$

where masked positions in $\bar{\mathbf{X}}$ are replaced with learned `[MASK]` embeddings and $\hat{\mathbf{X}} \in \mathbb{R}^{C_{\text{pre}} \times T}$ is the reconstructed signal in the 60-channel template space.

We assign each active branch y_{proc} a channel-importance vector $w^{(y_{\text{proc}})} \in [0,1]^{C_{\text{pre}}}$ to emphasize coarse process-relevant regions. Based on neuroscience priors, frontal, temporal, and central channels are emphasized for affective samples [13, 15], while central and parietal channels are emphasized for motor-related samples [14]. For the auxiliary residual branch encoder, all channels receive equal importance. We apply a smooth weighting function:

$$\text{weights}(w) = 0.5 + \sigma\big(\tau \cdot (w - 0.5)\big), \tag{10}$$

with sharpness parameter τ=epoch to gradually increase weighting strength.

Let $S(\mathbf{X}) \subseteq \{1, \ldots, C_{\text{pre}}\}$ denote the retained template-channel indices observed in sample $\mathbf{X}$. The reconstruction loss is computed on these available mapped channels:

$$\mathcal{L}_{\text{rec}} = \frac{1}{|S(\mathbf{X})|T} \sum_{i \in S(\mathbf{X})} \sum_{t=1}^{T} \text{weights}\big(w_i^{(y_{\text{proc}})}\big) \cdot \Big(\mathbf{X}_{i,t} - \hat{\mathbf{X}}_{i,t}\Big)^2. \tag{11}$$

The weighted MSE is computed over all available retained channels and time points, while absent or unmapped template channels are excluded.

The overall pretraining objective combines reconstruction and separation:

$$\mathcal{L}_{\text{BrainPro}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{sep}}. \tag{12}$$

Effect. This design preserves a unified topographic reconstruction space, avoids forcing the model to hallucinate unobserved channels, and injects coarse affective or motor-related spatial priors into self-supervised pretraining.

4 Experiment

4.1 Datasets and Pre-processing

Pretraining. We construct a large-scale pretraining corpus from diverse EEG datasets. Affective and motor-related data are used as the two primary mental-process groups because they have relatively stable scalp-level priors. All remaining datasets are assigned to an auxiliary residual group, which absorbs heterogeneous non-affect/non-motor variability during pretraining. The resulting corpus contains about **2,400 hours** of EEG recordings, with a complete dataset list and grouping provided in Appendix G.2. All recordings are segmented into 10-second clips, resampled to 200 Hz, and scaled to 0.1 mV following [6]. Noisy segments with absolute values greater than 10 are removed, yielding approximately **2,180 hours** of clean data. Further implementation details are provided in Appendix G.3.

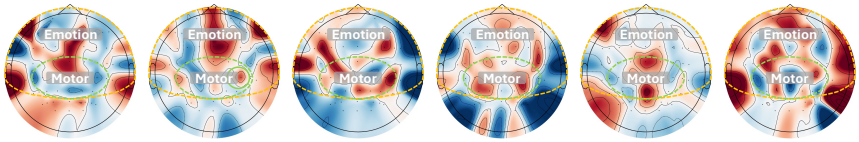
Downstream Evaluation. To comprehensively evaluate our method, we consider six representative downstream BCI tasks using nine publicly available datasets: emotion recognition, motor imagery, imagined speech, mental disorder diagnosis, mental stress detection, and attention detection. These tasks and their corresponding datasets are summarized in Table 4, with additional dataset-specific details in Appendix H.2. For consistency with pretraining, all EEG signals are resampled to 200 Hz. Further preprocessing details are provided in Appendix H.2.

Table 1: Comparison results of different methods on downstream tasks.

Methods	FACED (9-Class)			SEED-V (5-Class)			BCI-IV-2a (4-Class)		
	ACC-B	Kappa	F1-W	ACC-B	Kappa	F1-W	ACC-B	Kappa	F1-W
EEGNet	0.3692 ±0.0103	0.2880 ±0.0113	0.3693 ±0.0111	0.2408 ±0.0031	0.0536 ±0.0042	0.1908 ±0.0088	0.5521 ±0.0183	0.4028 ±0.0244	0.5346 ±0.0228
Conformer	0.4566 ±0.0108	0.3849 ±0.0121	0.4555 ±0.0098	0.3087 ±0.0095	0.1294 ±0.0134	0.2837 ±0.0229	0.4879 ±0.0183	0.3171 ±0.0244	0.4561 ±0.0206
BIOT	0.2992 ±0.0119	0.2105 ±0.0147	0.2954 ±0.0149	0.3575 ±0.0052	0.1971 ±0.0081	0.3636 ±0.0078	0.2392 ±0.0280	-0.0144 ±0.0373	0.1063 ±0.0060
EEGPT	0.2809 ±0.0116	0.1912 ±0.0129	0.2810 ±0.0120	0.2421 ±0.0048	0.0557 ±0.0077	0.2438 ±0.0055	0.3849 ±0.0226	0.1799 ±0.0301	0.3272 ±0.0336
LaBraM	0.5224 ±0.0116	0.4610 ±0.0126	0.5259 ±0.0108	0.3986 ±0.0200	0.2491 ±0.0256	0.4040 ±0.0197	0.5255 ±0.0329	0.3674 ±0.0439	0.5095 ±0.0405
CBraMod	0.5669 ±0.0094	0.5112 ±0.0110	0.5729 ±0.0105	0.3960 ±0.0033	0.2521 ±0.0048	0.4050 ±0.0052	0.5148 ±0.0402	0.3530 ±0.0536	0.5032 ±0.0512
CodeBrain	0.5801 ±0.0089	0.5242 ±0.0114	0.5821 ±0.0115	0.4025 ±0.0104	0.2564 ±0.0123	0.4095 ±0.0101	0.5024 ±0.0109	0.3366 ±0.0145	0.4952 ±0.0120
BrainPro	**0.5937** ±0.0087	**0.5418** ±0.0092	**0.6023** ±0.0061	**0.4078** ±0.0075	**0.2612** ±0.0089	**0.4115** ±0.0071	**0.5674** ±0.0148	**0.4232** ±0.0198	**0.5653** ±0.0169

Methods	SHU (2-Class)			Mental Arithmetic (2-Class)			Attention (2-Class)		
	ACC-B	AUC-PR	AUROC	ACC-B	AUC-PR	AUROC	ACC-B	AUC-PR	AUROC
EEGNet	0.5664 ±0.0522	0.6609 ±0.0099	0.6791 ±0.0085	0.5533 ±0.0280	0.5702 ±0.0539	0.5717 ±0.0456	0.6004 ±0.0123	0.6294 ±0.0288	0.6647 ±0.0180
Conformer	0.6167 ±0.0172	0.6763 ±0.0054	0.6927 ±0.0072	0.6867 ±0.0492	0.7691 ±0.0286	0.7039 ±0.0177	0.7198 ±0.0158	0.7819 ±0.0218	0.7992 ±0.0283
BIOT	0.4981 ±0.0044	0.5077 ±0.0025	0.5069 ±0.0060	0.5583 ±0.0306	0.5512 ±0.0516	0.5662 ±0.0250	0.6111 ±0.0411	0.7273 ±0.0132	0.7367 ±0.0162
EEGPT	0.5481 ±0.0139	0.5631 ±0.0165	0.5673 ±0.0197	0.5650 ±0.0341	0.5860 ±0.0805	0.5937 ±0.0762	0.6674 ±0.0560	0.8015 ±0.0372	0.8103 ±0.0303
LaBraM	0.6175 ±0.0172	0.6821 ±0.0214	0.6720 ±0.0253	0.6688 ±0.0279	0.7504 ±0.0649	0.7168 ±0.0452	0.6785 ±0.0223	0.7838 ±0.0307	0.7994 ±0.0198
CBraMod	0.6108 ±0.0233	0.6632 ±0.0277	0.6668 ±0.0338	0.7354 ±0.0410	0.8237 ±0.0225	0.7654 ±0.0203	0.6478 ±0.0258	0.7417 ±0.0175	0.7468 ±0.0198
CodeBrain	0.6205 ±0.0089	0.6819 ±0.0154	0.6833 ±0.0146	0.7111 ±0.0210	0.7984 ±0.0112	0.7510 ±0.0139	0.6988 ±0.0134	0.7518 ±0.0202	0.7782 ±0.0080
BrainPro	**0.6319** ±0.0107	**0.7102** ±0.0076	**0.7105** ±0.0076	**0.8083** ±0.0156	**0.8980** ±0.0052	**0.8512** ±0.0083	**0.7222** ±0.0291	**0.7975** ±0.0392	**0.8064** ±0.0259

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



(a) Affect (b) Affect (c) Motor (d) Motor (e) Shared (f) Shared

Figure 3: Visualization of the learned spatial filters for the Affect, Motor, and Shared encoders after pretraining. Each subplot shows a 2D scalp topography in which warmer colors indicate larger filter weights and cooler colors indicate smaller weights, reflecting the relative contribution of each scalp region to the encoder's representation. Emotion- and motor-related regions are outlined.

4.2 Baselines and Evaluation Metrics

Following prior studies [6, 8], we compare BrainPro against both non-foundation and foundation model baselines. All baselines share the same fine-tuning and evaluation setting. Among non-foundation approaches, we include EEGNet [16] and Conformer [17]. For foundation-model baselines, we consider: BIOT [9], LaBraM [6], EEGPT [7], CBraMod [8] and CodeBrain [18]. Further baseline details are in Appendix H.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 studies [6, 8], we report Balanced Accuracy (ACC-B), Cohen's Kappa, and Weighted F1 (F1-W).

4.3 Implementation Details

Pretraining. Pretraining is conducted end-to-end. 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 G.4 (Table 7).

Downstream Evaluation. We split each dataset into non-overlapping train/validation/test sets to avoid potential data leakage [8, 19, 20]. Models are trained on the training split, selected using the validation split, and evaluated once on the test split. We report mean±std over five seeds $\{0, 1, 2, 3, 4\}$. Experiments run on 5 NVIDIA A800 GPUs. Hyperparameters are provided in Appendix H.1. For downstream adaptation the shared encoder is always activated, and the affective, motor-related, or auxiliary residual branch is activated according to the downstream task family. Details are provided

Table 2: Ablation studies of the main components on affective and motor-related downstream tasks.

