SPERA: Spherical Prior EEG Foundation Model with Geometry- and Frequency-Aware Latent Prediction

Minsu Kim[1], Ye-Sung Kim[2], Hyeseong Jeon[2],
Wooseok Hyung[2], Joshua Lee[2], Chang-Hwan Im[3]*
[1]Department of Electronic Engineering, Hanyang University, Seoul, Republic of Korea
[2]Department of Artificial Intelligence, Hanyang University, Seoul, Republic of Korea
[3]Department of Biomedical Engineering, Hanyang University, Seoul, Republic of Korea

Abstract

Electroencephalography (EEG) provides a non-invasive measure of ongoing neural activity, but building general-purpose EEG models remains challenging due to the heterogeneity of subjects, devices, and electrode montages. Existing EEG foundation models predominantly rely on reconstruction-based objectives defined on the observed signal, which contains both neural and non-neural components. We introduce the Spherical Prior EEG Representation Architecture (SPERA[†]), an EEG foundation model that adopts the joint-embedding predictive architecture (JEPA) to predict in latent space. SPERA introduces a Legendre-polynomial spatial prior, incorporated into attention to encode varying scalp electrode geometries. Two further components adapt the model to EEG: factorized temporal and spatial attention interleaved with periodic full-attention blocks, and a relational spectral regularizer aligning latent similarity structure with spectral views. Pretrained on approximately 80,000 hours of EEG from 29,048 subjects across 106 datasets, SPERA achieves the highest average balanced accuracy across nine downstream tasks spanning clinical, cognitive, and BCI applications. SPERA further exhibits strong parameter efficiency under linear probing and robustness across varying recording conditions, suggesting its potential as a general-purpose backbone for diverse EEG analyses.

1 Introduction

Electroencephalography (EEG) reflects dynamic neural activity and underlying brain states. With high temporal resolution and ease of acquisition, EEG is widely used across clinical diagnosis, cognitive neuroscience, and brain-computer interface (BCI) applications [133, 88, 17, 146]. Deep learning models have achieved strong performance in interpreting these signals [2, 120].

Despite this progress, building a general-purpose EEG model remains challenging due to the heterogeneity of EEG data. EEG signals vary significantly across individuals [121] and are recorded under diverse experimental conditions with different devices and electrode configurations [19]. This makes it difficult to learn representations that transfer across datasets and subjects.

Inspired by foundation models in vision and language [112, 40], recent EEG foundation models such as LaBraM [57], CBraMod [150], and REVE [94] have explored scalable self-supervised pretraining, using masked reconstruction or token-prediction objectives whose targets are derived from the input signal. EEG recordings contain both neural activity and non-neural components such as muscle and ocular artifacts. Reconstruction-based objectives define their prediction targets on all

*Corresponding author. ich@hanyang.ac.kr
[†]Code: https://github.com/sees4321/SPERA_EEG

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

components present in the observed signal, which motivates considering prediction targets defined at the representation level. We therefore adopt the joint-embedding predictive architecture (JEPA) [8].

We propose the Spherical Prior EEG Representation Architecture (SPERA), a representation learning framework for EEG built on latent-space prediction. Beyond this objective, our central contribution is a Legendre polynomial spherical prior incorporated into attention, providing a learnable pairwise geometric bias that reflects the spherical structure of scalp EEG and applies to arbitrary montages. Two further components adapt the backbone to EEG: factorized temporal and spatial attention interleaved with periodic full-attention blocks for cross-axis interactions, and a relational spectral regularizer that induces frequency-aware latents by aligning pairwise similarity structure with spectral views.

Building on these components, we pretrain SPERA on a multi-domain EEG corpus comprising approximately 80,000 hours of recordings from 29,048 subjects across 106 datasets. We evaluate SPERA on nine diverse downstream tasks spanning clinical, cognitive, and BCI applications and compare it with prior EEG foundation models. We further investigate the scaling behavior of latent-space prediction under EEG-aware design, with a focus on parameter efficiency under linear probing (LP).

2 Related work

Self-supervised objectives for EEG Most existing EEG foundation models [57, 150, 94, 79, 75] rely on masked modeling objectives with targets derived from the input signal, including reconstructed signal patches and discrete neural tokens. Such targets therefore include both neural and non-neural components of the observed signal. JEPA models [8, 9] instead predict in latent space, and several EEG models have moved in this direction: latent prediction against a momentum-updated target encoder [87], representation alignment paired with auxiliary patch reconstruction [149], generative diffusion whose denoising activations serve as representations [110], and direct applications of JEPA to EEG [45, 53, 95]. However, none of these explicitly encode the pairwise geometric relationships between electrodes, which vary across montage configurations. We use a latent prediction objective over a montage-variable token grid, with masking of temporal regions and spatial neighborhoods defined by 3D electrode coordinates, and scale it to a multi-domain corpus spanning over a hundred datasets.

Electrode geometry in EEG modeling A central challenge for EEG foundation models is the heterogeneity of electrode montages across datasets. BIOT adds channel embeddings and relative position embeddings [155], CBraMod uses asymmetric conditional positional encoding to adapt to diverse EEG formats [150], and REVE introduces a 4D positional encoding scheme for arbitrary electrode arrangements [94]. Recent JEPA-style EEG models also introduce spatial inductive biases, including 3D coordinate-based positional embeddings [45], video-like spatio-temporal tokenization [53], and coordinate-aware channel mixing with Fourier features [95]. We instead encode electrode geometry directly into attention using a Legendre-based spherical prior, providing a learnable pairwise bias derived from the biophysical geometry of the scalp [102, 21].

Spatio-temporal modeling in EEG EEG signals exhibit distinct spatial and temporal dependencies, motivating attention factorization along each axis. A related factorized design has been used in video understanding, where TimeSformer and ViViT apply temporal and spatial attention sequentially within each block [18, 7]. In EEG, CBraMod models spatial and temporal dependencies in parallel using criss-cross attention [150], while CSBrain [162] combines cross-scale spatio-temporal tokenization with structured sparse attention over local temporal windows and anatomical regions. State-space models use Mamba-based backbones for EEG sequence modeling [151]. LuMamba uses learned-query cross-attention to unify variable electrode configurations before temporal modeling [20]. These approaches capture spatial and temporal structure in different ways. However, our setting introduces two additional design considerations: EEG channel montages vary across datasets, unlike the fixed spatial grids used in these video models, and purely factorized attention restricts direct cross-axis interaction. We address these considerations by combining explicit electrode-geometry encoding with a hybrid backbone that interleaves factorized and full-attention blocks.

