DEEPERBRAIN: A NEURO-GROUNDED EEG FOUNDATION MODEL TOWARDS UNIVERSAL BCI

Anonymous authors
Paper under double-blind review

ABSTRACT

Electroencephalography (EEG) foundation models promise general-purpose representations for Brain-Computer Interfaces (BCIs), yet most transfer effectively only under end-to-end fine-tuning and degrade markedly under *frozen-probing*—a protocol that diagnoses whether task-relevant information is directly accessible in a fixed representation. We attribute this gap to the reliance on generic sequence architectures that overlook the biophysical and dynamical structure of neural activity. We propose **DeeperBrain**, an EEG foundation model that injects neurophysiological inductive biases at two levels. Architecturally, a volume-conduction-aware channel encoding models spatial mixing from 3D electrode geometry, and a neurodynamics-aware temporal encoding captures slow oscillations and adaptive decay. For pretraining, we combine Masked EEG Reconstruction (MER) with Neurodynamics Statistics Prediction (NSP), an auxiliary self-supervised objective whose targets—spectral power, phase synchronization, cross-frequency coupling, and signal complexity—are computed directly from the signal and aligned with established neurophysiological descriptors. Pretrained on over 17,200 hours of heterogeneous EEG and evaluated on 12 downstream datasets spanning emotion, motor, sleep, seizure, speech, disorder, vigilance, and workload tasks, DeeperBrain achieves state-of-the-art or competitive fine-tuning performance and, notably, consistently superior frozen-probing performance. The results indicate that neurophysiological priors make downstream information more accessible without backbone adaptation, a property we view as a necessary step toward universal BCI. Our code will be publicly available on Github.

1 INTRODUCTION

Brain-Computer Interfaces (BCIs) establish direct communication pathways between the central nervous system and external devices (Schalk et al., 2004; Wu et al., 2013; Zhang et al., 2019b). Electroencephalography (EEG) is the predominant non-invasive modality for BCI, offering millisecond temporal resolution, portability, and cost-efficiency, and underpins applications from motor imagery (Altaheri et al., 2021; Dai et al., 2020) and affective recognition (Dadebayev et al., 2022; Gao et al., 2024; Zhao et al., 2025) to epilepsy diagnosis (Ahmad et al., 2022; Yıldız et al., 2022), sleep staging (Phan et al., 2021; Wang et al., 2024b; Zhou et al., 2024a), and depression detection (Yasin et al., 2021; Wang et al., 2025c). Achieving a *universal* BCI, however, remains challenging. EEG recordings are non-stationary and noisy (Saha & Baumert, 2020; Björk & Kugelberg, 1953; Mannan et al., 2018; Urigüen & Garcia-Zapirain, 2015; Jamil et al., 2021); acquisition hardware varies in channel count, reference scheme, and spatial alignment (Acharya & Acharya, 2019); and BCI paradigms recruit diverse neural mechanisms. These factors collectively impede cross-scenario generalization.

Inspired by large language and vision models (Bommasani et al., 2021; Min et al., 2023; Awais et al., 2025; Xu et al., 2025), the field has turned to EEG foundation models pretrained on large unlabeled corpora (Zhou et al., 2025a; Li et al., 2025; Gu et al., 2025; Lai et al., 2025; Yang et al., 2025), typically via masked signal reconstruction. Despite their statistical capacity, these data-driven models rely on ***end-to-end fine-tuning*** to adapt to downstream tasks and frequently exhibit limited efficacy under ***frozen-probing***, where the pretrained backbone is fixed and only a lightweight head is trained (Lee et al., 2025; Xiong et al., 2025; Wu et al., 2025). We argue that strong frozen-probing performance is a necessary diagnostic on the path toward universal BCI: it asks whether task-relevant

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

information is accessible in the fixed representation without scenario-specific parameter updates. We attribute the current deficiency to the reliance on generic sequence modeling, which prioritizes data-driven patterns while underutilizing the biophysical and dynamical principles that structure EEG.

Two principles shape this structure. First, *spatial organization* is governed by *volume conduction*: passive current spread through head tissues induces distance-dependent correlations among electrodes (Nunez & Srinivasan, 2006). Second, *temporal organization* spans a hierarchy from millisecond oscillations to second-level cognitive adaptations (Buzsáki, 2006; Fairhall et al., 2001). From a dynamical-systems perspective, macroscopic brain function is further characterized by low-dimensional *order parameters* such as spectral power (Buzsáki, 2006), functional connectivity (Varela et al., 2001), cross-frequency coupling (Canolty et al., 2006), and dynamic complexity (Richman & Moorman, 2000). Neglecting these priors yields neurobiologically agnostic representations that require extensive adaptation to recover meaningful features.

To bridge this gap, we propose **DeeperBrain**, a neuro-grounded EEG foundation model that integrates neurophysiological inductive biases into both architecture and learning objective. Architecturally, DeeperBrain uses **a channel positional encoding that models volume conduction via 3D electrode geometry and a learnable spatial decay kernel**, together with **a temporal positional encoding built from oscillatory and exponential-decay bases** that capture slow modulations and adaptation. For pretraining, we devise a dual-objective combining Masked EEG Reconstruction (MER) with **Neurodynamics Statistics Prediction (NSP)**, an auxiliary self-supervised objective whose targets are computed directly from the signal and aligned with the order parameters above. By explicitly encoding these physical and dynamical constraints, DeeperBrain encourages representations whose geometry makes downstream information accessible even under frozen-probing.

Our contributions are:

1. We propose **DeeperBrain**, a neuro-grounded EEG foundation model that systematically incorporates neurophysiological inductive biases into model design and learning objective. Built on the masked-reconstruction paradigm, its novelty lies in the prior injection rather than a new backbone.
2. We design **neurophysiologically grounded positional encodings**, a volume-conduction-aware channel encoding that models spatial mixing via 3D geometry, and a neurodynamics-aware temporal encoding that captures slow oscillations and adaptive dynamics via biologically plausible bases.
3. We devise **neurodynamics-guided pretraining**, a dual-objective strategy combining masked EEG reconstruction with neurodynamics statistics prediction, an auxiliary self-supervised objective that aligns representations with established neurophysiological descriptors.
4. We demonstrate **strong transferability under frozen-probing**. Pretrained on 14 diverse datasets, DeeperBrain achieves state-of-the-art or competitive fine-tuning performance across 12 downstream tasks and, notably, consistently superior frozen-probing performance, which we interpret as evidence that neurophysiological priors improve the accessibility of task-relevant information in a fixed representation.

2 METHOD

2.1 OVERVIEW

DeeperBrain is a neuro-grounded foundation model that learns transferable representations from unlabeled EEG. As illustrated in Fig. 1, the pipeline partitions the channel-reorganized EEG into non-overlapping spatiotemporal patches, masks a random subset, and projects the visible ones into dense tokens via a temporal convolutional backbone. We augment these tokens with two domain-specific positional encodings: a channel encoding that models volume conduction using 3D electrode geometry, and a temporal encoding that captures slow oscillatory and adaptive dynamics. A Transformer encoder then models global dependencies. The model is optimized with a dual-objective strategy: Masked EEG Reconstruction (MER) preserves local signal fidelity, while Neurodynamics Statistics Prediction (NSP) aligns representations with macroscopic neurophysiological descriptors.

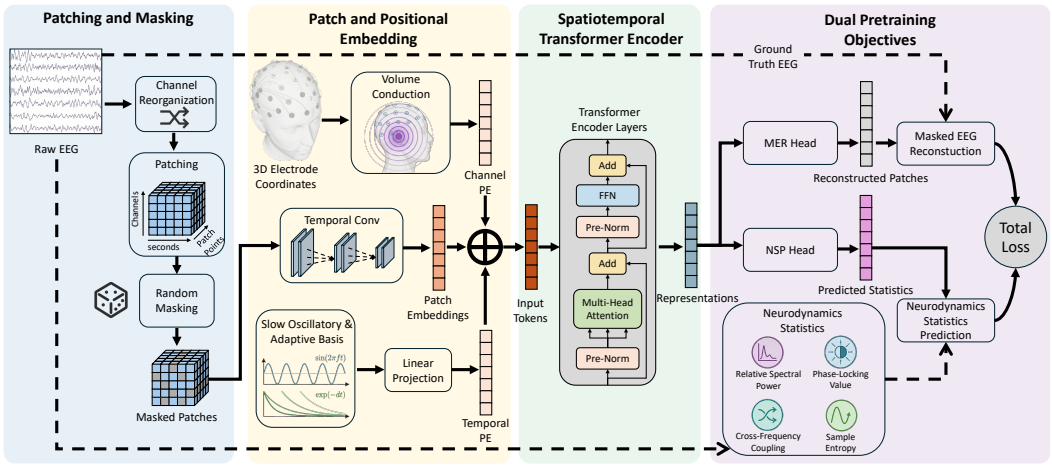


Figure 1: DeeperBrain overview.

2.2 Patching and Masking

To handle variable electrode montages, we apply stochastic *channel reorganization*: randomly permuting and subsampling channels to simulate variable-density recordings. The reorganized EEG is partitioned into non-overlapping one-second patches $\mathbf{X}_{\text{patch}} \in \mathbb{R}^{C\times S\times P}$ with $P = f_s$, the atomic unit for self-supervised learning. During pretraining, a fraction $\mu = 0.5$ of the $C \times S$ patches is masked uniformly and replaced with zero vectors in the raw voltage space. Formal notation in Appendix A.

2.3 Patch Embedding via Temporal Convolution

A lightweight temporal convolutional network transforms raw voltage patches into dense embeddings $\mathbf{E} \in \mathbb{R}^{C\times S\times D}$ via three convolutional blocks with Group Normalization and GELU. The first block (K_1=49, s_1=25) captures long-range structure, mimicking delta–theta bandpass filtering; subsequent blocks (K_2=K_3=3, unit stride) refine local features. Full architecture in Appendix A.

2.4 Neurophysiologically Grounded Positional Encodings

Standard positional encodings assume fixed geometric priors (e.g., a 2D grid in ViT (Dosovitskiy et al., 2020)) or purely learnable embeddings. EEG admits stronger biophysical inductive biases.

Channel encoding: modeling volume conduction. Scalp potentials arise from passive current spread through conductive tissues (Nunez & Srinivasan, 2006), so signal similarity decays with inter-electrode distance. Using 3D coordinates $\mathbf{P}_{\text{ch}} \in \mathbb{R}^{C\times 3}$ (MNI space), with pairwise distances $\mathbf{D}_{ij}$, we define a learnable spatial decay kernel (Grech et al., 2008):

$$\mathbf{K}_{ij} = \exp\left(-\frac{\mathbf{D}_{ij}}{\tau}\right), \quad \tau = \text{softplus}(\alpha) + \delta, \tag{1}$$

where τ is the decay scale (initialized to $\sim$8 cm). After row normalization, the smoothing $\tilde{\mathbf{p}}_i = \sum_j \bar{\mathbf{K}}_{ij}\mathbf{p}_j$ embeds each electrode in a *functional coordinate system*—a proximity-weighted average of anatomical locations that mimics volume-conducted mixing—and is projected to dimension D. Biophysical derivation in Appendix A.

Temporal encoding: slow modulations and adaptation. We construct a second-level temporal encoding from oscillatory and exponential-decay bases that capture the dissipative dynamics of neural adaptation (Buzsáki, 2006; Fairhall et al., 2001):

$$\psi_k^{\text{osc}}(t) = [\sin(2\pi f_k t),\, \cos(2\pi f_k t)], \qquad \psi_m^{\text{dec}}(t) = \exp(-d_m t), \tag{2}$$

where f_k spans 0.01–0.5 Hz (slow cognitive fluctuations) and d_m spans adaptation constants $\tau \in [1, 100]$ s. Both bases are held fixed during training and linearly projected to D dimensions. Biological motivation in Appendix A.