Methods	FACED (9-Class)			BCI-IV-2A (4-Class)		
	ACC-B	Kappa	F1-W	ACC-B	Kappa	F1-W
w/o reconstruction	0.5391 ±0.0151	0.3854 ±0.0202	0.5401 ±0.0177	0.4354 ±0.0329	0.2472 ±0.0439	0.3758 ±0.0548
w/o separation	0.5638 ±0.0138	0.4184 ±0.0185	0.5603 ±0.0107	0.4870 ±0.0184	0.3160 ±0.0246	0.4725 ±0.0219
w/ random retrieval	0.5747 ±0.0098	0.4329 ±0.0130	0.5674 ±0.0068	0.5125 ±0.0308	0.3500 ±0.0410	0.5052 ±0.0379
w/o spatial prior	0.5634 ±0.0271	0.5060 ±0.0319	0.5672 ±0.0330	0.5496 ±0.0143	0.3994 ±0.0191	0.5495 ±0.0141
w/ random process cue	0.5269 ±0.0288	0.4682 ±0.0341	0.5330 ±0.0329	0.5232 ±0.0065	0.3642 ±0.0087	0.5104 ±0.0127
w/ single encoder	0.5683 ±0.0241	0.5133 ±0.0281	0.5735 ±0.0272	0.5554 ±0.0142	0.4072 ±0.0189	0.5500 ±0.0176
w/o pretraining	0.4244 ±0.2751	0.3512 ±0.3088	0.4028 ±0.3202	0.4479 ±0.1066	0.2639 ±0.1421	0.4224 ±0.1454
BrainPro	**0.5937** ±0.0087	**0.5418** ±0.0092	**0.6023** ±0.0061	**0.5674** ±0.0148	**0.4232** ±0.0198	**0.5653** ±0.0169

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

in Appendix C. After loading pre-trained weights, we re-initialize temporal position embeddings with Xavier uniform initialization. An ablation is provided in Appendix M.

4.4 Primary Results

BrainPro improves downstream decoding across diverse BCI tasks. Table 1 compares BrainPro with non-foundation and foundation-model baselines across six downstream BCI task families. EEGNet and Conformer perform reasonably on some binary tasks but degrade on more complex multi-class settings, particularly in Kappa and weighted F1. BIOT and EEGPT show weaker transferability on several datasets, while LaBraM and CBraMod benefit from larger-scale pretraining. BrainPro achieves strong overall performance among the evaluated baselines across the reported benchmarks. 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. More results are provided in Appendix I. These results suggest that BrainPro improves downstream EEG decoding under the reported evaluation protocols.

4.5 Interpretation of Learned Spatial Filters

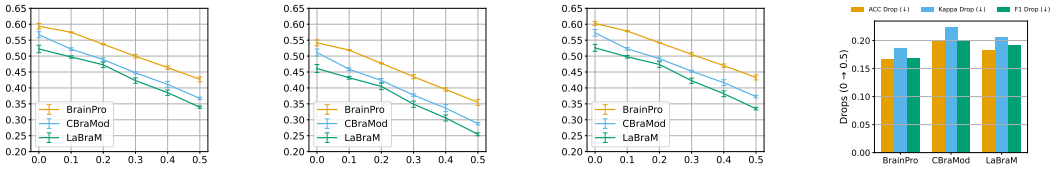
Spatial filters show process-consistent topographic patterns. Figure 3 visualizes the spatial filters learned by the Affect, Motor, and Shared encoders after pretraining. The learned filters show qualitatively distinct scalp patterns. The Affect encoder emphasizes frontal and temporal regions, consistent with affective EEG literature [13, 15], while the Motor encoder highlights central sensorimotor areas [14]. In contrast, the Shared encoder produces more distributed patterns that may reflect global EEG structure. These visualizations provide qualitative evidence that BrainPro learns mental-process-associated spatial structure over an topographically aligned basis. More details are provided in Appendix P.

4.6 Ablation Studies

Ablations support the contribution of major components. (1) Removing reconstruction degrades performance, showing that masked reconstruction provides useful self-supervised EEG structure. (2) Removing separation lowers decoding results, indicating that branch decoupling improves shared and process-associated representation learning. (3) Random retrieval performs worse than BrainPro, supporting the benefit of topographically aligned spatial filters. (4) Removing the spatial prior reduces performance, showing that process-guided region weighting adds value beyond retrieval alone. (5) Random process cues cause clear drops, suggesting that correct protocol-derived cues matter and that the gains are not from arbitrary branching. (6) The matched-size single encoder remains below BrainPro, indicating that process-conditioned multi-branch specialization is beneficial beyond simply increasing model size. (7) Removing pretraining leads to large drops and high variance, confirming the importance of large-scale self-supervised pretraining. Overall, all major components contribute to BrainPro's performance on affective and motor-related downstream tasks.

4.7 Channel-Drop Analysis

BrainPro exhibits smaller performance degradation under missing-channel conditions. Figure 4 evaluates decoding performance on FACED under progressively increasing channel-drop rates



(a) ACC-B (b) Kappa (c) F1-W (d) Drop from 0 to 0.5

Figure 4: Performance of BrainPro, CBraMod, and LaBraM under different levels of channel-drop rate.

Table 3: Effects of different encoder configurations on downstream performance.

	FACED (9-Class)			BCI-IV-2A (4-Class)		
Method	ACC-B	Kappa	F1-W	ACC-B	Kappa	F1-W
S	0.5635 ±0.0031	0.5088 ±0.0050	0.5710 ±0.0049	0.5527 ±0.0218	0.4036 ±0.0290	0.5504 ±0.0236
S + A	0.5937 ±0.0087	0.5418 ±0.0092	0.6023 ±0.0061	0.5499 ±0.0044	0.3999 ±0.0059	0.5469 ±0.0060
S + M	0.5828 ±0.0142	0.5296 ±0.0143	0.5904 ±0.0144	0.5674 ±0.0148	0.4232 ±0.0198	0.5653 ±0.0169
S + O	0.5821 ±0.0144	0.5292 ±0.0150	0.5881 ±0.0133	0.5525 ±0.0173	0.4033 ±0.0231	0.5491 ±0.0192
S + A + M	0.5988 ±0.0057	0.5477 ±0.0058	0.6098 ±0.0043	0.5797 ±0.0229	0.4396 ±0.0305	0.5764 ±0.0245
S + A + M + O	0.5824 ±0.0050	0.5272 ±0.0054	0.5841 ±0.0056	0.5660 ±0.0067	0.4213 ±0.0089	0.5631 ±0.0080

Note: Cyan highlight marks configurations used in BrainPro.

during inference, where a random subset of channels is removed at each drop level. Detailed results are in Table 13. This provides an assessment of how models behave when electrode coverage is incomplete. All metrics decrease as more channels are removed, and the differences between models become more pronounced at higher drop ratios. BrainPro consistently maintains higher performance than LaBraM and CBraMod across the entire drop spectrum, and its total ACC-B drop at 50% removal is the smallest among all models shown in Figure 4 (d). These results provide empirical evidence that the structured retrieval mechanism contributes to more stable downstream performance when electrodes are missing, supporting the intended behavior of BrainPro under montage variability.

4.8 Encoder-Combination Analysis

Table 3 reports decoding performance on FACED and BCI-IV-2A under different encoder combinations. Task-relevant process branches improve over the shared-only encoder: $\mathrm{S}{+}\mathrm{A}$ performs best among single-branch variants on FACED, while $\mathrm{S}{+}\mathrm{M}$ performs best on BCI-IV-2A. This suggests that the affective and motor-related encoders learn task-aligned mental-process-associated representations. Combining affective and motor branches ($\mathrm{S}{+}\mathrm{A}{+}\mathrm{M}$) can further improve performance, indicating complementary EEG information across branches. However, adding the auxiliary residual branch to all encoders does not always help, suggesting that this branch is a pragmatic auxiliary pathway instead of a coherent mental-process-associated encoder. An extended analysis further shows that the gains are not explained by MLP parameter changes alone (Appendix K).

5 Conclusion

We introduced BrainPro, a process-conditioned pretraining framework with topographic retrieval for large EEG models. BrainPro couples topographic retrieval with shared and process-associated encoders, an auxiliary residual branch, and process-guided region-weighted reconstruction to learn montage-consistent and process-sensitive EEG representations. Across nine public BCI benchmarks, BrainPro improves over evaluated baselines under the reported protocols. Spatial-filter visualizations, channel-drop tests, encoder-combination studies, and ablations suggest that the gains are associated with montage-consistent spatial modeling and process-conditioned specialization. BrainPro provides a practical framework for using coarse mental-process cues and topographic priors in large-scale EEG pretraining.

Acknowledgments and Disclosure of Funding

This research is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Grant No. MOE-T2EP20124-0001)

References

[1] Bradley J. Edelman, Shuailei Zhang, Gerwin Schalk, Peter Brunner, Gernot Müller-Putz, Cuntai Guan, and Bin He. Non-invasive brain-computer interfaces: State of the art and trends. *IEEE Reviews in Biomedical Engineering*, 18:26–49, 2025. doi: 10.1109/RBME.2024.3449790.

[2] Alfons Schnitzler, Stephan Salenius, Riitta Salmelin, Veikko Jousmäki, and Riitta Hari. Involvement of primary motor cortex in motor imagery: A neuromagnetic study. *NeuroImage*, 6(3): 201–208, 1997. ISSN 1053-8119.

[3] Vesa Putkinen, Sanaz Nazari-Farsani, Kerttu Seppälä, Tomi Karjalainen, Lihua Sun, Henry K Karlsson, Matthew Hudson, Timo T Heikkilä, Jussi Hirvonen, and Lauri Nummenmaa. Decoding music-evoked emotions in the auditory and motor cortex. *Cerebral Cortex*, 31(5): 2549–2560, 12 2020. ISSN 1047-3211.

[4] Abigail S. Greene, Corey Horien, Daniel Barson, Dustin Scheinost, and R. Todd Constable. Why is everyone talking about brain state? *Trends in Neurosciences*, 46(7):508–524, Jul 2023. ISSN 0166-2236.

[5] 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.

[6] 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.

[7] 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, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 39249–39280. Curran Associates, Inc., 2024.

[8] 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.

[9] 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, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 78240–78260. Curran Associates, Inc., 2023.

[10] 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.

[11] 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.

[12] Yi Ding, Neethu Robinson, Chengxuan Tong, Qiuhao 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.

[13] Soraia M. Alarcão and Manuel J. Fonseca. Emotions recognition using eeg signals: A survey. *IEEE Transactions on Affective Computing*, 10(3):374–393, 2019. doi: 10.1109/TAFFC.2017.2714671.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] Jingying Ma, Feng Wu, Qika Lin, Yucheng Xing, Chenyu Liu, Ziyu Jia, and Mengling Feng. Codebrain: Bridging decoupled tokenizer and multi-scale architecture for EEG foundation model. In *The Fourteenth International Conference on Learning Representations*, 2026.

[19] Xiang Li, Yazhou Zhang, Prayag Tiwari, Dawei Song, Bin Hu, Meihong Yang, Zhigang Zhao, Neeraj Kumar, and Pekka Marttinen. Eeg based emotion recognition: A tutorial and review. *Association for Computing Machinery*, 55(4), November 2022. ISSN 0360-0300.

[20] Yi Ding, Neethu Robinson, Su Zhang, Qiuhao 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 26304–26321. Curran Associates, Inc., 2023.

[30] 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)*, pages 1–5, 2024. doi: 10.1109/ISBI56570.2024.10635453.