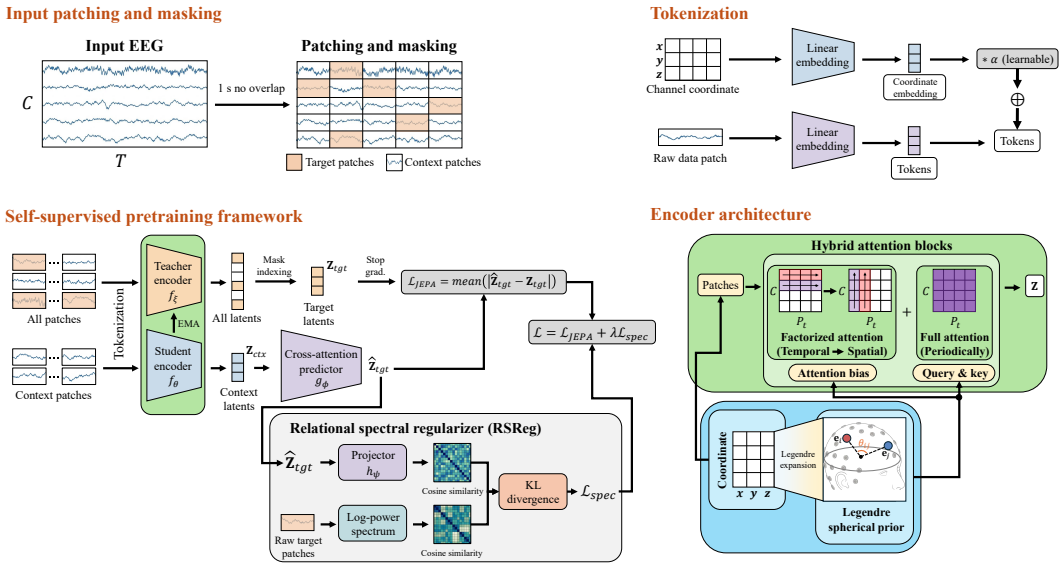

Figure 1: Overview of the SPERA pretraining framework.

Spectral objectives in EEG pretraining Frequency components are key features of EEG signals, motivating spectral objectives in several pretraining methods. LaBraM trains its tokenizer with vector-quantized neural spectrum prediction [57], and Pradeepkumar et al. [109] introduce time-frequency motif learning for frequency-aware EEG tokenization. We complement this line of work with a relational spectral regularizer that aligns pairwise similarity structure between latents and spectral views, without predicting absolute spectral values.

3 Methods

We denote an EEG recording as $\mathbf{X} \in \mathbb{R}^{C\times T}$, where C is the number of electrodes and T is the number of time samples. Each electrode is associated with a 3D anatomical coordinate, collectively represented as $\mathbf{E} \in \mathbb{R}^{C\times 3}$. To tokenize $\mathbf{X}$, each of the C channels is divided into P_t non-overlapping 1-second temporal patches at a sampling rate of $f_s = 200$ Hz, resulting in a total of $N = C \cdot P_t$ tokens. Each token is uniquely indexed by a (channel, time-patch) tuple (c, p) on the spatio-temporal grid. Unlike token sequences in language or vision, EEG signals have dataset-specific spatio-temporal grids, as channel montages vary across datasets. This cross-dataset variability and the two-axis structure of the token space motivate the architectural choices described in this section.

3.1 Self-supervised objective: spatio-temporal JEPA

JEPA objective Figure 1 illustrates the overall pretraining framework. Raw EEG contains muscle and ocular artifacts that persist after preprocessing, so an input-space reconstruction objective requires reproducing them alongside neural activity. We adopt JEPA [8], which predicts latent representations of masked target regions from visible context. SPERA consists of a student encoder f_θ and a teacher encoder f_ξ; the teacher receives no gradients and is updated as an exponential moving average (EMA) of f_θ. On the spatio-temporal grid, we partition the token indices into disjoint context and target sets, $\mathcal{C}$ and $\mathcal{T}$. The student encodes only the context tokens to produce $\mathbf{Z}_{\text{ctx}}$. The teacher encodes the full token sequence, and $\mathbf{Z}_{\text{tgt}}$ is obtained by selecting the teacher outputs at the target positions. A two-layer cross-attention predictor g_ϕ predicts $\hat{\mathbf{Z}}_{\text{tgt}}$ from $\mathbf{Z}_{\text{ctx}}$. For each target position, a query is formed by adding the target coordinate embedding to a single learnable vector shared across all target positions, and is refined by cross-attending to $\mathbf{Z}_{\text{ctx}}$. Following prior work [9], the model is trained to minimize the ℓ_1 distance in latent space:

$$\mathcal{L}_{\text{JEPA}} = \frac{1}{|\mathcal{T}|D} \sum_{(c,p)\in\mathcal{T}} \left\| \hat{\mathbf{z}}_{c,p} - \mathbf{z}_{c,p} \right\|_1 , \tag{1}$$

where $\hat{\mathbf{z}}_{c,p}$ and $\mathbf{z}_{c,p}$ denote the predicted and teacher latents at position (c, p), and D is the embedding dimension.

Masking strategy For each sample, temporal masking is activated with probability 0.8, and spatial masking is activated independently with probability 0.8. If neither is activated, we default to temporal masking. Temporal masks follow a semantic subsequence preserving (SSP) strategy [87]: rather than masking random individual time patches, we preserve contiguous blocks of patches to ensure the context carries coherent temporal structure. Spatial masks select a random electrode as a center and mask its k-nearest neighbors based on 3D Euclidean distance, with $k = \max(1, \lfloor \rho_s \cdot C \rceil)$ for a sampled spatial ratio ρ_s. Masking ratios are drawn uniformly from [0.15, 0.35] for time and [0.10, 0.30] for space (ρ_s).

Student augmentation We apply EEG-specific perturbations to the student's context input: temporal amplitude scaling and additive Gaussian noise, together with Gaussian jitter to electrode coordinates. Augmentation hyperparameters are reported in Appendix C.1.

3.2 Encoder architecture

Dense full attention allows interactions among all spatio-temporal token pairs but does not explicitly distinguish temporal and spatial dependencies. We adopt a hybrid backbone of factorized attention blocks interleaved with periodic full-attention blocks for cross-axis interactions.

Hybrid attention blocks The base block factorizes attention along each axis separately. A *temporal attention* pass operates independently over the P_t time patches within each channel, followed by a *spatial attention* pass over the C channels at each time patch. Every fourth block is a full self-attention block that attends jointly over all $N = C \cdot P_t$ tokens, enabling cross-axis interactions. Both attention block types share standard transformer design choices: a pre-norm architecture with layer normalization (LN), a SwiGLU MLP [128], and rotary position embeddings (RoPE) [139] applied along the temporal axis (in both the encoder and the predictor).

3.3 Electrode geometry encoding with spherical prior

To handle varying electrode montages across datasets, we incorporate simple coordinate embeddings and Legendre-based spherical priors. The Legendre prior is implemented in two forms: a pairwise Legendre bias for factorized attention blocks and Legendre anchor encoding for full-attention blocks. The two are assigned to attention types based on computational considerations analyzed in Section 4.5.

Coordinate embedding A direct approach to spatial conditioning is a per-token additive embedding derived from electrode coordinates. We normalize each coordinate onto the unit sphere and pass it through a linear projection. The resulting channel-wise embedding is scaled by a learnable scalar α, initialized to zero, and shared across all temporal patches of the same electrode.