2.5 Spatiotemporal Transformer Encoder

The tokens $\mathbf{Z}^{(0)} \in \mathbb{R}^{N \times D}$, $N = C \times S$, are processed by L pre-normalization Transformer layers (Wang et al., 2019):

$$\widetilde{\mathbf{Z}} = \mathbf{Z}^{(l-1)} + \mathrm{MHA}(\mathrm{LN}(\mathbf{Z}^{(l-1)})), \tag{3}$$

$$\mathbf{Z}^{(l)} = \widetilde{\mathbf{Z}} + \mathrm{FFN}(\mathrm{LN}(\widetilde{\mathbf{Z}})), \tag{4}$$

where MHA is multi-head self-attention and FFN a two-layer network with GELU. Self-attention lets each $(\text{channel}, \text{second})$ token integrate information across the spatiotemporal grid, modeling long-range interactions analogous to functional connectivity. The final output $\mathbf{H} = \mathbf{Z}^{(L)} \in \mathbb{R}^{N \times D}$ yields a contextualized representation encoding both local dynamics and global network structure.

2.6 Masked EEG Reconstruction (MER)

To preserve fine-grained temporal fidelity essential for transient events (epileptic spikes, sleep spindles), $\mathbf{H}$ is mapped back to the raw voltage space via a single linear head:

$$\hat{\mathbf{X}}_{\text{flat}} = \mathbf{H}\mathbf{W}_{\text{recon}}^{\top} + \mathbf{b}_{\text{recon}} \in \mathbb{R}^{N \times P}. \tag{5}$$

Reconstruction loss is computed only on masked patches $\mathcal{M} = \{(c, s) \mid \mathbf{M}_{c,s} = 1\}$ using Smooth L1 (Huber) loss (Huber, 1992):

$$\mathcal{L}_{\text{MER}} = \frac{1}{|\mathcal{M}|P} \sum_{(c,s)\in\mathcal{M}} \sum_{p=1}^{P} \rho(\hat{\mathbf{X}}_{c,s,p} - \mathbf{X}_{c,s,p}), \tag{6}$$

with $\beta = 1.0$. This balances precise waveform recovery (MSE-like for small residuals) with robustness to high-amplitude artifacts (MAE-like for large residuals).

2.7 Neurodynamics Statistics Prediction (NSP)

While MER preserves local fidelity, it does not explicitly enforce the modeling of multi-scale dynamical structure. We introduce *Neurodynamics Statistics Prediction* (NSP), an auxiliary self-supervised objective grounded in dynamical-systems theory (Strogatz, 2024; Breakspear, 2017). The brain is conceptualized as a high-dimensional nonlinear system whose macroscopic behavior is characterized by low-dimensional *order parameters* (Buzsáki, 2010). NSP targets are computed directly from the EEG signal via a fixed mapping $\Psi(\cdot)$, without human labels or teacher networks—analogous to MAE's reconstruction prior (He et al., 2022) but drawn from neurophysiology rather than pixel identity.

$\Psi(\cdot)$ extracts four neurophysiologically grounded statistics serving as proxies for canonical order parameters: **(1)** relative spectral power (5D), reflecting the attractor landscape (Buzsáki, 2006); **(2)** phase-locking value (10D), measuring large-scale functional integration (Varela et al., 2001); **(3)** cross-frequency coupling (3D), encoding coordination across temporal scales (Canolty et al., 2006); and **(4)** sample entropy (1D), capturing signal complexity (Richman & Moorman, 2000). These form the 19-dimensional target $\mathbf{Y}_{\text{NS}}$, predicted via a linear head with Smooth L1 loss:

$$\mathcal{L}_{\text{NSP}} = \frac{1}{CS \cdot 19} \sum_{c,s,k} \rho(\hat{\mathbf{Y}}_{\text{NS},c,s,k} - \mathbf{Y}_{\text{NS},c,s,k}). \tag{7}$$

The total objective is $\mathcal{L} = \lambda_{\text{MER}}\mathcal{L}_{\text{MER}} + \lambda_{\text{NSP}}\mathcal{L}_{\text{NSP}}$ with $\lambda_{\text{MER}} = \lambda_{\text{NSP}} = 1.0$. Detailed statistic definitions and formulas in Appendix A.

2.8 Task-specific Prediction Head

To evaluate representations, we attach a task-specific head to $\mathbf{H} \in \mathbb{R}^{N \times D}$. The head is selected from a small family depending on the task's temporal structure: (i) global average pooling over patches

Table 1: Pretraining datasets.

Dataset	Hz	Ch	# Smp	Hours
TUEG (clean) (Obeid & Picone, 2016)	256	19	1,109,545	9,246.2
PhysioNet 2018 (Goldberger et al., 2000a; Ghassemi et al., 2018)	200	6	872,089	7,267.4
Raw EEG Data (Trujillo, 2020)	256	64	30,978	42.7
Siena Scalp EEG (Detti et al., 2020)	256	27	50,749	140.8
10 OpenNeuro ds (OpenNeuro, 2025)	250–2048	28–128	375,292	545.9
Total			**2,438,653**	**17,243.0**

Table 2: Downstream BCI tasks (abbreviations used in Tables 3–4). Per-sample duration and split details in Appendix C.

Task	Dataset (abbr.)	Ch	# Smp	Labels
Emotion	FACED (FAC)	32	10,332	9-class
	SEED-V (S-V)	62	117,744	5-class
	SEED-VII (S-VII)	62	55,561	7-class
Motor imagery	PhysioNet-MI (PN-MI)	64	9,837	4-class
	BCIC-IV-2a (2a)	22	5,184	4-class
	SHU-MI (SHU)	32	11,988	2-class
Sleep	ISRUC (ISR)	6	89,240	5-class
Seizure	CHB-MIT (CHB)	16	326,993	2-class
Speech	BCIC2020-3 (BCI)	64	6,000	5-class
Disorder	MODMA (MOD)	128	1,066	2-class
Vigilance	SEED-VIG (VIG)	17	20,355	regression
Workload	MentalArith (MA)	20	1,707	2-class

followed by a linear classifier, or (ii) full-sequence flattening followed by a linear (or shallow MLP) classifier. We assess performance under *end-to-end fine-tuning* (entire model trained) and *frozen-probing* (backbone frozen, head only trained). The same head family and optimization budget are applied to all baselines for fair comparison.

3 Experiments

3.1 Pretraining Setup

Corpus. DeeperBrain is pretrained on 14 public datasets totaling over 17,200 hours (2,438,653 samples); see Table 1. The corpus spans clinical diagnostics, sleep, cognitive neuroscience, and perceptual tasks, with substantial heterogeneity in channel count (6–128), sampling rate (200–2048 Hz), and epoch duration (5–30 s). TUEG (Obeid & Picone, 2016) and PhysioNet 2018 (Goldberger et al., 2000a; Ghassemi et al., 2018) together account for approximately 96% of the total; the remaining twelve datasets—ten from OpenNeuro (OpenNeuro, 2025) (individually cited in Appendix B) and two clinical sources (Detti et al., 2020; Trujillo, 2020)—contribute diversity in montage and paradigm. The channel-reorganization step and the NSP prior partly mitigate this imbalance.

Preprocessing. All recordings are resampled to 200 Hz (PhysioNet 2018 already at 200 Hz), band-pass filtered (0.3–75 Hz), notch-filtered at 50/60 Hz, segmented into non-overlapping fixed-length samples, and normalized to a 100 μV scale following (Jiang et al., 2024b). TUEG is curated to a standard 19-channel 10–20 montage across its three sub-corpora, with edge segments trimmed; full per-dataset details are in Appendix B.

Implementation. PyTorch 2.8.0, CUDA 12.9, one NVIDIA RTX A5000; pretraining completes in $\sim$7 hours. Backbone: 12-layer Transformer, D=200, D_{ff}=800, 8 heads. Optimization: 1 s patches (200 samples), 50% mask ratio, 2 epochs, batch 16, AdamW (lr 5×10^{-4}, wd 5×10^{-2}), gradient clip norm 1.0. NSP pseudo-labels are precomputed and cached.

3.2 Downstream Tasks and Protocol

We evaluate on 12 datasets spanning eight BCI task categories (Table 2 and Appendix C). All signals are resampled to 200 Hz with 1 s patches. Splits follow LaBraM (Jiang et al., 2024b) or CBraMod (Wang et al., 2025a) where available for comparability; otherwise a subject-level

Table 3: End-to-end fine-tuning. Primary metric per dataset (%): Balanced Accuracy for classification; Pearson correlation $\times 100$ for SEED-VIG. **Bold** = best, underline = second best. Full multi-metric results in Appendix E.

	Emotion			Motor			Sleep	Seiz.	Speech	Dis.	Vig.	Wkld
Method	FAC	S-V	S-VII	PN-MI	2a	SHU	ISR	CHB	BCI	MOD	VIG	MA
EEGNet	40.90	29.61	25.28	58.14	44.82	58.89	71.54	56.58	44.13	61.90	57.01	67.70
EEGConf.	45.59	35.37	28.75	60.49	46.96	59.00	74.00	59.76	45.06	63.85	57.50	68.05
LaBraM	52.73	39.76	33.28	61.73	48.69	61.66	76.33	70.75	50.60	66.81	59.31	72.24
CBraMod	55.09	40.91	32.47	64.17	51.38	63.70	78.65	73.98	53.73	70.18	58.49	74.39
CSBrain	57.52	**41.97**	32.44	63.04	**56.57**	64.17	**79.25**	72.62	56.05	72.85	**63.14**	75.58
REVE	56.46	40.48	32.64	64.80	63.96	63.84	78.19	72.02	56.35	72.45	58.77	**76.60**
DeeperBrain	**60.32**	40.68	**33.60**	**65.17**	60.69	**64.31**	78.75	**74.34**	**57.06**	**76.31**	60.11	75.69

Table 4: Frozen-probing. Primary metric per dataset (%): Balanced Accuracy for classification; Pearson correlation $\times 100$ for SEED-VIG. **Bold** = best, underline = second best.

	Emotion			Motor			Sleep	Seiz.	Speech	Dis.	Vig.	Wkld
Method	FAC	S-V	S-VII	PN-MI	2a	SHU	ISR	CHB	BCI	MOD	VIG	MA
LaBraM	16.13	32.63	15.35	46.52	38.11	59.48	71.79	55.31	24.83	43.12	56.92	60.69
CBraMod	25.84	24.36	17.32	53.73	33.06	57.18	37.27	58.93	20.24	46.00	37.19	54.79
CSBrain	36.59	20.00	14.28	26.19	26.70	53.05	22.24	62.68	32.19	52.38	22.31	54.65
REVE	37.76	27.82	14.29	25.01	42.73	59.38	73.55	63.93	39.04	53.00	50.32	67.29
DeeperBrain	**50.96**	**35.08**	**23.26**	**56.57**	**51.01**	**60.81**	**74.10**	**64.95**	**43.33**	**63.36**	**59.85**	**68.05**

train/val/test split is used. We compare against two supervised baselines (EEGNet (Lawhern et al., 2018), EEGConformer (Song et al., 2022)) and four foundation-model baselines (LaBraM (Jiang et al., 2024b), CBraMod (Wang et al., 2025a), CSBrain (Zhou et al., 2025b), REVE (Ouahidi et al., 2025)). Compared baselines differ in pretraining scale, parameter count, and tokenization (Appendix D); the comparison is therefore a practical horizontal benchmark rather than a strictly controlled ablation. For each protocol we report the primary monitor metric (Balanced Accuracy for classification; Pearson correlation for SEED-VIG regression), with full multi-metric results in Appendix E. All results are mean$\pm$std over five seeds.

3.3 Results under End-to-End Fine-tuning

Table 3 reports the primary metric per dataset. DeeperBrain achieves the best result on 8 of 12 datasets and is competitive on the remainder, never falling below third. The results reveal three properties of the learned representation.

Pretraining extracts generalizable structure. DeeperBrain consistently outperforms supervised baselines (EEGNet, EEGConformer) by a large margin across all datasets. This gap validates that large-scale self-supervised pretraining captures transferable structure that supervised training on limited task-specific data cannot recover—the central premise of foundation models applied to EEG.