[31] Arman Savran, Koray Ciftci, Guillame 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.

[32] 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.

[33] 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.

[34] Logan Trujillo. Raw EEG Data. 2020. doi: 10.18738/T8/SS2NHB. URL `https://doi.org/10.18738/T8/SS2NHB`.

[35] 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.

[36] W. Zheng, W. Liu, Y. Lu, B. Lu, and A. Cichocki. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics*, pages 1–13, 2018. ISSN 2168-2267. doi: 10.1109/TCYB.2018.2797176.

[37] 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.

[38] 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.

[39] 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)*, pages 1–3. IEEE, 2017.

[40] 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.

[41] 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)*, pages 1–5. IEEE, 2017.

[42] 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, page 5975–5984, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085.

[43] 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)*, pages 1353–1357, 2021. doi: 10.1109/BIBM52615.2021.9669637.

[44] 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)*, pages 4167–4170, 2022. doi: 10.1109/EMBC48229.2022.9871630.

[45] 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)*, pages 946–949, 2021. doi: 10.1109/NER49283.2021.9441086.

[46] 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)*, pages 1–8, 2020. doi: 10.1109/IJCNN48605.2020.9206957.

[47] 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.

[48] 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.

[49] 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.

[50] 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`.

[51] 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.

[52] 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.

[53] 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`.

[54] Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4 (1), 2019. ISSN 2306-5729.

Appendix Contents

A More details about the temporal encoder in BrainPro . . . **15**

B More Details on Tokenization and the Transformer Used in This Paper . . . **15**

C Downstream Adaptation . . . **16**

D Theoretical Justification . . . **16**

D.1 Preliminaries . . . 17

D.2 Explicit Spatial Encoding via Spatial Filters . . . 17

D.3 Lack of Explicit Spatial Encoding in Self-Attention . . . 17

D.4 Retrieval-Based Alignment Restores Explicit Spatial Structure . . . 17

D.5 Conclusion . . . 18

E Related Work . . . **18**

E.1 Previous Methods for EEG Decoding . . . 18

E.2 EEG Foundation Models . . . 18

F More details for the Brain Region Definition . . . **19**

G More Details For Pretraining . . . **19**

G.1 Main Modules . . . 19

G.2 Pretraining Datasets . . . 20

G.3 Experiment Settings . . . 20

G.4 Hyper-parameters . . . 21

G.5 Compute resources . . . 21

H More Details For Downstream Tasks . . . **22**

H.1 Hyper-parameters . . . 22

H.2 Downstream Datasets . . . 22

H.3 Baselines . . . 23

H.4 Evaluation Metrics . . . 24

I More Experimental Results . . . **24**

I.1 SEED-VII . . . 24

I.2 Mental Disorder Diagnosis . . . 24

I.3 Speech . . . 25

I.4 Linear Probing on Representative Affective and Motor-related Tasks . . . 26

J Effects of Learning Rate . . . **27**

K Effect of Encoder Count and MLP Size on Performance . . . **27**

L Effects of Clear Data Pretraining . . . **27**

M Effects of Resetting the Temporal Position Embedding . . . **28**

N Effects of Different Token Selection Mode . . . **28**

O Parameters size and FLOPS . . . **29**

P Visualization of Learned Spatial Filters in Pretraining . . . **29**

Q Additional Comparisons and Diagnostics . . . **30**

Q.1 Further Analysis on SEED-VII . . . 30

Q.2 Data Scale and Separation-margin Sensitivity . . . 30

Q.3 Region Weighting, Spatial retrieval, and Fusion . . . 31

Q.4 Process Information and Dataset Identity . . . 32

Q.5 Branch Activation . . . 33

Q.6 Montage Mapping . . . 34

R Discussion . . . **34**

S Broader impacts . . . **35**

A More details about the temporal encoder in BrainPro

Following [6], a temporal CNN, $\mathcal{F}_{temp}$, extracts temporal dynamics along the time axis from the input EEG, $X \in \mathbb{R}^{C\times T}$, 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

$$\mathbf{Z}^{(\ell)} = \mathcal{F}_{temp}(\mathbf{Z}^{(\ell-1)}) = \mathrm{GELU}\Big(\mathrm{GroupNorm}\Big(\mathrm{Conv1D}^{(\ell)}(\mathbf{Z}^{(\ell-1)})\Big)\Big), \quad \ell = 1, \ldots, L_T. \tag{13}$$

where $\mathbf{Z}^{(0)} = \mathbf{X}$ and each $\mathrm{Conv1D}^{(\ell)}$ 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 $\mathbf{H}_{\mathrm{temp}} = \mathbf{Z}^{(L_T)} \in \mathbb{R}^{K_T\times C\times T}$ preserves temporal resolution (T) while expanding the per-channel feature depth to K_T.

Table 4: Overview of downstream BCI tasks and datasets.

BCI Task	Dataset	f_s	# Ch.	T	# Samples	Label
I.Emotion Recognition	FACED	250 Hz	32	10 s	10,332	9-class
	SEED-V	1000 Hz	62	1 s	117,744	5-class
	SEED-VII	1000 Hz	62	1 s	281,679	7-class
II.Motor Imagery	BCI-IV-2A	250 Hz	22	4 s	5,088	4-class
	SHU-MI	250 Hz	32	4 s	11,988	2-class
III.Imagined Speech	BCIC2020-3	256 Hz	64	4 s	5,250	5-class
IV.Mental Disorder Diagnosis	Mumtaz2016	256 Hz	19	4 s	8,860	2-class
V.Mental Stress Detection	MentalArithmetic	500 Hz	20	4 s	1,080	2-class
VI.Mental Attention Detection	ATTEN	500 Hz	28	4 s	4,680	2-class

Note: All datasets are preprocessed into fixed-length EEG segments following their official protocols.

B More Details on Tokenization and the Transformer Used in This Paper

Given the spatial representation from the retrieval-based spatial encoder, $\mathbf{H}_{\mathrm{spatial}}$, we divide it into N_p (possibly strided with the step being s_P) temporal patches of length P:

$$\mathbf{Z}_{\mathrm{patch}} \in \mathbb{R}^{N_p\times K_T\times K_{\mathrm{total}}\times P}, \quad N_p = \left\lfloor \frac{T-P}{s_P} \right\rfloor + 1. \tag{14}$$

Each patch is flattened and projected into a d-dimensional token using a learnable embedding matrix $\mathbf{E} \in \mathbb{R}^{d\times(K_T K_{\mathrm{total}} P)}$, followed by adding a positional embedding $\mathbf{p}_i \in \mathbb{R}^d$ onto the flattened and linearly projected tokens:

$$\mathbf{z}_i = \mathbf{E}\,\mathrm{Flatten}(\mathbf{Z}_{\mathrm{patch}}[i]) + \mathbf{p}_i, \quad i = 1, \ldots, N_p. \tag{15}$$

This yields a sequence of tokens $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^{N_p} \in \mathbb{R}^{N_p\times d}$, which serves as the input to the Transformer layers for modeling temporal contextual information.

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

$$\mathbf{Z}'^{(\ell)} = \mathbf{Z}^{(\ell)} + \mathrm{MSA}\Big(\mathrm{LN}(\mathbf{Z}^{(\ell)})\Big), \tag{16}$$

$$\mathbf{Z}^{(\ell+1)} = \mathbf{Z}'^{(\ell)} + \mathrm{MLP}\Big(\mathrm{LN}(\mathbf{Z}'^{(\ell)})\Big), \quad \ell = 0, \ldots, L-1, \tag{17}$$

where $\mathbf{Z}^{(0)} = \mathbf{Z}$ denotes the input token 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

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{18}$$

where d_k is the key dimension. Multiple attention heads operate in parallel and are concatenated to form richer contextual representations.

The MLP module consists of two fully connected layers with hidden dimension d_{ff} and a nonlinear GELU activation, applied independently to each token. Residual connections around both the MSA and MLP components stabilize training.

After L blocks, we obtain the final token sequence $\mathbf{Z}^{(L)} \in \mathbb{R}^{N_p \times d}$.

C Downstream Adaptation

Downstream adaptation follows a fixed branch-activation rule in all main experiments. The shared encoder is always activated. For affective downstream tasks, we activate the affective branch; for motor-imagery tasks, we activate the motor-related branch; for imagined-speech tasks, we activate the auxiliary residual branch. Only the activated encoders are fine-tuned, and their outputs are concatenated and fed to an MLP classifier [8]. For future applications where the process category is ambiguous or mixed, one possible strategy is to select from a fixed candidate set $\{\mathrm{S}, \mathrm{S{+}A}, \mathrm{S{+}M}, \mathrm{S{+}O}, \mathrm{S{+}A{+}M}\}$ using the validation split and choose the smallest near-best configuration.

Algorithm 1 BrainPro Pretraining Pipeline

Input: EEG segment $\mathbf{X} \in \mathbb{R}^{C \times T}$, montage mapping, mental-process cue y_{proc}
Params: Shared encoder $\mathcal{E}_{\mathrm{S}}$, mental-process encoders $\{\mathcal{E}_k\}_{k \in \{\mathrm{A,M,O}\}}$, mental-process decoders $\{\mathcal{D}_k\}_{k \in \{\mathrm{A,M,O}\}}$, spatial filter banks $\mathbf{W}_{\mathrm{C}}, \mathbf{W}_{\mathrm{R}}$, channel priors $w^{(y_{\text{proc}})}$, learnable input mask, mask ratio ρ

Step 1: Mask input
Replace a randomly selected fraction ρ of patches in $\mathbf{X}$ with learnable masks to obtain $\tilde{\mathbf{X}}$
Step 2: Representation extraction for separation
for each encoder $\mathcal{E}_j \in \{\mathcal{E}_{\mathrm{S}}, \mathcal{E}_{\mathrm{A}}, \mathcal{E}_{\mathrm{M}}, \mathcal{E}_{\mathrm{O}}\}$ **do**
  Temporal encoding: $\mathbf{H}_{\text{temp}} \leftarrow \mathcal{F}_{\text{temp}}(\tilde{\mathbf{X}})$
  Spatial retrieval: $\mathbf{H}_{\text{spatial}} \leftarrow \mathcal{F}_{\text{spatial}}(\mathbf{H}_{\text{temp}}, \mathbf{W}_{\mathrm{C}}, \mathbf{W}_{\mathrm{R}})$
  Patchify & embed: $\mathbf{Z} \leftarrow \mathrm{PatchifyEmbed}(\mathbf{H}_{\text{spatial}})$
  Transformer encoding: $\mathbf{Z}_j \leftarrow \mathrm{Transformer}(\mathbf{Z})$
  if $j \neq y_{\text{proc}}$ and $j \neq \mathrm{S}$ **then**
    Detach $\mathbf{Z}_j$ (inactive encoders, used only for separation loss)
  end if
end for
Step 3: Reconstruction
Concatenate outputs: $\mathbf{Z}^{\text{comb}}_{y_{\text{proc}}} \leftarrow \mathrm{Concat}(\mathbf{Z}_{\mathrm{S}}, \mathbf{Z}_{y_{\text{proc}}})$
Reconstruct EEG: $\hat{\mathbf{X}} \leftarrow \mathcal{D}_{y_{\text{proc}}}(\mathbf{Z}^{\text{comb}}_{y_{\text{proc}}})$
Step 4: Loss computation
Compute $\mathcal{L}_{\text{rec}}$ against the unmasked target on retained channels using Eq. (11)
Mental-process separation loss: $\mathcal{L}_{\text{sep}} \leftarrow$ margin-based cosine loss
Total loss: $\mathcal{L}_{\text{BrainPro}} \leftarrow \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{sep}}$
Step 5: Parameter update
Update $\mathcal{E}_{\mathrm{S}}$, $\mathcal{E}_{y_{\text{proc}}}$, $\mathcal{D}_{y_{\text{proc}}}$, and learnable mask parameters using $\mathcal{L}_{\text{BrainPro}}$

D Theoretical Justification

In this section, we provide a theoretical analysis showing that (1) retrieval-based spatial filtering enables *explicit* encoding of spatial information across heterogeneous EEG channel montages, whereas (2) self-attention yields *implicit* spatial representations that do not correspond to stable channel identities.