Legendre spatial bias Scalp EEG electrodes lie approximately on a spherical head surface, and we exploit this structure to build a biophysically motivated spatial prior on the sphere itself. Legendre polynomials provide a natural basis for parameterizing pairwise functions of angular separation on the sphere and have been used in conventional EEG analysis [102, 21]. We therefore use a truncated Legendre expansion to encode pairwise electrode geometry as an additive attention bias.

Within each factorized attention block, the spatial attention pass computes attention over C channels at a fixed time patch. To encode pairwise geometric relationships between electrodes, we introduce a learnable additive attention bias b_{ij} between channel i and channel j, defined as a truncated Legendre polynomial expansion of their cosine similarity on the unit sphere:

$$b_{ij} = \sum_{\ell=0}^{L} w_\ell \, P_\ell\big(\mathbf{e}_i^\top \mathbf{e}_j\big) \,, \tag{2}$$

where $\mathbf{e}$ denotes unit-normalized coordinates, L is the maximum polynomial degree, P_ℓ is the ℓ-th Legendre polynomial, and w_ℓ are learnable scalar coefficients. Unlike classical formulations that fix w_ℓ to a closed-form decay (e.g., $(2\ell + 1)/[\ell^m(\ell + 1)^m]$ in spherical splines), we let the coefficients

be learned end-to-end. This bias is pairwise, geometry-aware, and dataset-agnostic, requiring no assumption about the number or identity of electrodes.

Legendre anchor encoding The pairwise Legendre bias in Eq. (2) directly encodes the angular relationship between each pair of electrodes. In the dense implementation, applying this pairwise bias in full-attention blocks requires materializing an $N \times N$ score-bias matrix, which incurs substantial memory and computational overhead. We therefore construct an anchor-based alternative from the same Legendre geometry. Instead of expanding the relationship between every pair of electrodes, we represent each electrode by its Legendre-expanded similarities to a finite set of learnable anchor directions on the unit sphere. These anchor-based geometry encodings are then injected into the queries and keys of the full-attention blocks.

We compute a per-channel spatial feature $\boldsymbol{\phi}_c$ using a set of M learnable anchor points $\hat{\mathbf{a}}_m$ on the unit sphere, expanding the cosine similarity between each electrode and each anchor via Legendre polynomials:

$$g_{c,m} = \sum_{\ell=0}^{L} \omega_{m,\ell}\, P_\ell\big(\mathbf{e}_c^\top \hat{\mathbf{a}}_m\big)\,, \qquad \boldsymbol{\phi}_c = \mathrm{LN}(W_{\mathrm{proj}}\,[g_{c,1}, \ldots, g_{c,M}])\,, \tag{3}$$

$$\tilde{\mathbf{q}}_i = \mathbf{q}_i + \Gamma_Q \boldsymbol{\phi}_{c_i}, \qquad \tilde{\mathbf{k}}_j = \mathbf{k}_j + \Gamma_K \boldsymbol{\phi}_{c_j}, \qquad s_{ij} = \frac{\tilde{\mathbf{q}}_i^\top \tilde{\mathbf{k}}_j}{\sqrt{d}}, \tag{4}$$

where $\omega_{m,\ell}$ are learnable anchor-specific coefficients and W_{proj} is a linear projection. Here, c_i denotes the channel index of token i, and d is the attention head dimension. The geometry feature $\boldsymbol{\phi}_c$ is added to queries and keys via projections Γ_Q and Γ_K, applied after temporal RoPE. Expanding Eq. (4) gives four terms: content–content, an anchor-induced geometry–geometry term, and two content–geometry cross terms (see Appendix F.3 for details). The coefficients in Eqs. (2)–(3), as well as the anchor directions $\hat{\mathbf{a}}_m$, are shared globally across layers and heads. In the anchor-based formulation, layer-specific adaptation is introduced through the block-specific projections Γ_Q, Γ_K.

3.4 Relational spectral regularizer

Frequency components are key features of EEG signals, but incorporating spectral objectives into a JEPA framework is not straightforward. Direct spectrum reconstruction would introduce an absolute signal-derived target, whereas our goal is to regularize spectral relationships without reconstructing spectral values. Therefore, we introduce a relational spectral regularizer (RSReg) that aligns the pairwise similarity structure of the predicted representations with that of the corresponding spectral targets. This design is inspired by relation-preserving distillation [98, 97], which supervises pairwise relational structure. Here, the target relations are computed deterministically from spectral views of EEG patches, not from another network.

For each target token (c, p), $\mathbf{s}_{c,p}$ is the log-power spectrum of its 1-second raw patch, computed via a Hann-windowed FFT using frequency bins within the 1–45 Hz band, and z-score normalized across frequency bins. The predictor output $\hat{\mathbf{z}}_{c,p}$ is passed through a small projector h_ψ to obtain $\mathbf{r}_{c,p} = h_\psi(\hat{\mathbf{z}}_{c,p})$. We align the pairwise similarity matrices of $\mathbf{r}$ and $\mathbf{s}$ through a KL divergence loss:

$$\mathcal{L}_{\mathrm{spec}} = \mathrm{KL}\Big(\sigma\big(\mathbf{S}\mathbf{S}^\top/\tau_s\big) \;\Big\|\; \sigma\big(\mathbf{R}\mathbf{R}^\top/\tau_z\big)\Big)\,, \tag{5}$$

where $\mathbf{S}$ and $\mathbf{R}$ are row-wise ℓ_2-normalized matrices of $\mathbf{s}$ and $\mathbf{r}$, respectively, $\sigma(\cdot)$ denotes row-wise softmax, and τ_s, τ_z are temperature hyperparameters. Similarity matrices are computed per recording, with diagonal entries masked before softmax. The KL loss is averaged over target-token rows within each recording, then equally across recordings in the effective batch. This preserves relative spectral relationships among tokens without constraining their absolute spectral values, inducing frequency-aware latent representations through relational spectral supervision. The total training loss is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{JEPA}} + \lambda\, \mathcal{L}_{\mathrm{spec}}. \tag{6}$$

Table 1: Summary of downstream task datasets.

Task	Dataset	# Samples	# Channels	Duration (s)	# Classes
Imagined speech classification	BCIC2020-3	6,000	64	3	5
Motor imagery classification	BCIC-IV-2a	5,184	22	4	4
	PhysioNet-MI	9,837	64	4	4
Emotion recognition	FACED	10,332	32	10	9
Mental stress detection	MAT	1,707	20	5	2
Sleep staging	ISRUC	88,140	6	30	5
Mental disorder diagnosis	Mumtaz	7,143	19	5	2
EEG abnormality detection	TUAB	409,455	16	10	2
Event type classification	TUEV	112,491	16	5	6

4 Experiments

4.1 Pretraining

Dataset collection We collected a large-scale pretraining corpus spanning approximately 80,000 hours of EEG recordings from 29,048 subjects, yielding 2.67 million channel-hours and 9.60 billion tokens in total. To the best of our knowledge, this is the largest multi-domain EEG pretraining corpus reported for a general-purpose EEG foundation model. The corpus is organized into three categories (clinical, cognitive, and BCI), comprising 82.08%, 16.21%, and 1.71% of the total recording time, respectively. The pretraining dataset exhibits substantial heterogeneity in electrode montages, covering a wide range of channel counts and layouts across sources. A summary of the dataset composition and a complete list of included sources are provided in Appendix B.