Neurophysiological alignment broadens transfer. Against foundation-model baselines, DeeperBrain's gains are most pronounced where the target phenomena align with the NSP statistics. On MODMA (depression detection) and CHB-MIT (seizure detection), where spectral abnormalities and connectivity disruptions are clinically diagnostic, the NSP objective's emphasis on spectral organization, phase synchronization, and signal complexity provides a representational substrate that reconstruction-only pretraining does not. The advantage is not confined to clinical tasks: on FACED (emotion) and BCIC2020-3 (speech imagery), where rapid cognitive state transitions recruit distributed oscillatory networks, the same neurodynamical structure appears to facilitate decoding. This consistency across paradigms—from transient epileptic events to sustained affective states—suggests that the NSP targets function as a shared vocabulary of brain activity, enabling the model to abstract away from paradigm-specific waveforms toward invariant dynamical features.

Positional encodings mitigate montage heterogeneity. The robustness across datasets spanning 6 to 128 channels highlights the practical value of the channel encoding. By modeling volume conduction via 3D electrode geometry rather than fixed channel indices, DeeperBrain generalizes

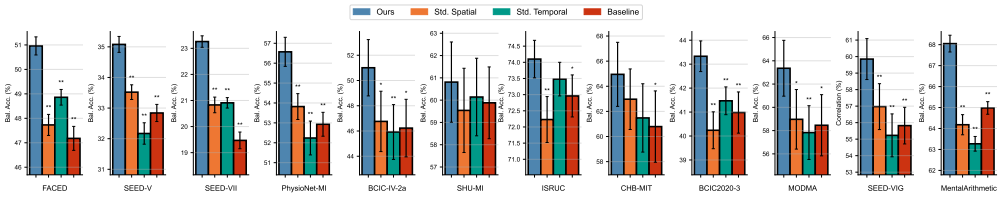


Figure 2: Positional-encoding ablation under frozen-probing (balanced accuracy). FT results and significance markers in Appendix F.

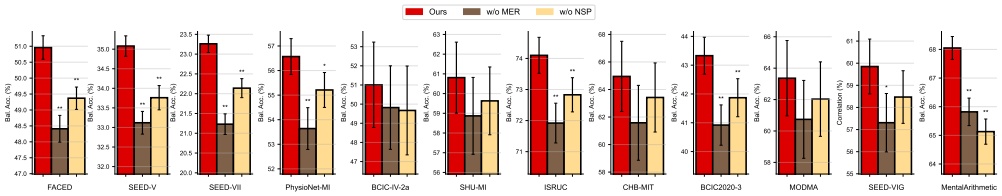


Figure 3: Pretraining-objective ablation under frozen-probing (balanced accuracy). FT results in Appendix F.

to unseen montages without architecture modification. The temporal encoding similarly aids tasks with distinct temporal characteristics—whether the target features are transient spikes (CHB-MIT), second-level sleep architectures (ISRUC), or sustained cognitive states (SEED-VIG). These results indicate that aligning the architecture's spatial and temporal structure with the biophysical and dynamical properties of EEG yields representations that transfer more broadly than those derived from generic sequence modeling.

3.4 Results under Frozen-Probing

Table 4 reports frozen-probing performance, where the backbone is fixed and only the head is trained. We frame frozen-probing as a *diagnostic* of whether task-relevant information is accessible in the fixed representation without backbone adaptation. Under this diagnostic, DeeperBrain achieves the best result on all 12 datasets, often by a wide margin.

Frozen features encode neurodynamical invariants. Unlike end-to-end fine-tuning, which allows the backbone to re-optimize for task-specific distributions, frozen-probing tests whether the pretrained features are separable and universally applicable out-of-the-box. The results indicate that DeeperBrain does not merely memorize statistical patterns from the pretraining corpus but encapsulates core neurodynamical properties—oscillatory states, functional integration, signal complexity—that are invariant across subjects and paradigms. The frozen features alone are sufficient to support high-performance decoding ranging from emotion recognition to mental workload estimation, where several baselines collapse to near-chance performance (e.g., LaBraM on FACED, 16.13% for a 9-class task). This suggests that the NSP objective, by explicitly aligning the representation with macroscopic order parameters during pretraining, imbues the feature space with a structure that downstream linear heads can readily exploit.

Smaller fine-tuning→frozen gap. While fine-tuning generally yields the highest absolute performance by adapting to task-specific nuances, the performance drop when switching to frozen-probing is remarkably smaller for DeeperBrain than for baselines. This resilience indicates that the dual-objective strategy—combining signal reconstruction with neurodynamics prediction—forces the model to disentangle meaningful physiological factors from noise during pretraining itself, rather than relying on downstream adaptation to recover this structure. The practical implication is that DeeperBrain can serve as a robust, off-the-shelf feature extractor for new tasks with minimal computational overhead and reduced risk of overfitting to small calibration datasets—a setting common in clinical BCI deployment where labeled recordings are scarce.

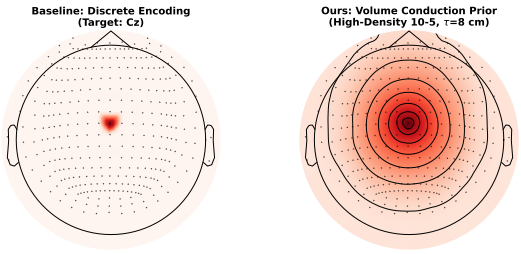


Figure 4: Spatial receptive fields on a 10-5 montage. Baseline (left): discrete impulse. DeeperBrain (right): continuous decay ($\tau \approx 8$ cm) aligned with volume conduction.

3.5 Ablations

Positional encodings. We compare the full DeeperBrain against three variants: (1) *Std. Spatial*—volume-conduction encoding replaced by a linear projection of 3D coordinates; (2) *Std. Temporal*—dynamic temporal basis replaced by sinusoidal embeddings; (3) *Baseline*—both standard. Fig. 2 shows that DeeperBrain outperforms all three variants across the 12 datasets under frozen-probing. The key observation is that merely providing 3D coordinates is insufficient: a linear projection can learn arbitrary spatial relationships from data but lacks the distance-decay prior that mirrors passive current spread through head tissues. By explicitly modeling volume conduction via the decay kernel $\mathbf{K}_{ij}$, DeeperBrain provides the Transformer with a physically grounded spatial smoothing that disentangles neural sources from volume-conducted mixtures—structure that the *Std. Spatial* variant must relearn from limited pretraining data. The temporal comparison reveals a parallel finding: generic sinusoids capture sequence order but are symmetric and directionless, whereas neural adaptation is fundamentally dissipative—an "arrow of time" governed by exponential decay. The oscillatory–decay basis encodes this asymmetry explicitly, aligning the model's temporal representation with biological timescales that sinusoidal embeddings cannot express. The *Baseline* variant, which strips both priors, suffers the largest degradation, confirming that the spatial and temporal encodings provide complementary benefits rooted in distinct aspects of neurophysiology.

Pretraining objective. We ablate the dual-objective by training with MER-only (*w/o NSP*) and NSP-only (*w/o MER*); Fig. 3. The full model achieves the highest performance on the majority of datasets, indicating that local signal fidelity and global dynamical alignment are complementary rather than redundant. The two objectives address different aspects of the representation: MER ensures that the model encodes the *what* of the signal—fine-grained waveform morphology essential for detecting transient events such as epileptic spikes and sleep spindles—while NSP guides the model to understand the *how* and *why* by predicting macroscopic order parameters that characterize brain states. Without MER, the model may overlook subtle temporal features that carry diagnostic value; without NSP, it retains local waveform information but lacks the global neurodynamical structure that makes downstream information accessible without backbone adaptation. Critically, the impact of removing NSP is most severe under frozen-probing, where the model cannot rely on downstream adaptation to recover the missing dynamical structure. This asymmetry directly supports our central claim: neurophysiological alignment during pretraining is what makes task-relevant information accessible in the fixed representation.

3.6 Visualization Analysis

We qualitatively examine whether the gains stem from genuine modeling of neurophysiological structure rather than statistical overfitting.

Volume conduction. Fig. 4 visualizes the effective spatial receptive field of Cz. The baseline yields a discrete impulse (assuming sensor independence), whereas DeeperBrain produces a smooth, isotropic decay aligned with volume conduction, encouraging aggregation over a physically meaningful neighborhood.

Temporal structure. Fig. 5 compares temporal encodings over a 100 s window. Standard sinusoids are symmetric and directionless; learnable embeddings start as unstructured noise; DeeperBrain's

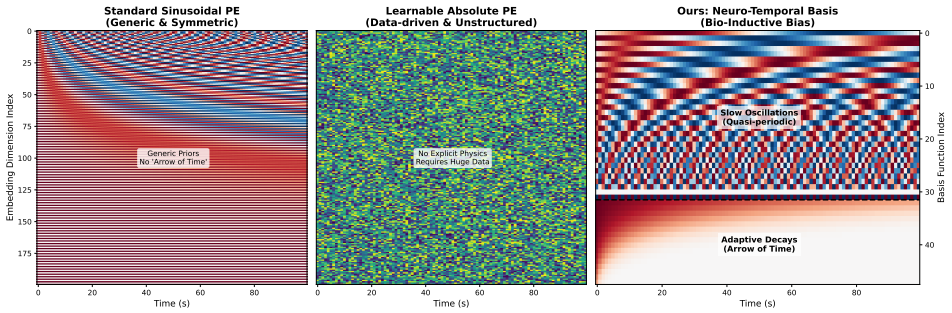


Figure 5: Temporal encodings over 100 s. **Left:** sinusoidal (symmetric). **Middle:** learnable (unstructured). **Right:** DeeperBrain—slow oscillations (top) and adaptive decays (bottom).

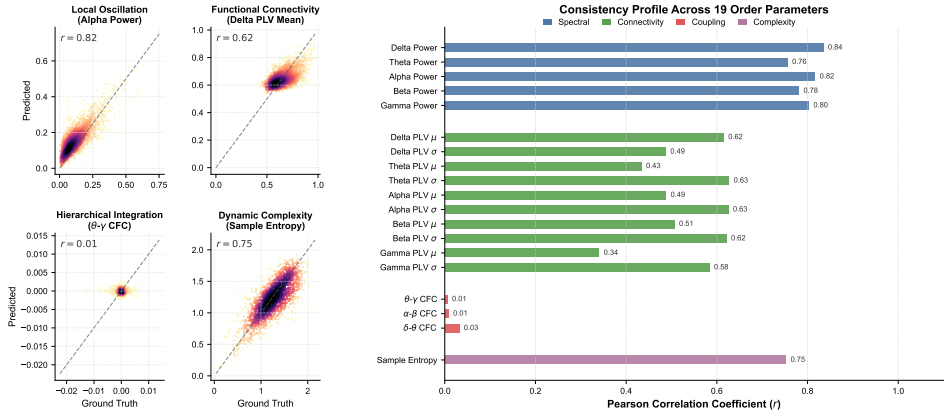


Figure 6: NSP zero-shot consistency on FACED. **Left:** scatter density (Alpha r=0.82, SampEn r=0.75). **Right:** per-metric profile across 19 statistics.

basis disentangles slow oscillations (2–100 s periods) from directional adaptive decays that model the dissipative "arrow of time" of neural adaptation, a structure absent from both alternatives.

Neurodynamical consistency. Fig. 6 reports zero-shot NSP consistency on FACED, comparing ground-truth statistics against predictions from 50%-masked inputs. DeeperBrain shows strong alignment for Alpha power (r=0.82) and sample entropy (r=0.75), moderate for PLV ($r\approx$0.60), and near-zero for cross-frequency coupling, which is a sparse, low-SNR statistic in passive-viewing scalp EEG. The per-metric profile indicates the NSP gain is distributed across several robust statistics rather than driven by a single component.

4 Conclusion

We presented DeeperBrain, a neuro-grounded EEG foundation model that integrates biophysical inductive biases—spatial volume conduction and multi-scale temporal adaptation—and an auxiliary self-supervised objective aligned with neurophysiological order parameters. Pretrained on over 17,000 hours of diverse unlabeled EEG, DeeperBrain achieves state-of-the-art or competitive fine-tuning performance across a broad spectrum of downstream tasks and consistently superior frozen-probing performance, indicating that neurophysiological priors make downstream information more accessible without backbone adaptation. We view these results as a principled step from purely data-driven empiricism toward neuro-grounded inductive biases in EEG foundation models, and as progress along the path toward universal BCI rather than its attainment.