D.1 Preliminaries

Let an EEG input be represented as $X \in \mathbb{R}^{C\times T}, \qquad x(t) = X_{:,t}$, where C is the number of channels and T is the number of time samples. Each channel c corresponds to a fixed scalp location $p_c \in \mathbb{R}^3$.

A spatial filter is parameterized by $w = (w_1, \ldots, w_C)^\top \in \mathbb{R}^C$, and produces the spatially filtered signal

$$y(t) = w^\top x(t) = \sum_{c=1}^{C} w_c X_{c,t}.$$

Self-attention operates on channel embeddings $H \in \mathbb{R}^{C\times d}$ via $Q = HW_Q, \quad K = HW_K, \quad V = HW_V$, and produces

$$Z = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V,$$

where W_Q, W_K, W_V are shared projection matrices.

D.2 Explicit Spatial Encoding via Spatial Filters

Proposition 1 (Explicit Spatial Encoding)**.** *If a model contains a learnable parameter w_c that always multiplies the same channel c, then the model explicitly encodes spatial information because each parameter corresponds to a fixed scalp location p_c.*

Proof. The mapping channel $\to$ scalp location is fixed: $c \mapsto p_c$. Since w_c always multiplies $X_{c,t}$, the contribution of scalp location p_c to the output $y(t)$ is governed by a dedicated parameter. Thus spatial information is tied directly to parameters, yielding an explicit spatial representation. □

D.3 Lack of Explicit Spatial Encoding in Self-Attention

Proposition 2 (Self-Attention Is Not Spatially Explicit)**.** *Self-attention cannot explicitly encode spatial structure because none of its learnable parameters are tied to channel identity or scalp location. The attention weights a_{ij} depend on input content rather than spatial location.*

Proof. Self-attention uses shared parameters W_Q, W_K, W_V for all channels. For channels i and j, the interaction weight

$$a_{ij} = \frac{\exp(q_i k_j^\top)}{\sum_{j'=1}^{C} \exp(q_i k_{j'}^\top)}$$

depends on the embeddings h_i and h_j, which depend on the input signal and not on fixed locations p_i or p_j. Thus a_{ij} does not represent a stable spatial relationship across datasets or subjects. Since no parameter is associated with a fixed channel index, self-attention lacks explicit spatial semantics. □

D.4 Retrieval-Based Alignment Restores Explicit Spatial Structure

Suppose dataset A has channels $\{1, \ldots, C_A\}$ with locations $\{p_c^{(A)}\}$, and dataset B has channels $\{1, \ldots, C_B\}$ with locations $\{p_c^{(B)}\}$. A retrieval function

$$\phi : \{1, \ldots, C_B\} \to \{1, \ldots, C_A\}$$

selects, for each channel c in dataset B, the most similar channel in dataset A:

$$\phi(c) = \arg\min_{j\in\{1,\ldots,C_A\}} d(p_c^{(B)}, p_j^{(A)}),$$

for distance measure $d(\cdot, \cdot)$.

The spatial filter learned on dataset A, $w^{(A)} \in \mathbb{R}^{C_A}$, is transferred to dataset B via

$$w_c^{(B)} = w_{\phi(c)}^{(A)}.$$

Proposition 3 (Retrieval Restores Explicit Spatial Encoding)**.** *If ϕ maps channels in dataset B to their spatially closest counterparts in dataset A, then the transferred filter $w^{(B)}$ preserves explicit spatial semantics across datasets.*

Proof. The retrieval function ϕ associates each channel c in dataset B with a channel $\phi(c)$ in dataset A that corresponds to the closest spatial location. Thus $w_c^{(B)}$ inherits the spatial meaning of $w_{\phi(c)}^{(A)}$. By Proposition 1, $w_{\phi(c)}^{(A)}$ explicitly encodes the spatial contribution of a fixed location. Therefore, $w_c^{(B)}$ is tied to a stable spatial meaning, yielding an explicit spatial representation across heterogeneous channel configurations. □

D.5 Conclusion

Spatial filters produce explicit spatial representations because each parameter corresponds to a fixed topographic location. Self-attention produces only implicit spatial structure because its interaction weights depend on sample-dependent feature similarity. Retrieval-based alignment allows spatial filters to retain explicit spatial semantics even when aggregating data across heterogeneous EEG montages.

E Related Work

E.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 [21]. With the rise of deep learning, neural architectures emerged to learn spatiotemporal EEG representations directly from raw signals [22]. CNN-based models captured spatial correlations between electrodes [16, 20], while recurrent networks such as LSTMs modeled temporal dynamics [23, 24]. Hybrid CNN-LSTM frameworks [25] and attention-based architectures [17, 26] further advanced decoding performance, and graph neural networks were introduced to exploit functional brain connectivity [12, 27, 28]. 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.

E.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 [5]. These models aim to learn universal neural representations that can be adapted to diverse downstream BCI tasks. Brant [29] and NeuroGPT [30], explored masked modeling and contrastive learning strategies to enhance cross-task generalization. BIOT [9] introduced a unified encoder that tokenizes heterogeneous biosignals into a sentence-like representation to support cross-dataset pretraining. LaBraM [6] 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 [7] 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 [8] 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 important limitations. Many existing models capture spatial interactions only implicitly or assume fixed channel configurations, limiting adaptability across heterogeneous montages. They also typically learn a single general-purpose representation, without explicitly modeling shared EEG structure and variation associated with different mental processes, such as affective and motor-related activity. These constraints motivate models that explicitly encode spatial dependencies while learning both shared and process-associated representations.

F 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 [12] 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 5.

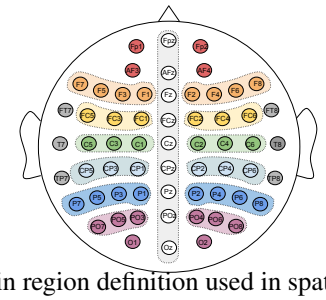


Figure 5: The brain region definition used in spatial learning block.

G More Details For Pretraining

G.1 Main Modules

The main modules in BrainPro framework are show in Table 5. The process-cue column indicates whether the operation depends on the sample's process label. Retrieval indices depend on the montage, while spatial filter parameters are encoder-specific.

Table 5: Overview of the main modules in BrainPro.

Module	**Input**	**Operation**	**Process cue**	**Purpose**
Topographic mapping	Dataset montage	Map supported channels to the 60-channel template	×	Establish a common spatial basis
Channel and region retrieval	Temporal features and template indices	Retrieve spatial filters from each encoder using montage-defined indices	×	Associate spatial parameters with scalp locations
Branch routing	EEG segment and process label	Route gradients to the shared and assigned process encoders	✓	Learn shared and process-associated representations
Separation loss	Branch representations	Apply margin-based cosine regularization with detached inactive outputs	✓	Discourage excessive similarity between branches
Reconstruction	Shared and active branch outputs	Decode the unmasked EEG target in template space using the active decoder	✓	Provide a masked reconstruction objective
Region weighting	Reconstruction errors, process priors, and epoch	Apply epoch-dependent channel weights, with uniform weights for the auxiliary group	✓	Emphasize coarse process-relevant regions

G.2 Pretraining Datasets

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

- **Emobrain** [31]: 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** [32]: EEG from 12 subjects (32 channels, 500 Hz) performing grasp-and-lift trials with a BrainAmp EEG amplifier. (11.72 h)
- **EEG Motor Movement/Imagery** [33]: 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** [11]: Data from 54 subjects conducting motor imagery tasks. It was recorded with 62 channels, 1000 Hz. (24.00 h)
- **HGD** [22]: Data from over 14 subjects, recorded 128 channels, 500 Hz. (7.49 h)
- **Raw EEG Data** [34]: EEG (64 channels, 256 Hz) collected during information-integration and multidimensional rule-based categorization tasks. (34.35 h)
- **Resting State EEG Data** [35]: 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** [10, 36, 37]: 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** [38]: 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** [39]: 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** [40]: Seizure-annotated EEG corpus (19-23 channels, 256 Hz) including onset, offset, channel, and seizure type information. (1138.53 h)
- **TUSL** [41]: TUEG subset with slowing event annotations (23 channels, 256 Hz), used in seizure detection error analysis. (20.59 h)
- **SHJT EEG Data** [42–46]: Data from over 140 subjects, recorded with the ESI NeuroScan System (62 channels, 1000 Hz). (342.23 h)

G.3 Experiment Settings

We describe the experimental setting for pretraining 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 pretraining 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-associated representation learning, we categorize the datasets into three groups, *affect*, *motor* and *other*, for brain-process decoupling training (see Table 6 for details). To make the pretraining 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. We pre-trained BrainPro for 30 epochs. 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.

Table 6: Categories of the pretraining 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, and TUSL

G.4 Hyper-parameters

The pretraining of BrainPro is in an end-to-end manner, and the hyper-parameter settings are shown in Table 7.

Table 7: Hyperparameters for pretraining.

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

G.5 Compute resources

Pretraining was performed on a server with 5 NVIDIA A800 GPUs, approximately 512 GB RAM, and 1 TB storage. The pretraining run took approximately 3 days, corresponding to about 360 A800

GPU-hours. Downstream fine-tuning experiments were run on the same hardware and required substantially less compute than pretraining.

H More Details For Downstream Tasks

H.1 Hyper-parameters

Table 8 lists the fine-tuned hyper-parameters for each downstream task, while the common training hyper-parameters are shown in Table 9.

Table 8: Tuned hyper-parameters for downstream tasks.