Preprocessing pipeline We apply a standardized preprocessing pipeline to all recordings. Power line noise is removed via a notch filter at 50 or 60 Hz depending on the recording country. EEG recordings are then bandpass filtered between 0.5 and 75 Hz and resampled to $f_s = 200$ Hz. Each channel is z-score normalized to account for cross-dataset variability in electrode impedance and amplifier gain. For datasets that provide digitized 3D electrode locations, we use the recorded coordinates directly; for those that specify only channel labels (e.g., the 10–20 system), we assign coordinates from the corresponding standard montage in MNE-Python [39]. In addition, amplitude values exceeding $\pm$ 15 standard deviations are clipped to suppress residual artifacts, as in [94]. Each recording is segmented into non-overlapping windows of either 10 or 60 seconds depending on the source task. Among the 60-second segments, those whose token count exceeds 4,096 are further split into 30-second chunks for training efficiency, preserving contiguous temporal structure.

Implementation details We train all models with AdamW [77]. The learning rate follows a cosine schedule with linear warmup, and the teacher encoder is updated via an EMA with a cosine momentum schedule. We pretrain four model sizes: SPERA-Tiny (18M), SPERA-Small (48M), SPERA-Base (194M), and SPERA-Large (396M). Full hyperparameter details are provided in Appendix C.1.

4.2 Experimental setup

Downstream task datasets We evaluate SPERA on nine datasets spanning diverse clinical, cognitive, and BCI tasks: BCIC2020-3 [56], BCIC-IV-2a [143], PhysioNet-MI [38, 122], FACED [23], mental arithmetic (MAT) [164], ISRUC [63], Mumtaz [89], TUAB [91], and TUEV [91]. These nine tasks are the intersection of the CBraMod and REVE evaluation suites. We adopt the same train/validation/test splits and preprocessing pipeline as baseline EEG foundation models [155, 57, 150, 94]. Table 1 summarizes these datasets.

Baseline models and evaluation protocol We compare SPERA against both non-foundation and foundation model baselines. Non-foundation baselines include EEGNet [70], EEGConformer [135], SPaRCNet [58], ContraWR [156], CNN-Transformer [100], FFCL [73], ST-Transformer [134], EEG2Rep [87], and EEGDM [110], while foundation model baselines include BIOT [155], LaBraM [57], CBraMod [150], REVE [94], CSBrain [162], CodeBrain [79], EEGMamba [151], GRAM [75], LuMamba [20], EEGPT [149], and Laya [95].

Table 2: Experimental results on nine downstream tasks. SPERA models are evaluated under the LPFT protocol, while baseline results are taken from the original publications using their reported downstream adaptation protocols. The results are reported as mean balanced accuracy ± standard deviation over five random seeds. The best performance for each task is highlighted in bold.

Methods	BCIC2020-3	BCIC-IV-2a	PhysioNet-MI	FACED	MAT
EEGNet	0.4413 ± 0.0096	0.4482 ± 0.0094	0.5814 ± 0.0125	0.4090 ± 0.0122	0.6770 ± 0.0116
EEGConformer	0.4506 ± 0.0133	0.4696 ± 0.0106	0.6049 ± 0.0104	0.4559 ± 0.0125	0.6805 ± 0.0123
SPaRCNet	0.4426 ± 0.0156	0.4635 ± 0.0117	0.5932 ± 0.0152	0.4673 ± 0.0155	0.6879 ± 0.0107
ContraWR	0.4257 ± 0.0162	0.4678 ± 0.0125	0.5892 ± 0.0133	0.4887 ± 0.0078	0.6631 ± 0.0097
CNN-Transformer	0.4533 ± 0.0092	0.4600 ± 0.0108	0.6053 ± 0.0118	0.4697 ± 0.0132	0.6779 ± 0.0268
FFCL	0.4678 ± 0.0197	0.4470 ± 0.0143	0.5726 ± 0.0092	0.4673 ± 0.0158	0.6798 ± 0.0142
ST-Transformer	0.4126 ± 0.0122	0.4575 ± 0.0145	0.6035 ± 0.0081	0.4810 ± 0.0079	0.6631 ± 0.0173
BIOT	0.4920 ± 0.0086	0.4748 ± 0.0093	0.6153 ± 0.0154	0.5118 ± 0.0118	0.6875 ± 0.0186
LaBraM-Base	0.5060 ± 0.0155	0.4869 ± 0.0085	0.6173 ± 0.0122	0.5273 ± 0.0107	0.6909 ± 0.0125
CBraMod	0.5373 ± 0.0108	0.5138 ± 0.0066	0.6417 ± 0.0091	0.5509 ± 0.0089	0.7256 ± 0.0132
CSBrain	0.6004 ± 0.0187	0.5657 ± 0.0071	0.6304 ± 0.0090	0.5752 ± 0.0042	0.7558 ± 0.0106
REVE-Base	0.5635 ± 0.0123	0.6396 ± 0.0095	0.6480 ± 0.0140	0.5646 ± 0.0164	0.7660 ± 0.0355
SPERA-Tiny	0.4827 ± 0.0173	0.5326 ± 0.0138	0.6117 ± 0.0093	0.4849 ± 0.0155	0.6971 ± 0.0229
SPERA-Small	0.5733 ± 0.0135	0.6085 ± 0.0156	0.6400 ± 0.0112	0.5132 ± 0.0167	0.7238 ± 0.0198
SPERA-Base	0.6504 ± 0.0122	0.6816 ± 0.0147	0.6726 ± 0.0072	0.5754 ± 0.0175	0.7778 ± 0.0216
SPERA-Large	$\mathbf{0.6883 \pm 0.0142}$	$\mathbf{0.7023 \pm 0.0144}$	$\mathbf{0.6817 \pm 0.0087}$	$\mathbf{0.5919 \pm 0.0186}$	$\mathbf{0.8021 \pm 0.0203}$