AI USE STATEMENT

In this work, we used generative AI tools for language editing of the manuscript. We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, cleaning or reformatting datasets, or interpreting results, and the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for suggesting the structure of the paper and drafting revisions of portions of the text. We have reviewed all AI-assisted work. All AI-assisted text was checked by the authors for technical accuracy and faithfulness to the reported experimental results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Jayant N Acharya and Vinita J Acharya. Overview of eeg montages and principles of localization. *Journal of Clinical Neurophysiology*, 36(5):325–329, 2019.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ijaz Ahmad, Xin Wang, Mingxing Zhu, Cheng Wang, Yao Pi, Javed Ali Khan, Siyab Khan, Oluwarotimi Williams Samuel, Shixiong Chen, and Guanglin Li. Eeg-based epileptic seizure detection via machine/deep learning approaches: A systematic review. *Computational Intelligence and Neuroscience*, 2022, 2022.

Hamdi Altaheri, Ghulam Muhammad, Mansour Alsulaiman, Syed Umar Amin, Ghadir Ali Altuwaijri, Wadood Abdul, Mohamed A Bencherif, and Mohammed Faisal. Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: A review. *Neural Computing and Applications*, pp. 1–42, 2021.

anonymous. Light hypnotherapy, 2025.

Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2245–2264, 2025.

Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. Uncovering the structure of clinical eeg signals with self-supervised learning. *Journal of Neural Engineering*, 18(4):046020, 2021.

A Björk and E Kugelberg. The electrical activity of the muscles of the eye and eyelids in various positions and during movement. *Electroencephalography and clinical neurophysiology*, 5(4): 595–602, 1953.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Michael Breakspear. Dynamic models of large-scale brain activity. *Nature neuroscience*, 20(3): 340–352, 2017.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

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

György Buzsáki. *Rhythms of the Brain*. Oxford university press, 2006.

György Buzsáki. Neural syntax: cell assemblies, synapsembles, and readers. *Neuron*, 68(3):362–385, 2010.

Hanshu Cai, Zhenqin Yuan, Yiwen Gao, Shuting Sun, Na Li, Fuze Tian, Han Xiao, Jianxiu Li, Zhengwu Yang, Xiaowei Li, et al. A multi-modal open dataset for mental-disorder analysis. *Scientific Data*, 9(1):178, 2022.

Ryan T Canolty, Erik Edwards, Sarang S Dalal, Maryam Soltani, Srikantan S Nagarajan, Heidi E Kirsch, Mitchel S Berger, Nicholas M Barbaro, and Robert T Knight. High gamma power is phase-locked to theta oscillations in human neocortex. *science*, 313(5793):1626–1628, 2006.

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

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

Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A Joshi, and Richard M Leahy. Neuro-gpt: Towards a foundation model for eeg. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2024.

Didar Dadebayev, Wei Wei Goh, and Ee Xion Tan. Eeg-based emotion recognition: Review of commercial eeg devices and machine learning techniques. *Journal of King Saud University-Computer and Information Sciences*, 34(7):4385–4401, 2022.

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

Muhammad Najam Dar, Muhammad Usman Akram, Sajid Gul Khawaja, and Amit N Pujari. Cnn and lstm-based emotion charting using physiological signals. *Sensors*, 20(16):4551, 2020.

Paolo Detti, Giampaolo Vatti, and Garazi Zabalo Manrique de Lara. Eeg synchronization analysis for seizure prediction: A study on data of noninvasive recordings. *Processes*, 8(7):846, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Yi Ding, Neethu Robinson, Su Zhang, 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, 2022.

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*, pp. 1–14, 2023.

Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. LUNA: Efficient and topology-agnostic foundation model for EEG signal analysis. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=uazfjnFL0G.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yang Du, Yongling Xu, Xiaoan Wang, Li Liu, and Pengcheng Ma. Eeg temporal–spatial transformer for person identification. *Scientific Reports*, 12(1):14378, 2022.

Adrienne L Fairhall, Geoffrey D Lewen, William Bialek, and Robert R de Ruyter van Steveninck. Efficiency and ambiguity in an adaptive neural code. *Nature*, 412(6849):787–792, 2001.

Zitao Fang, Chenxuan Li, Hongting Zhou, Shuyang Yu, Guodong Du, Ashwaq Qasem, Yang Lu, Jing Li, Junsong Zhang, and Sim Kuan Goh. Neuript: Foundation model for neural interfaces. *arXiv preprint arXiv:2510.16548*, 2025.

Navid Mohammadi Foumani, Geoffrey Mackellar, Soheila Ghane, Saad Irtza, Nam Nguyen, and Mahsa Salehi. Eeg2rep: Enhancing self-supervised eeg representation through informative masked inputs. *arXiv preprint arXiv:2402.17772*, 2024.

Pengxuan Gao, Tianyu Liu, Jia-Wen Liu, Bao-Liang Lu, and Wei-Long Zheng. Multimodal multi-view spectral-spatial-temporal masked autoencoder for self-supervised emotion recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1926–1930. IEEE, 2024.

Sanaz Ghaffari, Arian Yavari, Sara Bonyadian, Arsalan Ghofrani, and Russell Butler. Visual eeg study (brainvision $\rightarrow$ bids), 2025.

Mohammad M Ghassemi, Benjamin E Moody, Li-Wei H Lehman, Christopher Song, Qiao Li, Haoqi Sun, Roger G Mark, M Brandon Westover, and Gari D Clifford. You snooze, you win: the physionet/computing in cardiology challenge 2018. In *2018 Computing in Cardiology Conference (CinC)*, volume 45, pp. 1–4. IEEE, 2018.

Ary L. Goldberger, Luis A. Nunes Amaral, L Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and Harry Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101 23:E215–20, 2000a.

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

Roberta Grech, Tracey Cassar, Joseph Muscat, Kenneth P Camilleri, Simon G Fabri, Michalis Zervakis, Petros Xanthopoulos, Vangelis Sakkalis, and Bart Vanrumste. Review on solving the inverse problem in eeg source analysis. *Journal of neuroengineering and rehabilitation*, 5(1):25, 2008.

Xiao Gu, Yuxuan Shu, Jinpei Han, Yuxuan Liu, Zhangdaihong Liu, James Anibal, Veer Sangha, Edward Phillips, Bradley Segal, Hang Yuan, et al. Foundation models for biosignals: A survey. *Authorea Preprints*, 2025.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.

Zainab Jamil, Afshan Jamil, and Muhammad Majid. Artifact removal from eeg signals recorded in non-restricted environment. *Biocybernetics and Biomedical Engineering*, 41(2):503–515, 2021.

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.

Ziyu Jia, Youfang Lin, Jing Wang, Ronghao Zhou, Xiaojun Ning, Yuanlai He, and Yaoshuai Zhao. Graphsleepnet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification. In *IJCAI*, volume 2021, pp. 1324–1330, 2020.

Ziyu Jia, Youfang Lin, Jing Wang, Xiaojun Ning, Yuanlai He, Ronghao Zhou, Yuhan Zhou, and H Lehman Li-wei. Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1977–1986, 2021.

Muyun Jiang, Shuailei Zhang, Zhenjie Yang, Mengjun Wu, Weibang Jiang, Zhiwei Guo, Wei Zhang, Rui Liu, Shangen Zhang, Yong Li, et al. Elastiq: Eeg-language alignment with semantic task instruction and querying. *arXiv preprint arXiv:2509.24302*, 2025a.

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*, 2024a.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. NeuroLM: A universal multi-task foundation model for bridging the gap between language and EEG signals. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=Io9yFt7XH7.

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*, 2024b. URL https://openreview.net/forum?id=QzTpTRVtrP.

Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. Isruc-sleep: A comprehensive public dataset for sleep researchers. *Computer methods and programs in biomedicine*, 124:180–192, 2016.

Andy Jeesu Kim, Mara Mather, Santiago Morales, and Joshua Senior. Young adult resting state and auditory oddball task eeg data, 2025a.

Andy Jeesu Kim, Santiago Morales, Joshua Senior, and Mara Mather. Heartbeam: Older adult resting state and auditory oddball task eeg data, 2025b.

Isaac Kinley and Sue Becker. Cueing the future to reduce temporal discounting, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

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

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

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, 2018.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Na Lee, Konstantinos Barmpas, Yannis Panagakis, Dimitrios Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. Are large brainwave foundation models capable yet? insights from fine-tuning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=J5SbLoq7Uv.

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.

Hongqi Li, Yitong Chen, Yujuan Wang, Weihang Ni, and Haodong Zhang. Foundation models for cross-domain eeg analysis application: A survey. *arXiv preprint arXiv:2508.15716*, 2025.

Jiyao Liu, Li Zhang, Hao Wu, and Huan Zhao. Transformers for eeg emotion recognition. *arXiv preprint arXiv:2110.06553*, 2021a.

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*, 14(2):715–729, 2021b.

Fabien Lotte, Marco Congedo, Anatole Lécuyer, Fabrice Lamarche, and Bruno Arnaldi. A review of classification algorithms for eeg-based brain–computer interfaces. *Journal of neural engineering*, 4(2):R1, 2007.

Weiheng Lu, Chunfeng Song, Jiamin Wu, Pengyu Zhu, Yuchen Zhou, Weijian Mai, Qihao Zheng, and Wanli Ouyang. Unimind: Unleashing the power of llms for unified multi-task brain decoding. *arXiv preprint arXiv:2506.18962*, 2025.

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. *arXiv preprint arXiv:2506.09110*, 2025.

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, 2022.

Malik Muhammad Naeem Mannan, Muhammad Ahmad Kamran, and Myung Yung Jeong. Identification and removal of physiological artifacts from electroencephalogram signals: A review. *Ieee Access*, 6:30630–30652, 2018.

Dennis J McFarland, Charles W Anderson, K-R Muller, Alois Schlogl, and Dean J Krusienski. Bci meeting 2005-workshop on bci signal processing: feature extraction and translation. *IEEE transactions on neural systems and rehabilitation engineering*, 14(2):135–138, 2006.

María Melcón, Enrique Stern, Lydia Arana, and Almudena Capilla. Eeg data during three near-threshold visual detection tasks: a no-cue task, a noninformative cue task (50% validity), and an informative cue task (100% validity), 2025.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40, 2023.

Jianliang Min, Ping Wang, and Jianfeng Hu. Driver fatigue detection through multiple entropy fusion analysis in an eeg-based system. *PLoS one*, 12(12):e0188756, 2017.

Computational Neuroimaging and Neuroengineering Lab ar the University of Oklahoma. Resting eeg, 2025.

Paul L Nunez and Ramesh Srinivasan. *Electric fields of the brain: the neurophysics of EEG*. Oxford university press, 2006.

Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in neuroscience*, 10:196, 2016.

OpenNeuro. Openneuro: A free and open platform for validating and sharing bids-compliant mri, pet, meg, eeg, and ieeg data, 2025. URL `https://openneuro.org`.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. Reve: A foundation model for eeg–adapting to any setup with large-scale pretraining on 25,000 subjects. *arXiv preprint arXiv:2510.21585*, 2025.

Wei Yan Peh, Yuanyuan Yao, and Justin Dauwels. Transformer convolutional neural networks for automated artifact detection in scalp eeg. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3599–3602. IEEE, 2022.

Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. Seqsleepnet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(3):400–410, 2019.

Huy Phan, Oliver Y Chén, Minh C Tran, Philipp Koch, Alfred Mertins, and Maarten De Vos. Xsleepnet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5903–5915, 2021.

Huy Phan, Kaare Mikkelsen, Oliver Y Chén, Philipp Koch, Alfred Mertins, and Maarten De Vos. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering*, 69(8):2456–2467, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

Joshua S Richman and J Randall Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American journal of physiology-heart and circulatory physiology*, 278(6):H2039–H2049, 2000.

Simanto Saha and Mathias Baumert. Intra-and inter-subject variability in eeg-based sensorimotor brain computer interface: a review. *Frontiers in computational neuroscience*, 13:87, 2020.