Dataset	Learning Rate	Hidden Factor	Dropout	Token Selection
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	3e-4	1	0.1	Mean

Table 9: 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 / 1e-1 (BCIC2020-3)
Scheduler	CosineAnnealingLR
Cosine cycle epochs	50
Minimal learning rate	1e-6
Clipping gradient norm	1
Label smoothing (multi-class classification)	0.1

H.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 [47] 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 [8], subjects 1-80 are used for training, 81-100 for validation, and 101-123 for testing, ensuring subject-independent evaluation.

The **SEED-V** dataset [48] 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 [49] 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 [50] 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 applied subject-independent, ensuring that train, validation, and test partitions are disjoint at the subject level.

The **SHU-MI** dataset [51] 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 [8], subjects 1-15 are used for training, 16-20 for validation, and 21-25 for testing.

The **BCIC2020-3** dataset [52] 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 [53] 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 8,860 labeled samples. We follow [8] for subject-wise splits to ensure reliable evaluation.

The **Mental Arithmetic** dataset [54] 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 [12]. Following [8], 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 [26]. 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.

H.3 Baselines

EEGNet [16]: 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 [17]: 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 [9]: 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 [6]: 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 [7]: 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 [8]: 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.

CodeBrain [18]: It is a two-stage EEG foundation model that combines decoupled time-frequency tokenization with multi-scale representation learning. Its TFDual-Tokenizer learns separate temporal and frequency codebooks, while EEGSSM combines structured global convolution and sliding-window attention to capture long-range dependencies and local temporal patterns through masked token prediction.

H.4 Evaluation Metrics

In this section, we describe the evaluation metrics adopted in this work. Following LaBraM [6, 8], 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.

I More Experimental Results

I.1 SEED-VII

On the SEED-VII dataset (Table 10), 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. EEGNet 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 mental-process-separation contribute to consistent robustness across diverse emotion recognition tasks.

I.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.

Table 10: Results on the SEED-VII dataset.

Methods	ACC-B	Kappa	F1-W
EEGNet	0.1892 ± 0.0056	0.0562 ± 0.0071	0.1472 ± 0.0123
Conformer	0.2719 ± 0.0032	0.1566 ± 0.0039	0.2712 ± 0.0033
BIOT	0.3005 ± 0.0054	0.1859 ± 0.0070	0.3055 ± 0.0069
EEGPT	0.2213 ± 0.0060	0.0954 ± 0.0071	0.2147 ± 0.0120
LaBraM	**0.3346** ± 0.0122	**0.2256** ± 0.0146	**0.3391** ± 0.0121
CBraMod	0.3318 ± 0.0056	0.2236 ± 0.0065	0.3384 ± 0.0059
CodeBrain	0.3333 ± 0.0064	0.2249 ± 0.0085	0.3386 ± 0.0081
BrainPro	0.3315 ± 0.0038	0.2223 ± 0.0055	0.3350 ± 0.0045

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

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 11: Results on the Major Depressive Disorder dataset.

Methods	ACC-B	AUC-PR	AUROC
EEGNet	0.9113 ± 0.0104	0.9632 ± 0.0045	0.9512 ± 0.0096
Conformer	0.8422 ± 0.0344	0.9442 ± 0.0502	0.9613 ± 0.0064
BIOT	0.8789 ± 0.0190	0.9744 ± 0.0083	0.9664 ± 0.0136
EEGPT	0.8475 ± 0.0933	0.9695 ± 0.0076	0.9669 ± 0.0069
LaBraM	0.8986 ± 0.0018	0.9791 ± 0.0041	0.9754 ± 0.0050
CBraMod	0.8909 ± 0.0037	0.9769 ± 0.0047	0.9726 ± 0.0062
CodeBrain	0.9079 ± 0.0027	0.9826 ± 0.0051	**0.9804** ± 0.0057
BrainPro	**0.9161** ± 0.0054	**0.9831** ± 0.0038	0.9803 ± 0.0046

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

I.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 achieves 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 metnal-process-separation effectively capture subtle neural dynamics underlying imagined speech.

Table 12: Results on the Speech dataset.

Methods	ACC-B	Kappa	F1-W
EEGNet	0.2755 ± 0.0183	0.0943 ± 0.0228	0.2737 ± 0.0180
Conformer	0.3339 ± 0.0104	0.1673 ± 0.0130	0.3269 ± 0.0153
BIOT	0.3016 ± 0.0206	0.1270 ± 0.0257	0.2999 ± 0.0217
EEGPT	0.2787 ± 0.0061	0.0983 ± 0.0076	0.2758 ± 0.0060
LaBraM	0.4880 ± 0.0161	0.3600 ± 0.0201	0.4871 ± 0.0165
CBraMod	0.5061 ± 0.0211	0.3827 ± 0.0263	0.5050 ± 0.0212
CodeBrain	0.5124 ± 0.0043	0.3906 ± 0.0054	0.5117 ± 0.0033
BrainPro	**0.5253** ± 0.0084	**0.4067** ± 0.0105	**0.5244** ± 0.0090

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

Table 13: Robustness under random channel dropping on FACED. Values are mean ± std; Δ denotes the absolute score change relative to no channel dropping. Bold indicates the highest mean in each row.

Drop	BrainPro		CBraMod		LaBraM	
	Mean ± std	Δ	Mean ± std	Δ	Mean ± std	Δ
ACC-B ↑						
0%	**0.5937** ± 0.0087	0	0.5669 ± 0.0094	0	0.5224 ± 0.0116	0
10%	**0.5747** ± 0.0014	−0.0190	0.5218 ± 0.0042	−0.0451	0.4972 ± 0.0036	−0.0252
20%	**0.5373** ± 0.0015	−0.0564	0.4890 ± 0.0054	−0.0779	0.4725 ± 0.0086	−0.0499
30%	**0.4995** ± 0.0056	−0.0942	0.4473 ± 0.0035	−0.1196	0.4230 ± 0.0085	−0.0994
40%	**0.4639** ± 0.0054	−0.1298	0.4111 ± 0.0094	−0.1558	0.3851 ± 0.0084	−0.1373
50%	**0.4270** ± 0.0076	−0.1667	0.3671 ± 0.0036	−0.1998	0.3397 ± 0.0036	−0.1827
Kappa ↑						
0%	**0.5418** ± 0.0092	0	0.5112 ± 0.0110	0	0.4610 ± 0.0126	0
10%	**0.5187** ± 0.0015	−0.0231	0.4585 ± 0.0045	−0.0527	0.4319 ± 0.0043	−0.0291
20%	**0.4772** ± 0.0016	−0.0646	0.4234 ± 0.0060	−0.0878	0.4049 ± 0.0101	−0.0561
30%	**0.4352** ± 0.0063	−0.1066	0.3775 ± 0.0039	−0.1337	0.3490 ± 0.0097	−0.1120
40%	**0.3949** ± 0.0054	−0.1469	0.3365 ± 0.0111	−0.1747	0.3057 ± 0.0094	−0.1553
50%	**0.3547** ± 0.0093	−0.1871	0.2876 ± 0.0037	−0.2236	0.2544 ± 0.0046	−0.2066
F1-W ↑						
0%	**0.6023** ± 0.0061	0	0.5729 ± 0.0105	0	0.5259 ± 0.0108	0
10%	**0.5785** ± 0.0020	−0.0238	0.5228 ± 0.0041	−0.0501	0.4977 ± 0.0041	−0.0282
20%	**0.5419** ± 0.0011	−0.0604	0.4919 ± 0.0050	−0.0810	0.4736 ± 0.0088	−0.0523
30%	**0.5053** ± 0.0055	−0.0970	0.4525 ± 0.0029	−0.1204	0.4226 ± 0.0082	−0.1033
40%	**0.4699** ± 0.0053	−0.1324	0.4165 ± 0.0094	−0.1564	0.3817 ± 0.0091	−0.1442
50%	**0.4329** ± 0.0078	−0.1694	0.3728 ± 0.0040	−0.2001	0.3349 ± 0.0038	−0.1910

I.4 Linear Probing on Representative Affective and Motor-related Tasks

To further evaluate the quality of pretrained representations, we conduct linear probing on two representative downstream tasks corresponding to the primary process branches in BrainPro: FACED for affective EEG decoding and SHU-MI for motor imagery decoding. The pretrained encoder is frozen and only a linear classifier is trained. This protocol evaluates the linear separability of downstream labels without task-specific encoder adaptation. As shown in Table 14, BrainPro achieves the strongest frozen-encoder performance among the evaluated EEG foundation-model baselines on both datasets. On FACED, BrainPro improves ACC-B from 0.3146 to 0.3581 over CBraMod. On SHU-MI, BrainPro improves ACC-B from 0.5233 to 0.5992 over LaBraM. These results provide additional evidence that BrainPro learns more linearly separable pretrained representations for affective and motor-related EEG decoding. Full fine-tuning results are reported in Table 1, where task-specific adaptation further improves performance.

Table 14: Linear probing results on representative affective and motor-related downstream tasks.

Method	FACED (9-Class)			SHU-MI (2-Class)		
	ACC-B	Kappa	F1-W	ACC-B	AUC-PR	AUROC
BIOT (Fixed)	0.2775 ± 0.0318	0.1839 ± 0.0512	0.2599 ± 0.0271	0.5123 ± 0.0206	0.5215 ± 0.0197	0.5146 ± 0.0264
LaBraM (Fixed)	0.3004 ± 0.0458	0.2377 ± 0.0617	0.2943 ± 0.0375	0.5233 ± 0.0272	0.5329 ± 0.0284	0.5218 ± 0.0275
CBraMod (Fixed)	0.3146 ± 0.0346	0.2579 ± 0.0542	0.3077 ± 0.0298	0.5217 ± 0.0247	0.5304 ± 0.0253	0.5238 ± 0.0317
BrainPro (Fixed)	**0.3581 ± 0.0052**	**0.2749 ± 0.0060**	**0.3564 ± 0.0055**	**0.5992 ± 0.0104**	**0.6289 ± 0.0018**	**0.6483 ± 0.0023**

J 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.

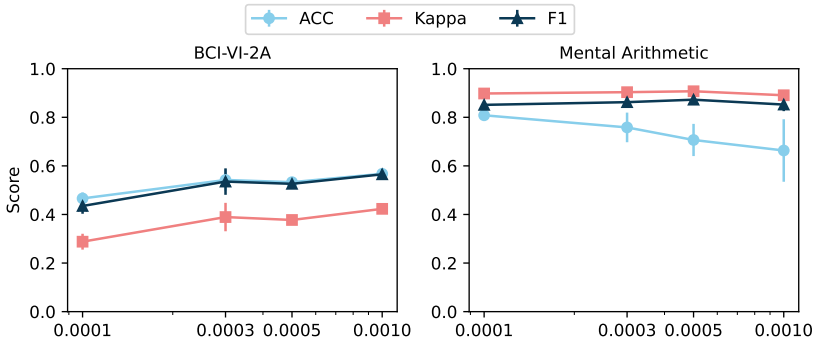


Figure 6: Effects of learning rate on BrainPro fine-tuning.