Methods	ISRUC	Mumtaz	TUAB	TUEV	Average
EEGNet	0.7154 ± 0.0121	0.9232 ± 0.0104	0.7642 ± 0.0036	0.3876 ± 0.0143	0.5941 ± 0.0037
EEGConformer	0.7400 ± 0.0133	0.9308 ± 0.0117	0.7758 ± 0.0049	0.4074 ± 0.0164	0.6128 ± 0.0044
SPaRCNet	0.7487 ± 0.0075	0.9316 ± 0.0095	0.7896 ± 0.0018	0.4161 ± 0.0262	0.6156 ± 0.0047
ContraWR	0.7402 ± 0.0126	0.9195 ± 0.0115	0.7746 ± 0.0041	0.4384 ± 0.0349	0.6119 ± 0.0053
CNN-Transformer	0.7363 ± 0.0087	0.9305 ± 0.0068	0.7777 ± 0.0022	0.4087 ± 0.0161	0.6133 ± 0.0045
FFCL	0.7277 ± 0.0182	0.9314 ± 0.0038	0.7848 ± 0.0038	0.3979 ± 0.0104	0.6085 ± 0.0044
ST-Transformer	0.7381 ± 0.0205	0.9135 ± 0.0103	0.7966 ± 0.0023	0.3984 ± 0.0228	0.6071 ± 0.0048
BIOT	0.7527 ± 0.0121	0.9358 ± 0.0052	0.7959 ± 0.0057	0.5281 ± 0.0225	0.6438 ± 0.0044
LaBraM-Base	0.7633 ± 0.0102	0.9409 ± 0.0079	0.8140 ± 0.0019	0.6409 ± 0.0065	0.6653 ± 0.0031
CBraMod	0.7865 ± 0.0110	0.9560 ± 0.0056	0.8289 ± 0.0022	0.6671 ± 0.0107	0.6898 ± 0.0051
CSBrain	0.7925 ± 0.0030	0.9643 ± 0.0155	0.8172 ± 0.0043	0.6903 ± 0.0059	0.7102
REVE-Base	0.7819 ± 0.0078	0.9644 ± 0.0097	0.8315 ± 0.0014	0.6759 ± 0.0229	0.7150 ± 0.0057
SPERA-Tiny	0.7719 ± 0.0124	0.9236 ± 0.0034	0.8104 ± 0.0044	0.7035 ± 0.0273	0.6687 ± 0.0064
SPERA-Small	0.7852 ± 0.0096	0.9450 ± 0.0048	0.8318 ± 0.0032	0.7384 ± 0.0258	0.7066 ± 0.0057
SPERA-Base	0.8137 ± 0.0075	0.9637 ± 0.0018	0.8488 ± 0.0035	0.7684 ± 0.0283	0.7502 ± 0.0050
SPERA-Large	$\mathbf{0.8142 \pm 0.0106}$	$\mathbf{0.9702 \pm 0.0016}$	$\mathbf{0.8608 \pm 0.0029}$	$\mathbf{0.7869 \pm 0.0286}$	$\mathbf{0.7665 \pm 0.0053}$

We evaluate SPERA under two protocols, LP and linear probing followed by fine-tuning (LPFT). LP freezes the pretrained encoder and trains a linear classification head on mean-pooled token features, while LPFT additionally fine-tunes the encoder after the LP stage. For each task, hyperparameters and checkpoints are selected on the validation set, and the held-out test set is used exclusively for final evaluation. Baseline results are extracted from the original publications [150, 94, 162]; their downstream adaptation protocols vary across studies (e.g., CBraMod uses an MLP classification head). Balanced accuracy is used as the primary metric; each experiment is repeated over five random seeds, and results are reported as mean ± standard deviation. For SPERA, the overall average is computed per seed as the unweighted mean across the nine datasets, and the reported standard deviation is taken across the resulting five seed-wise averages. Baseline results, including reported averages and standard deviations, are taken from the original publications.

4.3 Downstream task results

Table 2 summarizes LPFT performance across the nine downstream tasks and four SPERA model scales. Per-task Welch tests over five seeds, Holm-corrected across the nine comparisons, find significant gains for SPERA-Base on five tasks: BCIC2020-3, TUEV, BCIC-IV-2a, ISRUC, and TUAB. PhysioNet-MI is borderline ($p_{\mathrm{Holm}} = 0.052$) and no task shows a significant loss. Among the significant differences, the largest gains are observed on TUEV (+0.078; 6-class clinical event) and BCIC2020-3 (+0.050; 5-class imagined speech). The differences on FACED, MAT, and Mumtaz did not reach statistical significance after Holm correction. Mumtaz in particular shows limited headroom, with balanced accuracy remaining generally high across foundation models and model scales. SPERA

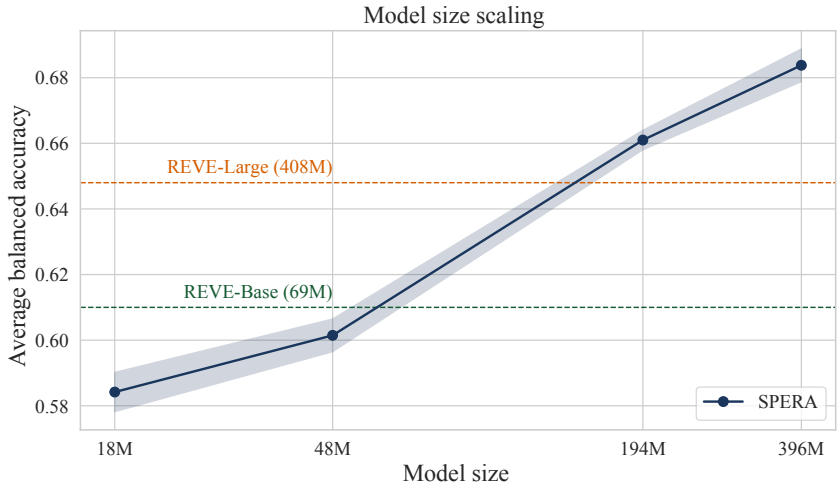

Figure 2: Model scaling under LP. Mean balanced accuracy across nine downstream tasks is shown for SPERA models ranging from 18M to 396M parameters. The shaded region denotes one standard deviation across five random seeds. Horizontal lines indicate the corresponding LP results of REVE-Base (69M) and REVE-Large (408M).

performs consistently on both BCIC-IV-2a (22 channels) and PhysioNet-MI (64 channels), two motor-imagery tasks with different electrode layouts. Per-task differences with confidence intervals, for both SPERA-Base and SPERA-Large, are reported in Appendix D.10.

4.4 Model scaling analysis

Table 2 shows that average downstream performance improves with model scale under LPFT. To further examine how the pretrained representations themselves scale, independently of full encoder fine-tuning, we evaluate the four model sizes (18M, 48M, 194M, and 396M) using LP across all nine downstream tasks. Per-task LP results are reported in Appendix D.9. Figure 2 shows that average balanced accuracy under LP improves with model size from 0.584 at 18M to 0.684 at 396M. The largest gain occurs between 48M and 194M (+0.060), which coincides with the largest jump in hidden dimension (512 $\rightarrow$ 896). The smaller improvement from 194M to 396M (+0.023) suggests diminishing returns at the current corpus size. A similar pattern is observed under LPFT in Table 2, where the average increases from 0.669 to 0.767 across the four scales, with a smaller gain from Base to Large. In addition, SPERA-Small (48M) achieves performance comparable to that of REVE-Base (69M), and SPERA-Base (194M) exceeds REVE-Large (408M) by +0.013 under LP despite using less than half the parameters.