Siavash Sakhavi, Cuntai Guan, and Shuicheng Yan. Learning temporal information for brain-computer interface using convolutional neural networks. *IEEE transactions on neural networks and learning systems*, 29(11):5619–5629, 2018.

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.

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.

Ali Hossam Shoeb. *Application of machine learning to epileptic seizure onset detection and treatment*. PhD thesis, Massachusetts Institute of Technology, 2009.

Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie. Transformer-based spatial-temporal feature learning for eeg decoding. *arXiv preprint arXiv:2106.11170*, 2021.

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

Steven H Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. Chapman and Hall/CRC, 2024.

Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, 2017.

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

Jose Antonio Urigüen and Begoña Garcia-Zapirain. Eeg artifact removal—state-of-the-art and guidelines. *Journal of neural engineering*, 12(3):031001, 2015.

Francisco Varela, Jean-Philippe Lachaux, Eugenio Rodriguez, and Jacques Martinerie. The brainweb: phase synchronization and large-scale integration. *Nature reviews neuroscience*, 2(4):229–239, 2001.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Brainbert: Self-supervised representation learning for intracranial recordings. In *The Eleventh International Conference on Learning Representations*, 2022.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL `https://openreview.net/forum?id=lvS2b8CjG5`.

Jialing Wang, Shiwei Cheng, Jieming Tian, and Yuefan Gao. A 2d cnn-lstm hybrid algorithm using time series segments of eeg data for motor imagery classification. *Biomedical Signal Processing and Control*, 83:104627, 2023a.

Jiquan Wang, Sha Zhao, Yangxuan Zhou, Haiteng Jiang, Zhenghe Yu, Tao Li, Shijian Li, and Gang Pan. Narcolepsy diagnosis with sleep stage features using psg recordings. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:3619–3629, 2023b.

Jiquan Wang, Sha Zhao, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Generalizable sleep staging via multi-level domain alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 265–273, 2024b.

Jiquan Wang, Sha Zhao, Haiteng Jiang, Yangxuan Zhou, Zhenghe Yu, Tao Li, Shijian Li, and Gang Pan. Caresleepnet: A hybrid deep learning network for automatic sleep staging. *IEEE Journal of Biomedical and Health Informatics*, 2024c.

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*, 2025a. URL https://openreview.net/forum?id=NPNUHgHF2w.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Shijian Li, and Gang Pan. Eegmamba: An eeg foundation model with mamba. *Neural Networks*, pp. 107816, 2025b.

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.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1810–1822, 2019.

Wenlong Wang, Feifei Qi, David Paul Wipf, Chang Cai, Tianyou Yu, Yuanqing Li, Yu Zhang, Zhuliang Yu, and Wei Wu. Sparse bayesian learning for end-to-end eeg decoding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15632–15649, 2023c.

Yilin Wang, Sha Zhao, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. M-mdd: A multi-task deep learning framework for major depressive disorder diagnosis using eeg. *Neurocomputing*, 636: 130008, 2025c.

Jiamin Wu, Zichen Ren, Junyu Wang, Pengyu Zhu, Yonghao Song, Mianxin Liu, Qihao Zheng, Lei Bai, Wanli Ouyang, and Chunfeng Song. Adabrain-bench: Benchmarking brain foundation models for brain-computer interface applications. *arXiv preprint arXiv:2507.09882*, 2025.

Wei Wu, Zhe Chen, Xiaorong Gao, Yuanqing Li, Emery N Brown, and Shangkai Gao. Probabilistic common spatial patterns for multichannel eeg analysis. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):639–653, 2014.

Zhaohui Wu, Gang Pan, and Nenggan Zheng. Cyborg intelligence. *IEEE Intelligent Systems*, 28(5): 31–33, 2013.

X. Memory reactivation levels remain unaffected by anticipated interference experiment 2 dataset, 2025.

Qinfan Xiao, Ziyun Cui, Chi Zhang, Siqi Chen, Wen Wu, Andrew Thwaites, Alexandra Woolgar, Bowen Zhou, and Chao Zhang. Brainomni: A brain foundation model for unified eeg and meg signals. *arXiv preprint arXiv:2505.18185*, 2025.

Wei Xiong, Jiangtong Li, Jie Li, and Kun Zhu. Eeg-fm-bench: A comprehensive benchmark for the systematic evaluation of eeg foundation models. *arXiv preprint arXiv:2508.17742*, 2025.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47 (3):1877–1893, 2025.

xx. Memory reactivation levels remain unaffected by anticipated interference, 2025.

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

Shihao Yang, Xiying Huang, Danilo Bernardo, Jun-En Ding, Andrew Michael, Jingmei Yang, Patrick Kwan, Ashish Raj, and Feng Liu. Foundation and large-scale ai models in neuroscience: A comprehensive review. *arXiv preprint arXiv:2510.16658*, 2025.

Sana Yasin, Syed Asad Hussain, Sinem Aslan, Imran Raza, Muhammad Muzammel, and Alice Othmani. Eeg based major depressive disorder and bipolar disorder detection using neural networks: A review. *Computer Methods and Programs in Biomedicine*, 202:106007, 2021.

İlkay Yıldız, Rachael Garner, Matthew Lai, and Dominique Duncan. Unsupervised seizure identification on eeg. *Computer methods and programs in biomedicine*, 215:106604, 2022.

Zhizhang Yuan, Daoze Zhang, Junru Chen, Geifei Gu, and Yang Yang. Brant-2: Foundation model for brain signals. *arXiv preprint arXiv:2402.10251*, 2024.

Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. In *Advances in Neural Information Processing Systems*, 2023.

Daoze Zhang, Zhizhang Yuan, Junru Chen, Kerui Chen, and Yang Yang. Brant-x: A unified physiological signal alignment framework. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4155–4166, 2024.

Ruilong Zhang, Qun Zong, Liqian Dou, and Xinyi Zhao. A novel hybrid deep learning scheme for four-class motor imagery classification. *Journal of neural engineering*, 16(6):066004, 2019a.

Shaomin Zhang, Sheng Yuan, Lipeng Huang, Xiaoxiang Zheng, Zhaohui Wu, Kedi Xu, and Gang Pan. Human mind control of rat cyborg's continuous locomotion with wireless brain-to-brain interface. *Scientific reports*, 9(1):1321, 2019b.

Zihan Zhang, Yu Bao, Tianyi Jiang, Xiao Ding, Xia Liang, Juntong Du, Yi Zhao, Kai Xiong, Bing Qin, and Ting Liu. Chisco-2.0, 2025.

Sha Zhao, Song Yi, Yangxuan Zhou, Jiadong Pan, Jiquan Wang, Jie Xia, Shijian Li, Shurong Dong, and Gang Pan. Wearable music2emotion: Assessing emotions induced by ai-generated music through portable eeg-fnirs fusion. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5627–5636, 2025.

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. *arXiv preprint arXiv:2503.00580*, 2025a.

Yangxuan Zhou, Sha Zhao, Jiquan Wang, Haiteng Jiang, Benyan Luo, Tao Li, Gang Pan, et al. Personalized sleep staging leveraging source-free unsupervised domain adaptation. *arXiv preprint arXiv:2412.12159*, 2024a.

Yangxuan Zhou, Sha Zhao, Jiquan Wang, Haiteng Jiang, Zhenghe Yu, Shijian Li, Tao Li, and Gang Pan. Simplifying multimodal with single eog modality for automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:1668–1678, 2024b.

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

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

A METHOD DETAILS

This appendix provides the detailed derivations and parameter specifications referenced in Section 2.

A.1 PATCHING AND MASKING: FORMAL NOTATION

Let the raw EEG be $\widetilde{\mathbf{X}} \in \mathbb{R}^{\widetilde{C}\times T}$. For each training instance we randomly permute the channel order and select a subset of size $C \in [1, \widetilde{C}]$ to simulate variable-density recordings:

$$\mathbf{X} = \texttt{channel_reorganization}(\widetilde{\mathbf{X}}) \in \mathbb{R}^{C\times T}. \tag{8}$$

The reorganized signal is partitioned into non-overlapping one-second segments. With sampling frequency f_s, each segment has $P = f_s$ points, yielding $\mathbf{X}_{\text{patch}} \in \mathbb{R}^{C\times S\times P}$ where $S = T/P$. Each vector $\mathbf{X}_{c,s,:} \in \mathbb{R}^P$ is a (channel $\times$ second) patch. Masked patches are replaced with a zero vector $\mathbf{0} \in \mathbb{R}^P$ in the raw voltage space (not a learnable token), preserving compatibility with the temporal-convolutional embedding. A binary mask $\mathbf{M} \in \{0,1\}^{C\times S}$ denotes masked positions.

A.2 PATCH EMBEDDING: ARCHITECTURE DETAILS

The masked input $\mathbf{X}_{\text{mask}} \in \mathbb{R}^{C\times S\times P}$ is reshaped to $\mathbb{R}^{1\times(CS)\times P}$ and processed by three convolutional blocks:

$$\mathbf{E}_{\text{time}} = \text{ConvBlock}_3 \circ \text{ConvBlock}_2 \circ \text{ConvBlock}_1(\mathbf{X}_{\text{mask}}). \tag{9}$$

Each block uses a 2D convolution with kernel $(1, K_i)$ and stride $(1, s_i)$, followed by Group Normalization (5 groups) and GELU. The first block (K_1=49, s_1=25) captures long-range structure and reduces length, mimicking delta–theta bandpass filtering; subsequent blocks (K_2=K_3=3, unit stride) refine local structures. The output $\mathbf{E}_{\text{time}} \in \mathbb{R}^{F\times(CS)\times L}$ with F=25 is flattened to patch embeddings $\mathbf{E} \in \mathbb{R}^{C\times S\times D}$ with $D = F \times L$.

A.3 CHANNEL ENCODING: BIOPHYSICAL DERIVATION

Scalp potentials arise from passive current spread through conductive tissues (Nunez & Srinivasan, 2006). Under the quasi-static approximation, the potential at scalp location $\mathbf{r}$ due to a dipole $\mathbf{q}$ at $\mathbf{r}_s$ is

$$\phi(\mathbf{r}) = \frac{1}{4\pi\sigma}\frac{\mathbf{q}\cdot(\mathbf{r}-\mathbf{r}_s)}{\|\mathbf{r}-\mathbf{r}_s\|^3}, \tag{10}$$

so signal similarity decays with inter-electrode distance. From a biophysical perspective, EEG can be modeled as a linear mixture $\mathbf{x} = \mathbf{A}\mathbf{s} + \boldsymbol{\epsilon}$, where $\mathbf{A} \in \mathbb{R}^{C\times Q}$ is the lead field matrix and $\mathbf{s}$ denotes source activity. Under isotropic and uncorrelated sources ($\mathbb{E}[\mathbf{s}\mathbf{s}^\top] = \mathbf{I}$), the covariance of observed signals satisfies $\Sigma_{ij} \propto \exp(-\|\mathbf{p}_i - \mathbf{p}_j\|/\tau)$ (Nunez & Srinivasan, 2006); our kernel $\mathbf{K}_{ij}$ approximates this expected spatial correlation structure. The decay scale α is initialized so that τ corresponds to a typical inter-electrode distance of 8 cm, with numerical-stability constant $\delta = 10^{-5}$.

A.4 TEMPORAL ENCODING: BIOLOGICAL MOTIVATION

Higher-order brain states involve slow modulations (seconds to minutes), including alpha fluctuations and adaptation (Buzsáki, 2006; Fairhall et al., 2001). Generic sinusoidal or learnable encodings treat time as symmetric and lack an inductive bias for the dissipative "arrow of time" of neural adaptation. The oscillatory basis captures slow cognitive and experimental-state fluctuations such as block-wise attentional engagement or sleep stage transitions that persist even after high-pass filtering of raw EEG. The decay basis spans adaptation constants covering sensory adaptation (seconds) to cognitive maintenance (tens of seconds), grounded in the observation that neural responses to sustained input often follow $r(t) \propto e^{-t/\tau}$. The concatenated vector $\boldsymbol{\psi}(t) \in \mathbb{R}^{2K+M}$ (with K=M=8) is linearly projected to D dimensions and broadcast across channels, providing a principled, non-adaptive temporal prior that complements the fast features extracted by the convolutional backbone.