K Effect of Encoder Count and MLP Size on Performance

Using multiple encoders improves performance over single-encoder variants, while enlarging the MLP alone does not. Table 15 shows how downstream performance varies with different encoder counts and MLP sizes. Models that use only a single encoder perform worse on both FACED and BCI-IV-2A, even when their MLP size is increased to match or exceed that of multi-encoder configurations. In contrast, BrainPro, which uses two encoders, consistently outperforms the single-encoder variants despite having a comparable MLP ratio. Although adding a third encoder can provide additional gains in some settings, the main performance improvement arises from using more than one encoder rather than from simply increasing model capacity. These results indicate that architectural structure plays a larger role in performance than scaling the MLP alone, specifically incorporating multiple complementary encoders.

Table 15: Effect of encoder count and relative MLP size on performance. MLP Ratio is computed w.r.t. BrainPro.

Encoder No.	FACED				BCI-IV-2A			
	ACC-B	Kappa	F1-W	MLP Ratio	ACC-B	Kappa	F1-W	MLP Ratio
1	0.5635 ± 0.0031	0.5088 ± 0.0050	0.5710 ± 0.0049	0.51	0.5527 ± 0.0218	0.4036 ± 0.0290	0.5504 ± 0.0236	0.59
1	0.5808 ± 0.0048	0.5255 ± 0.0068	0.5840 ± 0.0079	1.02	0.5403 ± 0.0206	0.3871 ± 0.0275	0.5368 ± 0.0245	1.16
2	0.5937 ± 0.0087	0.5418 ± 0.0092	0.6023 ± 0.0061	1.00	0.5674 ± 0.0148	0.4232 ± 0.0198	0.5653 ± 0.0169	1.00
3	0.5938 ± 0.0069	0.5427 ± 0.0069	0.6039 ± 0.0042	0.93	**0.5889** ± 0.0165	**0.4519** ± 0.0220	**0.5868** ± 0.0157	1.07
3	**0.5988** ± 0.0057	**0.5477** ± 0.0058	**0.6098** ± 0.0043	**1.49**	0.5797 ± 0.0229	0.4396 ± 0.0305	0.5764 ± 0.0245	1.41

Note: **Bold** marks the best performance per dataset block. Cyan highlight marks BrainPro.

L Effects of Clear Data Pretraining

Figure 7 illustrates the effect of noisy samples during pretraining. 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 pretraining 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 pretraining. This emphasizes that high-quality pretraining data is critical for enabling BrainPro to learn robust and generalizable EEG representations.

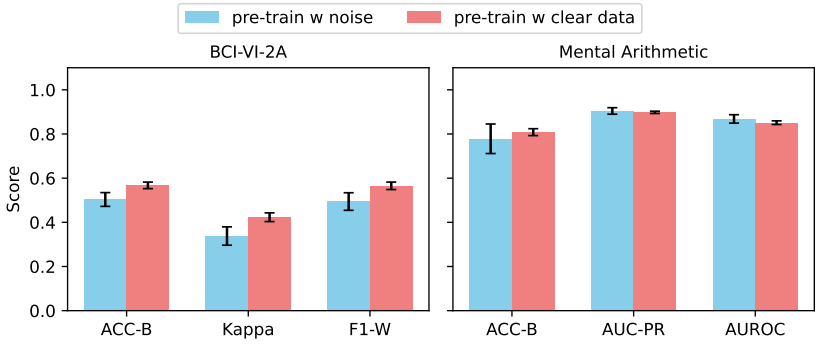


Figure 7: Effects of noisy samples on BrainPro pretraining.

M Effects of Resetting the Temporal Position Embedding

Table 16 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 pretraining. The motivation behind this strategy lies in the observation that, although pretraining captures general mental-process patterns, the temporal ordering of such patterns 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 16: Effect of resetting temporal position embedding on BCI-IV-2A and Mental Arithmetic datasets.

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

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

N Effects of Different Token Selection Mode

Table 17 compares different token selection strategies for constructing segment-level representations. On BCI-IV-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 17: 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	**0.5674** ± 0.0148	**0.4232** ± 0.0198	**0.5653** ± 0.0169	0.7367 ± 0.0315	0.8353 ± 0.0175	0.7822 ± 0.0141
Mean	0.4090 ± 0.0470	0.2120 ± 0.0627	0.3664 ± 0.0703	**0.8083** ± 0.0156	**0.8980** ± 0.0052	**0.8512** ± 0.0083
All	0.5373 ± 0.0315	0.3831 ± 0.0421	0.5295 ± 0.0335	0.5883 ± 0.0347	0.6753 ± 0.0131	0.6293 ± 0.0160

Note: Bold indicates the best performance for each dataset.

O Parameters size and FLOPS

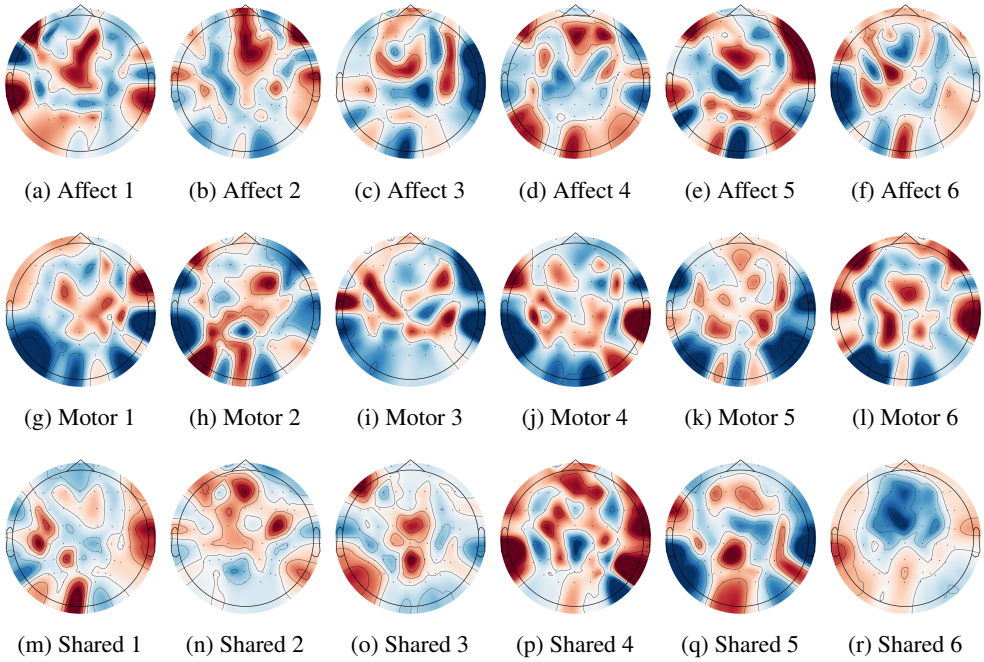
Table 18 compares the parameter counts and FLOPs of different EEG models, using the BCI-IV-2A dataset as the example for evaluation. Lightweight baselines such as EEGNet and Conformer have very few parameters (0.005M and 0.17M, respectively) and correspondingly low FLOPs, making them efficient but limited 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.

Table 18: Comparison of model complexity.

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

P Visualization of Learned Spatial Filters in Pretraining

As shown in Figure 8, the affect encoders predominantly emphasize frontal and temporal regions, which are consistent with neural substrates associated with emotional processing [13, 15], whereas the motor encoders focus on central and parietal regions, aligning with sensorimotor activity [14]. 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 mental processes. These results demonstrate that the proposed pretraining strategy learns spatial filters that not only support downstream performance but also provide qualitative evidence that the learned spatial filters are consistent with coarse neurophysiological priors.



(a) Affect 1 (b) Affect 2 (c) Affect 3 (d) Affect 4 (e) Affect 5 (f) Affect 6

(g) Motor 1 (h) Motor 2 (i) Motor 3 (j) Motor 4 (k) Motor 5 (l) Motor 6

(m) Shared 1 (n) Shared 2 (o) Shared 3 (p) Shared 4 (q) Shared 5 (r) Shared 6

Figure 8: Visualization of learned filters for three groups of encoders after pretraining: Affect (top row), Motor (middle row), and Shared (bottom row).

Q Additional Comparisons and Diagnostics

Q.1 Further Analysis on SEED-VII

Figure 9 examines branch selection and class-level performance on SEED-VII. S+A achieves the highest mean scores among the three branch configurations. Its ACC-B is 0.3315, compared with 0.3346 for LaBraM. The recall comparison shows class-dependent differences between BrainPro and LaBraM. These results identify SEED-VII as a benchmark for further improvement.

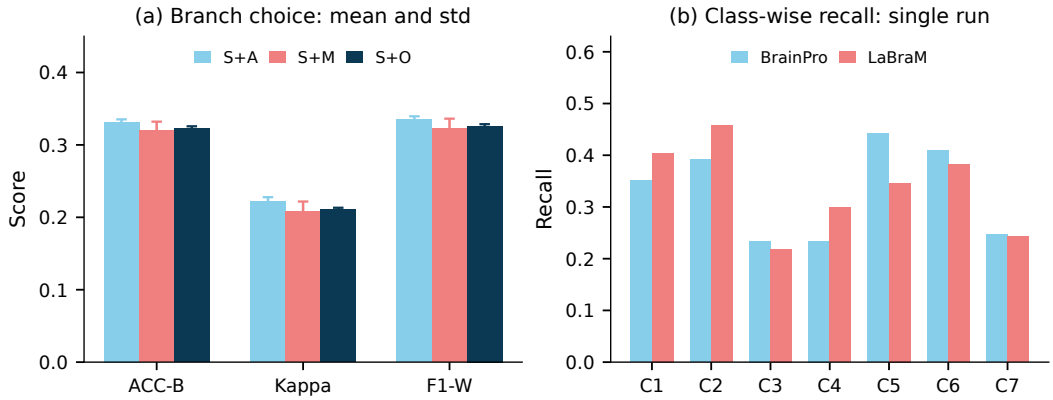


Figure 9: SEED-VII analysis. (a) Performance of three branch configurations. (b) Class recall (7 classes) calculated as $n_{ii}/\sum_j n_{ij}$ from single-run confusion matrices.