4.5 Ablation study and efficiency analysis

To understand the contribution of each component, we conduct ablation studies on the encoder architecture, RSReg, and electrode geometry encoding. Given the substantial computational cost of repeatedly pretraining the SPERA-Base model (194M), we adopt the SPERA-Small model (48M) as a proxy, evaluated on motor imagery, sleep staging, and EEG abnormality detection tasks using LP. We jointly analyze accuracy (Table 3) and computational efficiency (Table 4) to assess both representational quality and viability under realistic compute budgets.

Component ablations We ablate three components: the hybrid attention backbone, RSReg, and the Legendre-based spatial priors. Neither full nor factorized attention alone matches our periodic hybrid design, and a variant that concentrates full-attention blocks at the end of the network ("Hybrid + late full" in Table 3) improves over full-only but still underperforms periodic interleaving. This suggests that periodic cross-axis interaction benefits EEG modeling. Removing RSReg lowers average accuracy by 0.012, with the largest drop on ISRUC (-0.021), suggesting a benefit of frequency-aware supervision. The Legendre prior improves average accuracy by 0.024 over coordinate embeddings

Table 3: Ablation studies on the encoder architecture, RSReg, and electrode geometry encoding. All models are based on SPERA-Small (48M) and evaluated using LP on motor imagery, sleep staging, and EEG abnormality detection tasks. Results are reported as mean balanced accuracy ± standard deviation over five random seeds. The best result per column is in **bold**; the second-best result is underlined.

Variant	PhysioNet-MI	ISRUC	TUAB	Average
Encoder architecture				
Full attention only	0.4450 ± 0.0124	0.7097 ± 0.0218	0.7749 ± 0.0024	0.6432 ± 0.0084
Factorized attention only	0.4833 ± 0.0091	0.7153 ± 0.0187	0.7865 ± 0.0066	0.6617 ± 0.0073
Hybrid + late full	0.4871 ± 0.0181	0.6973 ± 0.0155	0.7860 ± 0.0045	0.6568 ± 0.0081
RSReg				
No RSReg	0.5182 ± 0.0184	0.6994 ± 0.0205	0.7759 ± 0.0046	0.6645 ± 0.0093
Electrode geometry encoding				
Coordinate embedding only	0.4821 ± 0.0150	0.7009 ± 0.0268	0.7744 ± 0.0056	0.6525 ± 0.0104
Legendre bias only	$\underline{0.5193 \pm 0.0106}$	$\mathbf{0.7320 \pm 0.0182}$	$\mathbf{0.7929 \pm 0.0034}$	$\mathbf{0.6814 \pm 0.0071}$
Legendre anchor encoding only	0.5071 ± 0.0197	0.7137 ± 0.0201	0.7855 ± 0.0028	0.6688 ± 0.0094
Learnable pairwise MLP	0.4463 ± 0.0203	0.6679 ± 0.0264	0.7671 ± 0.0067	0.6271 ± 0.0113
REVE spatial encoding	0.4921 ± 0.0174	0.6910 ± 0.0212	0.7607 ± 0.0091	0.6479 ± 0.0091
Gaussian (geodesic)	0.4512 ± 0.0161	0.6977 ± 0.0225	0.7705 ± 0.0089	0.6398 ± 0.0101
SPERA-Small	$\mathbf{0.5227 \pm 0.0112}$	$\underline{0.7206 \pm 0.0240}$	$\underline{0.7869 \pm 0.0062}$	$\underline{0.6767 \pm 0.0110}$

Table 4: Computational efficiency and memory profiling using the SPERA-Small model. All measurements are conducted on a single NVIDIA RTX 5000 Ada GPU with a batch size of 1 and a sequence length $N = 3{,}840$ (64 channels $\times$ 60 patches). 'Mem' denotes peak GPU memory.

Architecture	Legendre prior	Params (M)	Matmul FLOPs (G)	Mem (GB)	Throughput (tok/s)
Full	Bias	37.9	653.0	2.51	20,260
	Anchor encoding	38.7	653.1	1.61	75,769
Factorized	Bias	50.6	399.0	1.70	97,043
	Anchor encoding	51.3	399.1	1.78	90,714
Hybrid	Bias	47.4	462.5	2.26	52,263
	Anchor encoding	48.2	462.6	1.66	96,165
	Off	47.4	462.5	1.59	100,259
SPERA-Small	Full (anchor encoding) + factorized (bias)	47.6	462.5	1.95	92,866

alone (0.6767 vs. 0.6525), and outperforms both REVE's spatial encoding substituted for our Legendre prior (0.6479) and a Gaussian bias on geodesic electrode distance (0.6398). The lower performance of the pairwise MLP suggests that the structured Legendre parameterization is beneficial under the present training setup. Further geometry variants are reported in Appendix F.1. We next examine the computational trade-offs that determine which Legendre implementation is used in each attention type.

Accuracy-efficiency trade-off Factorized attention reduces the matrix-multiplication cost substantially at our sequence length (N= 3,840), from 653 GFLOPs for full attention to 399 GFLOPs, despite having more parameters due to the additional axis-specific projections (50.6M vs. 37.9M). The hybrid architecture lies between the two at 462.5 GFLOPs, reflecting its combination of factorized blocks with periodic full-attention blocks. Here, matrix-multiplication FLOPs include linear projections and the attention matrix multiplications, with one multiply-accumulate counted as two FLOPs.

The Legendre implementation has little effect on the matrix-multiplication cost but strongly affects realized throughput. Under full attention, the pairwise Legendre bias requires an $N \times N$ score-bias matrix and prevents the fused attention path used in our implementation (i.e., FlashAttention via PyTorch SDPA), reducing throughput to 20,260 tok/s despite nearly identical matrix-multiplication FLOP counts for the bias- and anchor-based implementations (653.0 vs. 653.1 GFLOPs). The anchor encoding instead injects geometry features into the queries and keys while preserving the fused attention path, reaching 75,769 tok/s. In factorized attention, the smaller attention matrices make the

explicit pairwise bias practical, achieving 97,043 tok/s compared with 90,714 tok/s for the anchor encoding.

These observations motivate our final design: we use the anchor encoding in full-attention blocks to preserve throughput and the pairwise Legendre bias in factorized layers, where its computational cost is small. The resulting model achieves accuracy comparable to that of the bias-only hybrid configuration while providing 1.78× higher throughput (92,866 vs. 52,263 tok/s) and 14% lower peak memory (1.95 vs. 2.26 GB), at the cost of a 0.005 decrease in average balanced accuracy.

4.6 Additional analyses

We provide further analyses in the appendix, including controlled comparisons of the pretraining objective and corpus (Appendix E.1), temporal positional encoding and masking strategy (Appendices E.2 and E.3), analyses of the Legendre spherical prior (Appendix F), few-shot performance (Appendix G), noise and ocular-artifact robustness (Appendix H), and learned representation analysis (Appendix I).