A.5 NSP: Detailed Statistic Definitions

Formally, with latent state $\mathbf{z}(t)$ evolving as $d\mathbf{z}/dt = \mathbf{F}(\mathbf{z}, \mathbf{u}, t)$ and observed as $\mathbf{x}(t) = \mathcal{H}(\mathbf{z}(t)) + \boldsymbol{\epsilon}(t)$, we approximate the order parameters $\mathbf{y}(t) = \Phi(\mathbf{z}(t))$ directly from EEG via $\Psi(\mathbf{x}(t)) \approx \mathbf{y}(t)$. The four statistics computed by $\Psi(\cdot)$ are:

(1) Relative spectral power (5D). Spectral power distribution reflects the attractor landscape, with specific frequency bands linked to cognitive states (Buzsáki, 2006). We compute the PSD via real FFT, integrate over five canonical bands (delta 1–4 Hz, theta 4–8 Hz, alpha 8–13 Hz, beta 13–30 Hz, gamma 30–75 Hz), and normalize to a probability simplex: $p_k = \frac{\int_{f \in \text{band}_k} \text{PSD}(f)\, df}{\sum_{k'} \int_{f \in \text{band}_{k'}} \text{PSD}(f)\, df}$.

(2) Phase-locking value (10D). Phase synchronization measures large-scale functional integration (Varela et al., 2001). For each band, we extract analytic phases $\phi_c(t)$ via the Hilbert transform and compute the PLV matrix $\text{PLV}_{ij} = |\frac{1}{P} \sum_t e^{j(\phi_i(t) - \phi_j(t))}|$, summarized by the mean and standard deviation of its upper-triangular elements.

(3) Cross-frequency coupling (3D). CFC encodes coordination across temporal scales, such as theta–gamma interactions (Canolty et al., 2006). For canonical pairs (theta–gamma, alpha–gamma, beta–gamma), we compute $\text{CFC} = \frac{1}{P} \sum_t \cos(\phi_{\text{low}}(t))\, A_{\text{high}}(t)$, where ϕ_{low} and A_{high} are the low-frequency phase and high-frequency amplitude.

(4) Sample entropy (1D). Signal complexity reflects the richness of neural dynamics (Richman & Moorman, 2000). We compute $\text{SampEn}(m, r) = -\log(C^{m+1}(r)/C^m(r))$ with m=2, r=0.2 std(x), yielding a measure of signal unpredictability.

B Detailed Preprocessing Pipelines

TUEG (clean subset). We standardize each recording to the 19-channel 10–20 montage (FP1, FP2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, T6, FZ, CZ, PZ) across the three TUEG subcorpora (`01_tcp_ar`, `02_tcp_le`, `03_tcp_ar_a`); recordings missing any required channel are discarded. Recordings are resampled to 200 Hz, band-pass filtered at 0.3–75 Hz, and notch-filtered at 60 Hz. After trimming the first and last 60 s of each recording to remove edge artifacts, signals are segmented into non-overlapping 30 s windows; recordings shorter than 60 s after trimming are discarded. This yields 1,109,545 samples totaling 9,246.2 hours.

PhysioNet 2018. Sleep EEG from 994 subjects (training set only), used at its native 200 Hz with the same band-pass and notch filtering, segmented into 30 s windows.

Siena Scalp EEG. Originally 31 channels; we use the standard 10–20 subset of 27 channels, resampled to 200 Hz, filtered, and segmented into 10 s windows.

OpenNeuro datasets. We incorporate ten datasets from the OpenNeuro platform (OpenNeuro, 2025). These were selected to maximize paradigm and montage diversity within the OpenNeuro repository: they span perceptual decision-making, imagined speech, memory, emotional processing, vigilance, and resting-state paradigms—domains underrepresented in the clinical TUEG and sleep-focused PhysioNet corpora—and provide high-density montages (28–128 channels) that complement the 19- and 6-channel recordings dominating the pretraining pool. All datasets are publicly available under open licenses. Each is resampled to 200 Hz, band-pass filtered (0.3–75 Hz), notch-filtered, and segmented into 5–8 s windows. Original sampling rates range from 250 to 2048 Hz. The ten datasets are: ds006171 (Melcón et al., 2025) (near-threshold visual detection with cueing manipulations), ds006317 (Zhang et al., 2025) (Chisco-2.0 imagined-speech EEG), ds006367 (xx, 2025) (memory reactivation under anticipated interference, Exp. 1), ds006370 (X, 2025) (memory reactivation under anticipated interference, Exp. 2), ds006437 (anonymous, 2025) (LIGHT hypnotherapy EEG), ds006446 (Kinley & Becker, 2025) (future-cueing and temporal discounting), ds006466 (Kim et al., 2025b) (HeartBEAM older-adult resting-state and auditory oddball), ds006480 (Kim et al., 2025a) (HeartBEAM young-adult resting-state and auditory oddball), ds006525 (Neuroimaging & ar the University of Oklahoma, 2025) (resting-state EEG), and ds006547 (Ghaffari et al., 2025) (visual EEG study, BrainVision → BIDS).

Raw EEG Data. 64-channel cognitive-task recordings, resampled to 200 Hz, filtered, segmented into 5 s windows.

All recordings are normalized to a 100 μV scale (divided by 100) so that values predominantly lie in $[-1, 1]$, following (Jiang et al., 2024b).

C Downstream Dataset Details

Table 5 extends Table 2 with sampling rate, per-sample duration, label structure, and the subject-level train/val/test split. Splits follow LaBraM (Jiang et al., 2024b) or CBraMod (Wang et al., 2025a) where available; otherwise a subject-level split is used to ensure no subject leakage across partitions. For SEED-V and SEED-VII, trials are split per session (5:5:5 and 10:5:5 respectively), following the standard protocol of (Liu et al., 2021b; Jiang et al., 2024a). For BCIC2020-3, the predefined training/validation/test partition provided by the challenge organizers is used. For MODMA, the split balances major-depression (MDD) and normal-control (NC) subjects across partitions to preserve class balance.

Table 5: Downstream dataset details, including subject-level train/val/test splits.

Task	Dataset	Hz	Ch	Dur	# Smp	Labels	Train	Val	Test
Emotion	FACED (Chen et al., 2023)	250	32	10s	10,332	9-class	Sub 1–80	Sub 81–100	Sub 101–123
	SEED-V (Liu et al., 2021b)	1000	62	1s	117,744	5-class	Trials/session 5:5:5		
	SEED-VII (Jiang et al., 2024a)	1000	62	5s	55,561	7-class	Trials/session 10:5:5		
Motor	PhysioNet-MI (Goldberger et al., 2000b; Schalk et al., 2004)	160	64	4s	9,837	4-class	Sub 1–70	Sub 71–89	Sub 90–109
	BCIC-IV-2a (Brunner et al., 2008)	250	22	4s	5,184	4-class	Sub 1–5	Sub 6–7	Sub 8–9
	SHU-MI (Ma et al., 2022)	250	32	4s	11,988	2-class	Sub 1–15	Sub 16–20	Sub 21–25
Sleep	ISRUC (Khalighi et al., 2016)	200	6	30s	89,240	5-class	Sub 1–80	Sub 81–90	Sub 91–100
Seizure	CHB-MIT (Goldberger et al., 2000b; Shoeb, 2009)	256	16	10s	326,993	2-class	Sub 1–19	Sub 20–21	Sub 22–23
Speech	BCIC2020-3 (Jeong et al., 2022)	256	64	3s	6,000	5-class	Predefined partition		
Disorder	MODMA (Cai et al., 2022)	250	128	15s	1,066	2-class	15MDD+18NC	4MDD+5NC	5MDD+6NC
Vigilance	SEED-VIG (Min et al., 2017)	200	17	8s	20,355	regression	Sub 1–15	Sub 16–19	Sub 20–23
Workload	MentalArith (Goldberger et al., 2000b; Zyma et al., 2019)	500	20	5s	1,707	2-class	Sub 1–28	Sub 29–32	Sub 33–36

All downstream EEG signals are resampled to 200 Hz with a 1 s patch duration (200 samples), consistent with the pretraining configuration. For evaluation, we report the primary monitor metric per task: Balanced Accuracy for classification and Pearson correlation for SEED-VIG regression. Task-specific metrics (Cohen's Kappa, Weighted F1 for multi-class; AUC-PR, AUROC for binary; R2, RMSE for regression) are reported in Appendix E. All results are mean$\pm$std over five independent random seeds.

Evaluation protocol. For *end-to-end fine-tuning*, the entire model (backbone + head) is optimized. For *frozen-probing*, the pretrained backbone is fixed and only the task-specific head is trained. Both protocols share the same optimizer (AdamW, weight decay 5×10^{-2}), learning-rate schedule (cosine annealing from $\{1, 3, 5\}\times10^{-4}$, selected on validation), batch size (32 or 64 depending on dataset scale), and epoch budget (50). The same protocol is applied to DeeperBrain and to every baseline for fair comparison.

D Baseline Configurations

Table 6 reports, for each foundation-model baseline, the pretraining corpus size, parameter count, embedding dimension, and source. Because these models were published with different pretraining data, parameter counts, and tokenization schemes, the comparison in Tables 3–4 is a practical horizontal benchmark rather than a strictly controlled ablation; DeeperBrain's advantage holds across baselines of varying scale.

Table 6: Foundation-model baseline configurations (as reported in their original papers).

Model	Pretrain hrs	# Params	Embed dim	Source
LaBraM (Jiang et al., 2024b)	∼2,534	∼5.8M	200	public
CBraMod (Wang et al., 2025a)	∼9,000	∼4.0M	200	public
CSBrain (Zhou et al., 2025b)	∼9,000	∼4.9M	200	public
REVE (Ouahidi et al., 2025)	61,415	∼12M	200	public
DeeperBrain (ours)	17,243	∼5.8M	200	–

E Full Multi-Metric Results

Tables 7 and 8 report the full three-metric results for end-to-end fine-tuning and frozen-probing, respectively (mean±std, %). **Bold** denotes the best and underline the second best. For SEED-VIG, correlation values are multiplied by 100 for readability; the standard range is $r \in [-1, 1]$.

Table 7: End-to-end fine-tuning, full metrics (mean±std, %).