Q.2 Data Scale and Separation-margin Sensitivity

Data scaling. As shown in Figure 10, mean scores increase across the 10%, 20%, 50%, and 100% data settings on both tasks. FACED ACC-B increases from 0.5736 to 0.5937. BCI-IV-2A ACC-B increases from 0.5356 to 0.5674. These changes correspond to gains of 2.01 and 3.18 percentage

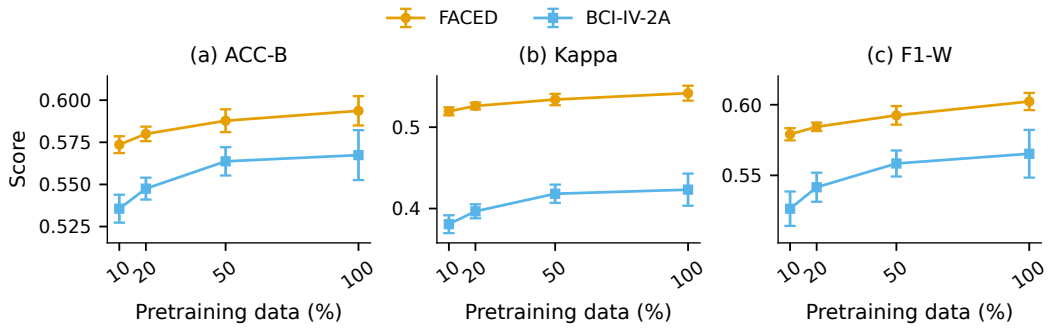


Figure 10: Effect of pretraining data size on FACED and BCI-IV-2A. Each panel uses a restricted vertical range to show performance changes.

points, respectively. Increasing the data fraction from 50% to 100% yields further gains of 0.59 and 0.37 percentage points. The full retained corpus contains 2,180.38 hours. Proportional estimates for the four data fractions are approximately 218, 436, 1,090, and 2,180 hours. The results show consistent improvements with increasing pretraining data within the evaluated range.

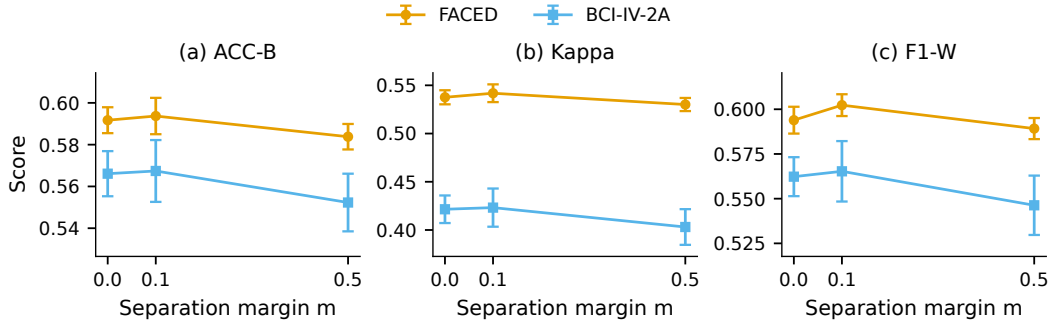


Figure 11: Effect of the separation margin on FACED and BCI-IV-2A. The three panels show ACC-B, Kappa, and F1-W.

Separation margin. The default margin is fixed at $m = 0.1$. We further compare $m \in \{0, 0.1, 0.5\}$ on both emotion and motor imagery tasks. As shown in Figure 11, margins of 0 and 0.1 yield similar mean scores. A margin of 0.5 yields lower mean scores. In $\max(\text{sim}_{\cos} - m, 0)$, a larger margin reduces the penalty for a given similarity. The observed performance decrease at $m = 0.5$ is consistent with weaker separation pressure.

Q.3 Region Weighting, Spatial retrieval, and Fusion

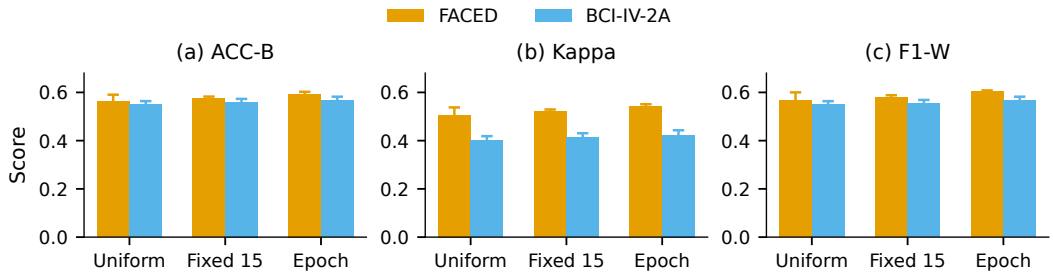


Figure 12: Comparison of regional weighting schedules on FACED and BCI-IV-2A. The settings use uniform weighting, fixed sharpness of 15, and epoch-dependent sharpness.

Region weighting. As shown in Figure 12, epoch-dependent sharpness achieves the highest mean scores among the three weighting settings on both tasks. Relative to fixed sharpness, ACC-B increases by 1.90 % on FACED and 0.74 % on BCI-IV-2A. The sharpness parameter controls the contrast between regional weights. The weights remain bounded between 0.5 and 1.5 throughout training. The schedule therefore increases regional emphasis within a fixed weight range.

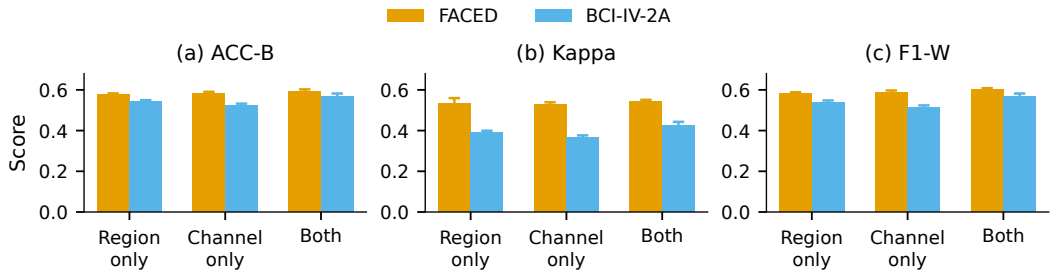


Figure 13: Comparison of spatial retrieval configurations on FACED and BCI-IV-2A. Region only uses regional retrieval. Channel only uses channel retrieval. Both combines the two spatial scales.

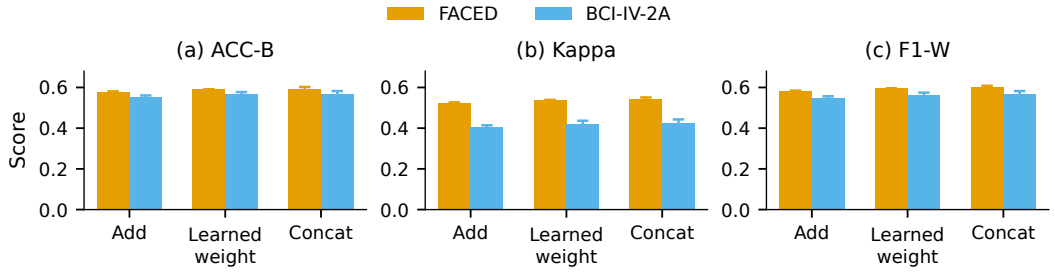


Figure 14: Comparison of feature fusion methods on FACED and BCI-IV-2A. The alternatives are addition, learned weighting, and concatenation.

Complementary spatial scales. Combined channel and region retrieval achieves the highest mean scores across all three metrics on both tasks shown in Figure 13. This result supports the complementary roles of the two spatial scales in the evaluated architecture.

Fusion. Concatenation retains fine-scale and coarse-scale features as separate components before Transformer encoding. As shown in Figure 14, concatenation achieves the highest mean scores among the three evaluated fusion methods. Learned weighting yields similar performance on BCI-IV-2A. Its ACC-B is 0.5667, compared with 0.5674 for concatenation. These results support concatenation in the evaluated configuration.

Q.4 Process Information and Dataset Identity

Process categories may correlate with source datasets, devices, laboratories, and preprocessing pipelines. We assess this confounding through dataset identity prediction within process families and process prediction across datasets.

Probe settings. We train class-balanced logistic regression probes on standardized, mean-pooled features from frozen encoders. Dataset identity prediction uses an approximately 80/20 split within each dataset. The test set is a randomly positioned block of sampled windows ordered by their original corpus indices to avoid data leakage. We evaluate five block-selection seeds. For process prediction, we use a leave-one-dataset-out protocol. Each probe is trained on the remaining datasets and evaluated on the held-out dataset. Predictions are pooled across held-out datasets to calculate accuracy, balanced accuracy, and macro-F1.

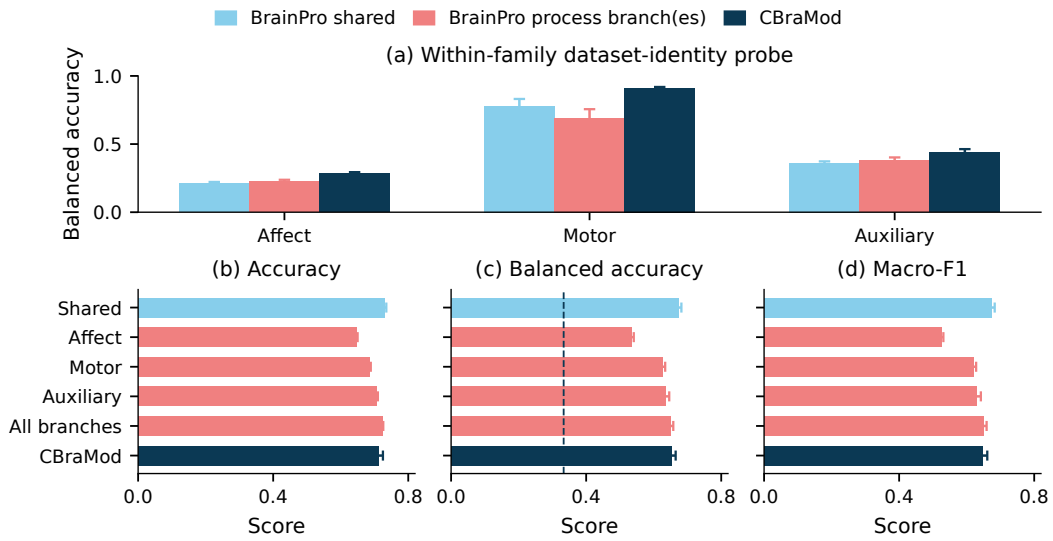


Figure 15: Linear probes on frozen representations. (a) Dataset identity prediction within each process family. The matched process branches are A, M, and O. Panels (b), (c), and (d) show process prediction with one dataset excluded from probe training. Dataset identity comparisons are made within each family. The dashed line indicates chance balanced accuracy of $1/3$.

Results. Figure 15(a) shows lower dataset identity prediction scores for the matched BrainPro branches than for CBraMod in all three process families. The affective and auxiliary branches yield higher scores than the shared encoder. The shared encoder achieves the highest process balanced accuracy of 0.675. CBraMod achieves 0.652. The process-specific encoders achieve balanced accuracies between 0.535 and 0.634. All exceed the three-class chance level of $1/3$. These results indicate that BrainPro representations contain both dataset-specific and process-associated information. Process prediction transfers across dataset boundaries, while dataset identity remains decodable within process families. Establishing independence from device, laboratory, and preprocessing effects requires evaluations that control these factors directly.