5 Limitations and future work

While our scaling analysis covers four model sizes from 18M to 396M parameters, we did not investigate scaling at the billion-parameter scale due to limited computational resources. Furthermore, the diminishing returns observed beyond 194M suggest that further gains may require expanded data rather than parameters alone, but we did not systematically disentangle the contributions of corpus size, model capacity, and architectural priors to the observed scaling behavior. A systematic characterization of these factors remains an important direction for future investigation. Beyond scale, the spherical prior assumes surface electrodes; extending it to intracranial recordings would require incorporating radial information, as discussed in Appendix F.5.

We evaluate SPERA on nine diverse downstream tasks spanning clinical, cognitive, and BCI domains; however, the tasks are focused on classification problems over fixed-duration EEG segments ranging from 3 to 30 seconds. Further validation is therefore needed to assess generalization to continuous decoding, regression targets, and long-horizon clinical prediction. In addition, demographic metadata such as age, sex, ethnicity, and clinical status are not consistently available across our pretraining datasets, limiting a comprehensive fairness analysis of the learned representations across subpopulations. Addressing these gaps will be important before clinical or consumer deployment.

6 Conclusion

We present SPERA, a foundation model for EEG that learns representations through latent-space prediction. SPERA encodes scalp electrode geometry into attention through a Legendre-polynomial spatial prior, supported by a hybrid attention backbone and RSReg.

Across nine downstream tasks, SPERA-Base exceeds the strongest baseline (REVE-Base; Table 2) by 0.035 in average balanced accuracy. The largest significant gains are observed on fine-grained multi-class tasks, including imagined speech and clinical event classification. The scaling analysis further shows that SPERA-Base exceeds REVE-Large under LP despite using less than half the parameters. Controlled ablations show that latent prediction outperforms a matched reconstruction objective and that the Legendre prior outperforms the alternative geometry encodings we evaluated.

Overall, the results support latent-space prediction, geometry-aware spatial modeling, and frequency-aware supervision as complementary components for EEG representation learning across heterogeneous datasets and electrode montages. These findings provide a basis for future work on scalable and transferable EEG representation learning.

Acknowledgments and Disclosure of Funding

This work was supported by the Samsung Science & Technology Foundation Program (SRFC-IT2401-05). The authors declare no competing interests.

References

[1] Blanca Aguado-Lopez, Ana F. Palenciano, Jose M. G. Penalver, Paloma Diaz-Gutierrez, David Lopez-Garcia, Chiara Avancini, Luis F. Ciria, and Maria Ruz. Proactive selective attention across competition contexts. OpenNeuro dataset, 2024.

[2] 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*, 35(20):14681–14722, 2023.

[3] Edilberto Amorim, Wei-Long Zheng, Mohammad M Ghassemi, Mahsa Aghaeeaval, Pravinkumar Kandhare, Vishnu Karukonda, Jong Woo Lee, Susan T Herman, Adithya Sivaraju, Nicolas Gaspard, et al. The international cardiac arrest research consortium electroencephalography database. *Critical Care Medicine*, 51(12):1802–1811, 2023.

[4] Edilberto Amorim, Wei-Long Zheng, Jong Woo Lee, Susan Herman, Mohammad Ghassemi, Adithya Sivaraju, Nicolas Gaspard, Jeannette Hofmeijer, Michel J. A. M. van Putten, Matthew Reyna, Gari Clifford, and Brandon Westover. I-CARE: International cardiac arrest research consortium database. PhysioNet dataset, version 2.1, 2023.

[5] Pietro Aricò, Fabio Aloise, Francesca Schettini, Serenella Salinari, Donatella Mattia, and Febo Cincotti. Influence of P300 latency jitter on event related potential-based brain-computer interface performance. *Journal of Neural Engineering*, 11(3):035008, 2014.

[6] Bruno Aristimunha, Igor Carrara, Pierre Guetschel, Sara Sedlar, Pedro Rodrigues, Jan Sosulski, Divyesh Narayanan, Erik Bjareholt, Quentin Barthelemy, Robin Tibor Schirrmeister, Reinmar Kobler, Emmanuel Kalunga, Ludovic Darmet, Cattan Gregoire, Ali Abdul Hussain, Ramiro Gatti, Vladislav Goncharenko, Anton Andreev, Jordy Thielen, Davoud Hajhassani, Katelyn Begany, Thomas Moreau, Yannick Roy, Vinay Jayaram, Alexandre Barachant, and Sylvain Chevallier. Mother of all BCI benchmarks. Zenodo software, version 1.5.0, 2026.

[7] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[8] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15619–15629, 2023.

[9] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[10] Anahit Babayan, Miray Erbey, Deniz Kumral, Janis D Reinelt, Andrea MF Reiter, Josefin Röbbig, H Lina Schaare, Marie Uhlig, Alfred Anwander, Pierre-Louis Bazin, et al. A mind-brain-body dataset of MRI, EEG, cognition, emotion, and peripheral physiology in young and old adults. *Scientific Data*, 6(1):180308, 2019.

[11] Imad J. Bajwa, Andre S. Nilsen, René Skukies, Arnfinn Aamodt, Gernot Ernst, Johan F. Storm, and Bjørn E. Juel. A repeated awakening study exploring the capacity of complexity measures to capture dreaming during propofol sedation. OpenNeuro dataset, 2024.

[12] Alexandre Barachant. Alex motor imagery dataset. Zenodo dataset, 2012.

[13] Luis Alberto Barradas Chacón and Selina C. Wriessnegger. Toonfaces. OpenNeuro dataset, 2022.

[14] Matthew Barras and Liam Booth. Cognitive workload 8-level arithmetic. OpenNeuro dataset, 2026.

[15] Matthew Barras and Liam Booth. Multimodal cognitive workload n-back task, 4 difficulties. OpenNeuro dataset, 2026.

[16] Cemre Baykan and Alexander C. Schütz. Electroencephalographic responses to the number of objects in partially occluded and uncovered scenes. OpenNeuro dataset, 2024.

[17] Martha Ann Bell and Kimberly Cuevas. Using EEG to study cognitive development: Issues and practices. *Journal of cognition and development*, 13(3):281–294, 2012.

[18] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 813–824. PMLR, 2021.

[19] Nima Bigdely-Shamlo, Tim Mullen, Christian Kothe, Kyung-Min Su, and Kay A Robbins. The PREP pipeline: standardized preprocessing for large-scale EEG analysis. *Frontiers in neuroinformatics*, 9:16, 2015.

[20] Danaé Broustail, Anna Tegon, Thorir Mar Ingolfsson, Yawei Li, and Luca Benini. LuMamba: Latent unified Mamba for electrode topology-invariant and efficient EEG modeling. *arXiv preprint arXiv:2603.19100*, 2026.

[21] Claudio Carvalhaes and J Acacio De Barros. The surface Laplacian technique in EEG: Theory and methods. *International Journal of Psychophysiology*, 97(3):174–188, 2015.

[22] James F. Cavanagh and Trevor C. J. Jackson. Mood Manipulation and PST, Experiment 1. OpenNeuro dataset, 2022.

[23] 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.