	FACED, 9-class			SEED-V, 5-class			SEED-VII, 7-class		
Method	Bal.Acc	Kappa	F1	Bal.Acc	Kappa	F1	Bal.Acc	Kappa	F1
EEGNet	40.90±1.22	33.42±2.51	41.24±1.41	29.61±1.02	10.06±1.43	27.49±0.98	25.28±0.33	12.85±0.44	25.36±0.36
EEGConformer	45.59±1.25	38.58±1.86	45.14±1.07	35.37±1.12	17.72±1.74	34.87±1.36	28.75±0.28	17.47±0.30	28.02±0.28
LaBraM	52.73±1.07	46.98±1.88	52.88±1.02	39.76±1.38	23.86±2.09	39.74±1.11	33.28±0.33	22.85±0.44	33.36±0.36
CBraMod	55.09±0.89	50.41±1.22	56.18±0.93	40.91±0.97	25.69±1.43	41.01±1.08	32.47±0.36	21.47±0.38	32.77±0.32
CSBrain	57.52±0.42	52.04±0.36	57.96±0.31	**41.97**±0.33	**27.85**±0.34	**42.80**±0.23	32.44±0.38	21.45±0.49	32.74±0.47
REVE	56.46±1.64	50.80±1.91	56.59±1.72	40.48±0.23	26.22±0.21	41.45±0.17	32.64±0.34	21.72±0.37	33.02±0.31
DeeperBrain	**60.32**±0.34	**54.99**±0.36	**60.58**±0.33	40.68±0.31	26.43±0.36	41.54±0.25	**33.60**±0.25	**22.86**±0.28	**33.95**±0.23

	PhysioNet-MI, 4-class			BCIC-IV-2a, 4-class			SHU-MI, 2-class		
Method	Bal.Acc	Kappa	F1	Bal.Acc	Kappa	F1	Bal.Acc	AUC-PR	AUROC
EEGNet	58.14±1.25	44.68±1.99	57.96±1.15	44.82±0.94	26.93±1.21	42.26±1.08	58.89±1.77	63.11±1.42	62.83±1.52
EEGConformer	60.49±1.04	47.36±1.71	60.62±0.95	46.96±1.06	29.24±1.41	45.33±1.28	59.00±1.07	63.70±0.93	63.51±1.01
LaBraM	61.73±1.22	49.12±1.92	61.77±1.41	48.69±0.85	31.59±1.54	47.58±1.03	61.66±1.92	67.61±0.83	66.04±0.91
CBraMod	64.17±0.91	52.22±1.69	64.27±1.00	51.38±0.66	35.18±0.94	49.84±0.85	63.70±1.51	71.39±0.88	69.88±0.68
CSBrain	63.04±0.90	50.71±1.20	63.08±0.95	56.57±0.71	42.09±0.93	56.37±0.87	64.17±0.37	**72.30**±0.79	**72.00**±0.58
REVE	64.80±1.40	53.06±1.87	64.84±1.70	**63.96**±0.95	**51.94**±1.26	**63.39**±1.10	63.84±0.34	69.90±0.72	70.63±0.42
DeeperBrain	**65.17**±0.37	**53.56**±0.49	**65.38**±0.34	60.69±1.40	47.59±1.86	60.64±1.50	**64.31**±0.47	70.99±0.53	70.95±0.43

	ISRUC, 5-class			CHB-MIT, 2-class			BCIC2020-3, 5-class		
Method	Bal.Acc	Kappa	F1	Bal.Acc	AUC-PR	AUROC	Bal.Acc	Kappa	F1
EEGNet	71.54±1.21	70.40±1.73	75.13±1.24	56.58±1.06	19.14±1.82	80.48±1.36	44.13±0.96	30.16±1.23	44.13±1.02
EEGConformer	74.00±1.33	71.43±1.62	76.34±1.51	59.76±1.41	22.09±2.15	82.26±1.70	45.06±1.33	31.33±1.83	44.88±1.54
LaBraM	76.33±1.02	72.31±1.82	78.10±1.33	70.75±3.58	32.87±4.02	86.79±1.99	50.60±1.55	38.00±2.42	50.54±2.05
CBraMod	78.65±1.10	74.42±1.52	80.11±0.99	73.98±2.84	36.89±3.82	88.92±1.54	53.73±1.08	42.16±1.63	53.83±0.96
CSBrain	**79.25**±0.30	74.06±1.02	79.90±0.91	72.62±1.15	51.64±4.49	89.15±3.21	56.05±4.38	45.07±5.47	56.05±4.41
REVE	78.19±0.78	75.00±1.56	80.11±0.99	72.02±4.95	53.97±6.38	90.41±0.97	56.35±1.23	45.43±1.54	56.33±1.24
DeeperBrain	78.75±0.60	**75.01**±0.32	**80.89**±0.40	**74.34**±1.49	**65.51**±4.27	**94.09**±0.77	**57.06**±0.78	**46.33**±0.97	**57.04**±0.78

	MODMA, 2-class			SEED-VIG, regression			MentalArith, 2-class		
Method	Bal.Acc	AUC-PR	AUROC	Corr.*	R2	RMSE↓	Bal.Acc	AUC-PR	AUROC
EEGNet	61.90±2.40	64.50±3.00	62.30±4.70	57.01±1.67	23.66±0.84	28.28±0.74	67.70±1.16	57.63±1.02	73.21±1.08
EEGConformer	63.85±2.20	66.40±3.25	64.70±4.60	57.50±1.39	23.44±0.91	28.50±0.83	68.05±1.23	58.29±1.34	74.24±1.28
LaBraM	66.81±7.29	73.43±5.82	68.83±6.22	59.31±0.98	24.32±0.85	27.62±0.48	72.24±2.05	70.78±3.77	84.11±2.46
CBraMod	70.18±2.34	75.92±3.05	72.37±4.91	58.49±2.78	24.82±3.67	27.35±0.83	74.39±2.15	71.68±2.93	83.90±3.27
CSBrain	72.85±2.60	77.90±3.60	75.10±5.00	**63.14**±3.56	23.63±5.19	27.74±0.94	75.58±1.06	66.96±2.21	84.78±2.97
REVE	72.45±2.10	77.83±3.21	74.62±4.75	58.77±1.17	25.44±1.90	27.42±0.35	**76.60**±3.55	**74.70**±8.07	84.50±5.14
DeeperBrain	**76.31**±2.59	**81.55**±3.67	**78.85**±5.19	60.11±0.72	**34.14**±1.62	**25.77**±0.32	75.69±2.32	72.04±4.52	**86.21**±2.19

*SEED-VIG correlation values multiplied by 100 for readability (standard range $r \in [-1, 1]$).

F Additional Ablation Figures

Figures 7 and 8 show the positional-encoding and objective ablations under end-to-end fine-tuning (the frozen-probing counterparts appear in the main text). Significance markers: "*" $p < 0.05$, "**" $p < 0.01$.

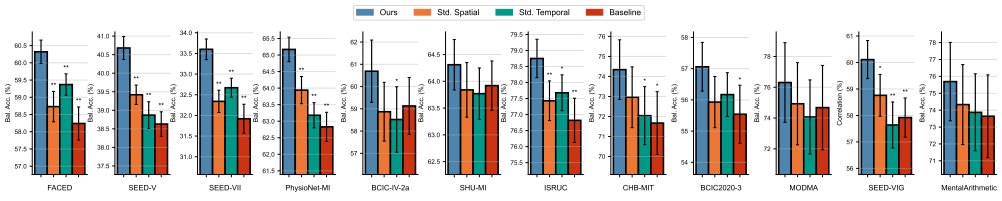


Figure 7: Positional-encoding ablation under end-to-end fine-tuning.

Table 8: Frozen-probing, full metrics (mean±std, %).

	FACED, 9-class			SEED-V, 5-class			SEED-VII, 7-class		
Method	Bal.Acc	Kappa	F1	Bal.Acc	Kappa	F1	Bal.Acc	Kappa	F1
LaBraM	16.13±10.04	5.61±11.21	9.88±13.37	32.63±0.53	15.45±0.69	32.62±0.87	15.35±1.99	1.29±2.42	7.09±6.15
CBraMod	25.84±1.36	16.32±1.60	21.77±1.46	24.36±0.38	5.29±0.47	20.39±1.18	17.32±0.27	3.48±0.35	14.23±1.43
CSBrain	36.59±0.96	28.56±1.07	36.59±0.94	20.00±0.00	0.00±0.00	8.87±0.00	14.28±0.01	-0.01±0.01	4.76±0.72
REVE	37.76±0.54	29.77±0.65	37.63±0.63	27.82±0.72	9.66±0.87	27.56±1.15	14.29±0.00	0.00±0.00	3.76±0.42
DeeperBrain	**50.96**±0.37	**44.38**±0.40	**50.36**±0.36	**35.08**±0.26	**18.75**±0.29	**35.39**±0.29	**23.26**±0.22	**10.71**±0.24	**22.83**±0.53
	PhysioNet-MI, 4-class			BCIC-IV-2a, 4-class			SHU-MI, 2-class		
Method	Bal.Acc	Kappa	F1	Bal.Acc	Kappa	F1	Bal.Acc	AUC-PR	AUROC
LaBraM	46.52±0.58	28.70±0.78	46.62±0.45	38.11±1.04	17.48±1.39	33.50±1.66	59.48±3.31	65.28±1.21	**66.92**±1.38
CBraMod	53.73±0.39	38.30±0.54	53.91±0.46	33.06±0.63	10.74±0.84	21.73±0.59	57.18±2.22	64.96±0.75	64.85±0.49
CSBrain	26.19±1.60	2.04±2.86	15.52±6.61	26.70±1.05	2.27±1.40	19.29±3.78	53.05±0.82	54.95±1.78	54.69±0.91
REVE	25.01±0.02	0.01±0.03	10.03±0.05	42.73±1.12	23.63±1.49	38.66±1.31	59.38±1.14	66.77±0.35	65.63±0.83
DeeperBrain	**56.57**±0.73	**42.10**±0.98	**56.74**±0.72	**51.01**±2.23	**34.68**±2.98	**49.23**±3.09	**60.81**±1.80	**67.01**±0.27	66.57±0.25
	ISRUC, 5-class			CHB-MIT, 2-class			BCIC2020-3, 5-class		
Method	Bal.Acc	Kappa	F1	Bal.Acc	AUC-PR	AUROC	Bal.Acc	Kappa	F1
LaBraM	71.79±0.95	67.94±1.08	74.55±0.90	55.31±3.45	28.76±3.90	89.08±1.95	24.83±2.61	6.03±3.26	22.76±5.82
CBraMod	37.27±5.90	24.87±5.63	37.16±3.23	58.93±2.85	36.66±7.10	86.20±2.55	20.24±0.52	0.30±0.65	8.74±2.12
CSBrain	22.24±2.25	3.01±3.18	22.61±3.11	62.68±4.42	8.13±3.45	75.48±4.18	32.19±1.40	15.23±1.75	30.96±2.18
REVE	73.55±0.47	68.78±0.27	75.72±0.42	63.93±3.70	**46.68**±4.06	**92.11**±1.30	39.04±9.59	23.80±11.99	36.32±14.87
DeeperBrain	**74.10**±0.58	**70.07**±0.60	**76.71**±0.68	**64.95**±2.55	38.46±4.84	88.73±3.75	**43.33**±0.64	**29.17**±0.80	**42.69**±0.73
	MODMA, 2-class			SEED-VIG, regression			MentalArith, 2-class		
Method	Bal.Acc	AUC-PR	AUROC	Corr.*	R2	RMSE↓	Bal.Acc	AUC-PR	AUROC
LaBraM	43.12±7.52	43.41±9.28	30.02±9.01	56.92±1.56	18.51±4.44	28.66±0.80	60.69±9.18	69.67±2.02	**83.06**±1.94
CBraMod	46.00±3.23	37.55±1.73	36.81±3.62	37.19±3.20	9.12±3.17	30.27±0.53	54.79±1.98	47.15±2.70	65.40±1.73
CSBrain	52.38±10.83	55.94±11.95	52.42±10.82	22.31±3.89	4.38±1.89	31.06±0.31	54.65±3.77	38.53±10.19	59.32±6.29
REVE	53.00±3.48	61.32±9.19	56.75±9.15	50.32±4.35	10.61±13.46	29.94±2.27	67.29±7.89	**70.84**±2.13	82.69±1.54
DeeperBrain	**63.36**±2.40	**70.37**±1.00	**70.39**±1.13	**59.85**±1.24	**31.43**±1.52	**27.52**±0.30	**68.05**±0.40	54.57±2.52	74.79±0.87

*SEED-VIG correlation values multiplied by 100 for readability (standard range $r \in [-1, 1]$).

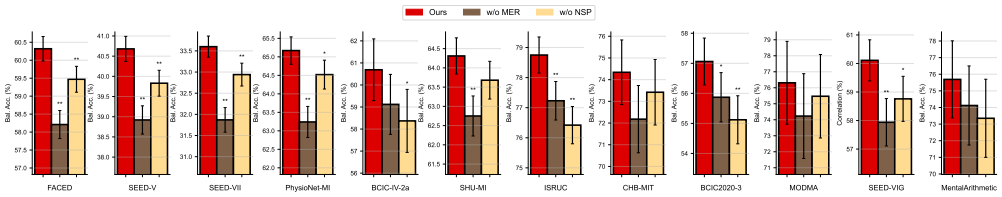


Figure 8: Pretraining-objective ablation under end-to-end fine-tuning.

G Related Work

G.1 Traditional EEG Decoding

EEG decoding infers cognitive, affective, or pathological states from scalp-recorded electrical activity. Early methods relied on handcrafted features—time-domain statistics, spectral power, and connectivity measures—combined with classifiers such as support vector machines or linear discriminant analysis (McFarland et al., 2006; Lotte et al., 2007; Wu et al., 2014; Wang et al., 2023c). These approaches require substantial domain expertise and generalize poorly across subjects and recording conditions.