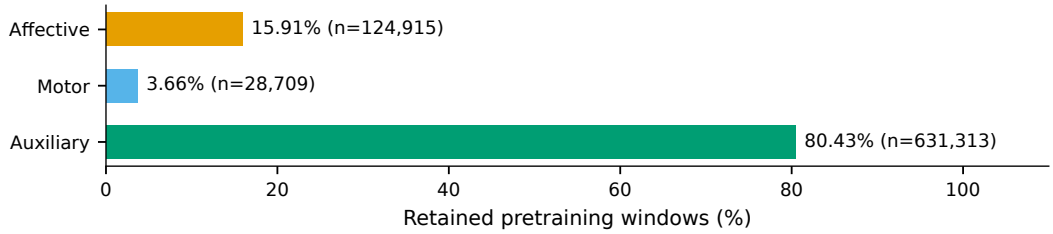


Figure 16: Branch exposure after data cleaning. Bars show each process family's proportion of retained pretraining windows. Labels give the corresponding window counts.

Q.5 Branch Activation

We quantify branch usage during pretraining to characterize the balance of encoder exposure. Data cleaning retains 784,937 windows, corresponding to 92.64% of generated windows and 2,180.38 hours. As shown in Figure 16, the affective, motor, and auxiliary groups contribute 124,915, 28,709, and 631,313 retained windows, respectively. Their proportions are 15.91%, 3.66%, and 80.43%. Under sampling proportional to retained window counts, these proportions give the activation probabilities of the corresponding process encoders. The shared encoder is active for every sample. Pretraining exposure is therefore uneven, with the auxiliary encoder receiving the most samples and the motor encoder receiving the fewest. Despite this imbalance, BrainPro improves downstream

decoding performance across the evaluated tasks. Balanced or hierarchical process sampling may further improve performance.

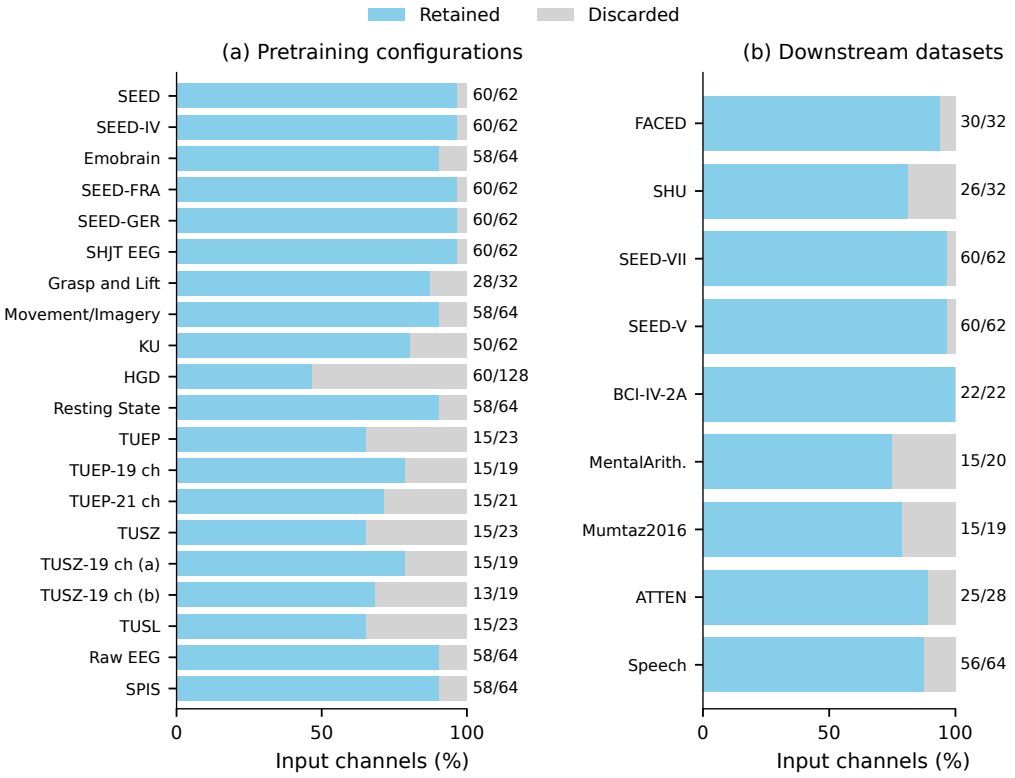


Figure 17: Montage coverage for (a) 19 pretraining configurations and (b) nine downstream datasets. Skyblue indicates retained channels. Light gray indicates discarded channels. Labels give retained and input channel counts. Retained template indices are unique in all evaluated mappings. Repeated TUEP and TUSZ rows represent distinct montage configurations.

Q.6 Montage Mapping

Figure 17 shows that the mapping preserves most input channels in all evaluated configurations except HGD. HGD retains 60 of 128 channels. The high retention across the other configurations supports broad montage compatibility while preserving channel coverage. All evaluated mappings have unique retained template indices. Unsupported channels are excluded.

R Discussion

New insights. This work shows that incorporating coarse process conditioning into large EEG models yields consistent improvements across diverse decoding tasks. The spatial filters learned by the affective, motor-related, and shared encoders form distinct and interpretable topographies, and the channel-drop analysis suggests that these spatial representations remain effective when electrodes are missing. The encoder-combination results indicate that process-associated branches capture task-aligned and complementary features beyond the shared encoder.

Limitations. Despite these advances, BrainPro has several limitations. First, the current process taxonomy is coarse, mainly covering affective, motor-related data. It does not explicitly model finer-grained processes such as attention, memory, language, or workload, which may limit branch specialization beyond the affective and motor-related settings emphasized in this work. Second, the auxiliary residual branch groups heterogeneous non-affective and non-motor data into a single pathway.

Future work. Several directions remain open. Extending BrainPro to richer process annotations, potentially through hierarchical process labels or dynamically allocated process-associated branches, could improve flexibility and interpretability. Incorporating multimodal neurophysiological signals such as fNIRS, MEG, eye tracking, or peripheral physiological signals may further enrich representation learning. Improving efficiency through parameter sharing, pruning, adapter-based tuning, or distillation will be important for real-time and on-device BCI applications. Overall, BrainPro provides a practical step toward scalable EEG pretraining that combines montage-aware spatial modeling with coarse process-conditioned representation learning.

S Broader impacts

BrainPro may improve robust EEG representation learning across heterogeneous montages and protocols, benefiting BCI, affective computing, and clinical decision-support research. However, EEG data are sensitive biometric and health-related signals, so use requires informed consent, secure data handling, and ethical oversight. Future use should evaluate bias across subjects, devices, and populations, and avoid non-consensual monitoring or high-stakes standalone decisions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions: topographic spatial retrieval, process-conditioned encoder branches, and process-guided region-weighted reconstruction. The empirical claims are limited to the reported BCI benchmarks and are supported by the main results, ablations, channel-drop analysis, encoder-configuration studies, and spatial-filter visualizations; we also clarify that BrainPro uses coarse mental-process cues and does not identify latent neural states.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a limitations discussion (Appendix R). It states that the current process taxonomy is coarse, mainly covers affective and motor-related data, does not explicitly model finer-grained processes such as attention, memory, language, and workload, and uses an auxiliary residual branch for heterogeneous non-affective and non-motor data.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present formal theoretical results as main contributions. It provides mathematical formulations and appendix-level justification to motivate retrieval-based spatial encoding in Appendix D, while the main claims are methodological and empirical rather than formal theoretical guarantees.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the model architecture, topographic spatial retrieval, process-conditioned branches, masking strategy, reconstruction and separation objectives, downstream adaptation protocol, datasets, preprocessing, evaluation metrics, data-split protocol, random seeds, optimizer, and hyperparameters. Additional details are provided in Section 4 and Appendices G-H, including dataset descriptions, pretraining settings, downstream settings, and baseline information.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The experiments use released EEG datasets, and code is provided.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper reports the pretraining corpus construction, preprocessing pipeline, segmentation length, resampling frequency, scaling procedure, masking ratio, optimizer, learning-rate schedule, number of epochs, model components, downstream datasets, evaluation metrics, non-overlapping train/validation/test split protocol, validation-based model selection, five random seeds, and hardware. Full hyperparameter settings are provided in Appendices G–H.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports mean $\pm$ standard deviation over five random seeds $\{0, 1, 2, 3, 4\}$ for the main downstream results, ablation studies, encoder-configuration analysis, and appendix experiments. The standard deviation is computed across the five independent runs and captures variability due to random initialization and training randomness under the same experimental protocol.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports that pretraining was conducted on 5 NVIDIA A800 GPUs for approximately 3 days, corresponding to roughly 360 A800 GPU-hours, on a machine with approximately 512 GB RAM and 1 TB storage (Appendix G.5).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

Justification: The study uses existing public or controlled-access EEG datasets and does not involve new data collection, deception, crowdsourcing, or intervention on human participants. We follow provider-specific dataset terms, use non-overlapping train/validation/test splits to reduce data leakage, and discuss limitations and responsible-use considerations.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses potential benefits for EEG representation learning, BCI research, assistive technologies, and robust decoding under heterogeneous montages in Appendix S. It also acknowledges potential risks of EEG modeling, including privacy concerns, misuse for intrusive monitoring or affective inference, and limitations for clinical or high-stakes deployment without additional validation.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: We do not release a high-risk generative model, scraped dataset, or system intended for deployment in high-stakes settings. The paper focuses on EEG representation learning and discusses responsible-use considerations, including that the model should not be used for clinical diagnosis, surveillance, or high-stakes decision-making without proper consent, privacy protection, and task-specific validation.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: We cite the original sources for the public and controlled-access EEG datasets and baseline methods used in this study, and dataset details are summarized in the Appendix G.2 and H.2. Because the study uses many third-party EEG datasets with heterogeneous access conditions, including datasets requiring registration or application, we do not enumerate every dataset license in the checklist and do not redistribute third-party datasets; users should obtain each dataset from its original provider and comply with the corresponding terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce the BrainPro model and document its architecture, pretraining objective, topographic spatial retrieval module, downstream adaptation protocol, preprocessing settings, hyperparameters, and evaluation protocol in the main paper and appendices. Any released implementation materials are intended to document and facilitate use of the core BrainPro components, while third-party EEG datasets must be obtained from their original providers.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We do not conduct crowdsourcing experiments or collect new human-subject data. All experiments use existing public or controlled-access EEG datasets, and participant instructions, compensation, and consent procedures are governed by the original dataset collection studies cited in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve new human-subject recruitment, intervention, or data collection. The experiments use existing public or controlled-access EEG datasets, for which ethics approvals, consent procedures, and risk disclosures are governed by the original dataset providers and cited dataset papers.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method, experiments, and scientific contributions do not involve LLMs as an important, original, or non-standard component. Any use of standard writing, editing, or formatting assistance, if applicable, does not affect the methodology, experimental design, scientific claims, or originality of the research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.