Deep learning has since shifted the field toward end-to-end representation learning (LeCun et al., 2015). Convolutional Neural Networks capture local spatiotemporal patterns (Schirrmeister et al., 2017; Sakhavi et al., 2018; Lawhern et al., 2018; Ding et al., 2022), while Recurrent Neural Networks model long-range temporal dependencies for sleep staging and BCI control (Wang et al., 2018; Phan et al., 2019). Hybrid CNN-LSTM models combine local feature extraction with sequential modeling (Supratak et al., 2017; Zhang et al., 2019a; Dar et al., 2020; Li et al., 2022; Wang et al., 2023a). The Transformer architecture has been introduced to model global spatiotemporal interactions via self-attention (Song et al., 2021; Liu et al., 2021a; Du et al., 2022; Phan et al., 2022),

with several works integrating CNNs and Transformers to jointly exploit local invariance and global context (Song et al., 2022; Peh et al., 2022; Wang et al., 2023b; Zhou et al., 2024b; Wang et al., 2024c). Graph Neural Networks explicitly model inter-channel relationships using predefined or learned connectivity (Jia et al., 2020; 2021; Ding et al., 2023). Despite success on specific benchmarks, these supervised methods are constrained by the cost of labeled data and poor cross-dataset generalization, motivating the shift toward self-supervised pretraining.

G.2 EEG Foundation Models

Foundation models have transformed vision, language, and multimodal learning (Bommasani et al., 2021; Devlin et al., 2018; He et al., 2022; Radford et al., 2021; Achiam et al., 2023; Kirillov et al., 2023; Brooks et al., 2024). In the EEG domain, early self-supervised efforts focused on pretext tasks to extract invariant representations from unlabeled data (Banville et al., 2021). BENDR (Kostas et al., 2021) employs contrastive learning on generic EEG embeddings, while MAEEG (Chien et al., 2022) uses a masked autoencoder for signal reconstruction. Subsequent works extended these concepts to intracranial recordings: BrainBERT (Wang et al., 2022) reconstructs masked stereo-EEG spectrograms, and Brant (Zhang et al., 2023) introduces a Transformer for long-term dependency modeling, with Brant-2 (Yuan et al., 2024) and Brant-X (Zhang et al., 2024) further incorporating scalp EEG and multi-modal alignment.

Recent research increasingly targets scalp EEG. BIOT (Yang et al., 2023) employs a linear transformer for joint pretraining across biosignals; EEG2Rep (Foumani et al., 2024) predicts masked segments in a latent space; LaBraM (Jiang et al., 2024b) introduces neural token prediction for EEG patches; Neuro-GPT (Cui et al., 2024) uses a decoder-only architecture with autoregressive pretraining; and EEGPT (Wang et al., 2024a) employs a dual objective for universal feature extraction. Approaches such as NeuroLM (Jiang et al., 2025b), UniMind (Lu et al., 2025), and ELASTIQ (Jiang et al., 2025a) explore harnessing Large Language Models by treating EEG as a foreign language. Architecturally, CBraMod (Wang et al., 2025a) uses criss-cross attention with asymmetric positional encoding, EEGMamba (Wang et al., 2025b) and CodeBrain (Ma et al., 2025) leverage state space models, and LUNA (Döner et al., 2025) unifies diverse electrode layouts via learned queries. Other innovations include mixture-of-experts in NeurIPT (Fang et al., 2025), structured sparse attention in CSBrain (Zhou et al., 2025b), discrete tokenization in BrainOmni (Xiao et al., 2025), and massive-scale pretraining in REVE (Ouahidi et al., 2025). While these models improve transferability, they predominantly adapt generic architectures from vision or language and rely heavily on end-to-end fine-tuning, typically struggling under frozen-probing protocols (Lee et al., 2025; Xiong et al., 2025; Wu et al., 2025). DeeperBrain complements this line by explicitly encoding neurophysiological priors into both architecture and objective.

H Pretraining Hyperparameters

Table 9 summarizes the full pretraining configuration. DeeperBrain is implemented in Python 3.13.5 with PyTorch 2.8.0 and CUDA 12.9, and is pretrained on a single machine with an Intel Xeon Gold 6226R CPU and one NVIDIA RTX A5000 GPU, completing in approximately 7 hours. The backbone is a 12-layer pre-normalization Transformer encoder with embedding dimension $D{=}200$, feed-forward hidden dimension $D_{\text{ff}}{=}800$, and 8 self-attention heads. Input EEG is partitioned into non-overlapping 1-second spatiotemporal patches (200 time steps at 200 Hz). A random patch-wise masking strategy with a mask ratio of 50% replaces masked patches with zero vectors in the raw voltage space. The model is trained for 2 epochs with batch size 16. The two pretraining objectives are weighted equally, $\lambda_{\text{MER}} = \lambda_{\text{NSP}} = 1.0$. Optimization uses AdamW with base learning rate $5{\times}10^{-4}$, weight decay $5{\times}10^{-2}$, and default moments $\beta_1{=}0.9$, $\beta_2{=}0.999$. No learning-rate warmup is used, following the stability of the pre-normalization design (Wang et al., 2019). Gradient clipping with maximum norm 1.0 stabilizes training. NSP pseudo-labels are precomputed and cached to avoid recomputation overhead.

The channel encoding initializes the decay scale α so that τ corresponds to a typical inter-electrode distance of 8 cm, with numerical-stability constant $\delta = 10^{-5}$. The temporal encoding uses $K{=}8$ logarithmically spaced oscillatory frequencies spanning 0.01–0.5 Hz and $M{=}8$ decay constants spanning adaptation time constants $\tau \in [1, 100]$ s; both bases are held fixed during training. NSP targets are computed per 1-second patch: relative spectral power over five canonical bands (delta

Table 9: Pretraining hyperparameters.

Component	Configuration
Backbone	12-layer pre-norm Transformer
Embedding dim D	200
FFN hidden dim D_{ff}	800
Attention heads	8
Patch length	1 s (200 samples)
Mask ratio	0.5
Epochs	2
Batch size	16
Optimizer	AdamW
Learning rate	5×10^{-4}
Weight decay	5×10^{-2}
β_1, β_2	0.9, 0.999
LR warmup	none
Gradient clip norm	1.0
λ_{MER}, λ_{NSP}	1.0, 1.0
Smooth L1 β	1.0
Hardware	1× NVIDIA RTX A5000
Runtime	∼7 hours

1–4 Hz, theta 4–8 Hz, alpha 8–13 Hz, beta 13–30 Hz, gamma 30–75 Hz), phase-locking value summarized per band (mean and std of the upper-triangular PLV matrix, 10D), cross-frequency coupling for canonical theta–gamma, alpha–gamma, and beta–gamma pairs (3D), and sample entropy with m=2, r=0.2 std(x) (1D), yielding 19-dimensional targets per patch.

I Discussion

I.1 Implications

DeeperBrain advances EEG foundation modeling by bridging deep learning and established neurophysiological principles. Whereas prior approaches rely on general-purpose sequence modelers to implicitly infer neural patterns, DeeperBrain explicitly integrates volume-conduction-aware spatial priors and multi-scale temporal dynamics, aligning the model structure with the physical and functional organization of the brain. This principled approach enhances performance across diverse downstream tasks and improves neuroscientific interpretability, ensuring that learned representations encode canonical markers including spectral organization, functional connectivity, cross-frequency coordination, and dynamic complexity.

From an engineering perspective, the robust performance under both end-to-end fine-tuning and frozen-probing indicates that neurodynamics-guided self-supervision is more effective than reconstruction alone. This has practical implications for deploying universal BCIs where labeled data are scarce and inter-subject variability is high: a single pretrained DeeperBrain model can serve as a universal encoder for a wide range of BCI tasks, requiring minimal task-specific adaptation and lowering the barrier for real-world application.

I.2 Trade-offs of Fixed Neurophysiological Priors

The NSP objective aligns the pretrained representation with a fixed set of established EEG descriptors—spectral power, phase synchronization, cross-frequency coupling, and signal complexity. This alignment is a deliberate design choice: it injects domain knowledge that pure data-driven pretraining cannot infer from limited corpora, and the empirical results confirm its effectiveness across diverse downstream tasks. However, this choice introduces an inherent trade-off. By supervising the encoder to preserve information that the existing EEG literature already considers useful, NSP may bias the representation toward current analytical conventions. Information not well summarized by band power, global PLV statistics, the chosen CFC definitions, or sample entropy—for instance, higher-order nonlinear interactions or non-stationary transient dynamics—may receive less emphasis than a purely data-driven model might discover. The current benchmark suite, while spanning eight task categories, cannot fully rule out this bias because many evaluated tasks are naturally compatible with the selected descriptors. A key question for future work is whether NSP's advantage persists, diminishes, or transforms as pretraining data scales and diversifies: if handcrafted super-

vision mainly compensates for limited data, its benefit may narrow at larger scale; conversely, if the priors encode genuine neurophysiological structure, they should continue to complement data-driven learning. Resolving this requires scaling experiments and evaluation on paradigms whose relevant variables are not captured by current descriptors—for example, active BCI protocols (P300, SSVEP) and tasks involving novel cognitive mechanisms. Per-component ablations of the NSP statistics would further clarify whether the gain is distributed across all four descriptors or driven by a subset, and whether removing any single component harms tasks unrelated to that descriptor.

I.3 Future Directions

Several avenues extend the current framework. First, DeeperBrain relies on standardized 3D electrode coordinates; integrating coordinate estimation or learned spatial priors would broaden applicability to settings where exact montage geometry is unavailable. Second, the NSP objective currently utilizes four classical statistics; expanding this set to higher-order dynamics such as metastability and criticality would further align representations with complex brain states. Third, while stochastic channel reorganization accommodates variable channel counts, explicitly modeling montage topology shifts would enhance robustness to heterogeneous acquisition protocols.

The pretraining corpus, while spanning 14 heterogeneous datasets, retains demographic biases toward specific populations due to the geographic concentration of data sources; extending to globally representative EEG data would enhance fairness and generalizability. Finally, the current reliance on fixed-duration patches limits flexibility; future work will explore streaming or variable-length inference mechanisms to facilitate real-time BCI applications. The probe-head family used for evaluation is task-dependent rather than a single standardized linear probe, reflecting the diversity of BCI output structures; a unified probing protocol would enable cleaner cross-model comparison.

J Reproducibility Statement

To support reproducibility, upon acceptance we will release: (i) the DeeperBrain source code, including the model architecture, the dual-objective pretraining pipeline (MER + NSP), the NSP pseudo-label computation module $\Psi(\cdot)$, and the stochastic channel-reorganization routine; (ii) the pretrained backbone weights; (iii) the task-specific prediction-head implementations and the per-task head configurations used in our experiments; (iv) the data preprocessing scripts that reproduce the per-dataset pipelines in Appendix B, including resampling, filtering, segmentation, and the 100 μV normalization; (v) the subject-level train/val/test splits for all 12 downstream datasets (Appendix C); and (vi) the evaluation scripts that compute the full multi-metric results reported in Appendix E.

All pretraining hyperparameters are listed in Table 9 (Appendix H); the downstream evaluation protocol (optimizer, learning-rate schedule, batch size, epoch budget) is specified in Appendix C. All 14 pretraining datasets are publicly available: TUEG (Obeid & Picone, 2016), PhysioNet 2018 (Goldberger et al., 2000a; Ghassemi et al., 2018), Siena Scalp EEG (Detti et al., 2020), Raw EEG Data (Trujillo, 2020), and ten OpenNeuro datasets (OpenNeuro, 2025) (individually cited in Appendix B). All 12 downstream datasets are likewise public and are cited in Table 5. The model is implemented in PyTorch 2.8.0 with CUDA 12.9 and pretrained on a single NVIDIA RTX A5000 in approximately 7 hours. Each downstream result is the mean$\pm$std over five independent random seeds. No human or animal subjects were newly collected for this work; all datasets used are pre-existing public releases with their own ethical approvals.