[24] Xiaoli Chen, Rui Wang, Phyllis Zee, Pamela L Lutsey, Sogol Javaheri, Carmela Alcántara, Chandra L Jackson, Michelle A Williams, and Susan Redline. Racial/ethnic differences in sleep disturbances: the multi-ethnic study of atherosclerosis (MESA). *Sleep*, 38(6):877–888, 2015.

[25] Hohyun Cho, Minkyu Ahn, Sangtae Ahn, Moonyoung Kwon, and Sung Chan Jun. EEG datasets for motor imagery brain-computer interface. *GigaScience*, 6(7):gix034, 2017.

[26] Ian Daly, Nicoletta Nicolaou, Duncan Williams, Faustina Hwang, Alexis Kirke, Eduardo Miranda, and Slawomir J. Nasuto. An EEG dataset recorded during affective music listening. OpenNeuro dataset, 2020.

[27] Arnaud Delorme. Go-nogo categorization and detection task. OpenNeuro dataset, 2021.

[28] Arnaud Delorme and Claire Braboszcz. Meditation vs thinking task. OpenNeuro dataset, 2021.

[29] Paolo Detti. Siena scalp EEG database. PhysioNet dataset, version 1.0.0, 2020.

[30] 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.

[31] Giovanni M. Di Liberto, Michael P. Broderick, Ole Bialas, and Edmund C. Lalor. EEG responses to continuous naturalistic speech. OpenNeuro dataset, 2023.

[32] Guido Dornhege, Benjamin Blankertz, Gabriel Curio, and Klaus-Robert Müller. Boosting bit rates in non-invasive EEG single-trial classifications by feature combination and multi-class paradigms. *IEEE Transactions on Biomedical Engineering*, 51(6):993–1002, 2004.

[33] Pauline Dreyer, Aline Roc, Léa Pillette, Sébastien Rimbert, and Fabien Lotte. A large EEG database with users' profile information for motor imagery brain-computer interface research. *Scientific Data*, 10(1):580, 2023.

[34] Josef Faller, Carmen Vidaurre, Teodoro Solis-Escalante, Christa Neuper, and Reinhold Scherer. Autocalibration and recurrent adaptation: Towards a plug and play online ERD-BCI. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 20(3):313–319, 2012.

[35] Lukas Gehrke, Sezen Akman, Albert Chen, Pedro Lopes, and Klaus Gramann. Prediction error. OpenNeuro dataset, 2021.

[36] Irini Giannopulu and Haruo Mizutani. Neural kinesthetic contribution to motor imagery of body parts: Tongue, hands, and feet. *Frontiers in Human Neuroscience*, 15:602723, 2021.

[37] Alessandro T Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M Cichy. A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022.

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

[39] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Lauri Parkkonen, and Matti S Hämäläinen. MNE software for processing MEG and EEG data. *NeuroImage*, 86:446–460, 2014.

[40] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[41] Tijl Grootswagers, Amanda Robinson, Sofia Shatek, and Thomas Carlson. Features-EEG. OpenNeuro dataset, 2022.

[42] Tijl Grootswagers, Amanda Robinson, Sofia Shatek, and Thomas Carlson. EEG-attention-rsvp-exp1. OpenNeuro dataset, 2023.

[43] Tijl Grootswagers, Amanda Robinson, Sofia Shatek, and Thomas Carlson. EEG-attention-rsvp-exp2. OpenNeuro dataset, 2023.

[44] Tijl Grootswagers, Ivy Zhou, Amanda Robinson, Martin Hebart, and Thomas Carlson. Human electroencephalography recordings from 50 subjects for 22,248 images from 1,854 object concepts. OpenNeuro dataset, 2021.

[45] Pierre Guetschel, Thomas Moreau, and Michael Tangermann. S-JEPA: Towards seamless cross-dataset transfer through dynamic spatial attention. *arXiv preprint arXiv:2403.11772*, 2024.

[46] Cameron D. Hassall, Laurence T. Hunt, and Clay B. Holroyd. Average Task Value. OpenNeuro dataset, 2022.

[47] Cameron D. Hassall, Yan Yan, and Laurence T. Hunt. Drum trainer. OpenNeuro dataset, 2022.

[48] Cameron D. Hassall, Yan Yan, and Laurence T. Hunt. Steer the Ship. OpenNeuro dataset, 2022.

[49] Marleen Haupt, Monika Graumann, Santani Teng, Carina Kaltenbach, and Radoslaw M. Cichy. Braille letters - EEG. OpenNeuro dataset, 2024.

[50] Jason Helbing, Dejan Draschkow, and Melissa L.-H. Võ. Search superiority recollection familiarity. OpenNeuro dataset, 2024.

[51] Verónica Henao Isaza, Valeria Cadavid Castro, Luisa María Zapata Saldarriaga, Yorguin-Jose Mantilla-Ramos, Jazmín Ximena Suarez Revelo, Carlos Andrés Tobón Quintero, and John Fredy Ochoa Gómez. Longitudinal EEG test-retest reliability in healthy individuals. OpenNeuro dataset, 2026.

[52] Ulrich Hoffmann, Jean-Marc Vesin, Touradj Ebrahimi, and Karin Diserens. An efficient P300-based brain-computer interface for disabled subjects. *Journal of Neuroscience Methods*, 167(1):115–125, 2008.

[53] Amirabbas Hojjati, Lu Li, Ibrahim Hameed, Anis Yazidi, Pedro G Lind, and Rabindra Khadka. From video to EEG: Adapting joint embedding predictive architecture to uncover visual concepts in brain signal analysis. *arXiv preprint arXiv:2507.03633*, 2025.

[54] Elif Isbell, Amanda N. Peters, Dylan M. Richardson, and Nancy E. R. De León. Cognitive electrophysiology in socioeconomic context in adulthood: An EEG dataset. OpenNeuro dataset, 2025.

[55] Simin Jamshidi, Arturo Espinoza, Soura Dasgupta, and Nandakumar Narayanan. EEG Mortality Dataset in Parkinson's Disease. OpenNeuro dataset, 2025.

[56] 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.

[57] Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. *arXiv preprint arXiv:2405.18765*, 2024.

[58] Jin Jing, Wendong Ge, Shenda Hong, Marta Bento Fernandes, Zhen Lin, Chaoqi Yang, Sungtae An, Aaron F Struck, Aline Herlopian, Ioannis Karakis, et al. Development of expert-level classification of seizures and rhythmic and periodic patterns during EEG interpretation. *Neurology*, 100(17):e1750–e1762, 2023.

[59] Jin Jing, Zhen Lin, Chaoqi Yang, Ashley Chow, Sohier Dane, Jimeng Sun, and M. Brandon Westover. HMS - harmful brain activity classification, 2024. Kaggle.

[60] Michael J. Kahana, Joseph H. Rudoler, Lynn J. Lohnas, Karl Healey, Ada Aka, Adam Broitman, Elizabeth Crutchley, Patrick Crutchley, Kylie H. Alm, Brandon S. Katerman, Nicole E. Miller, Joel R. Kuhn, Yuxuan Li, Nicole M. Long, Jonathan Miller, Madison D. Paron, Jesse K. Pazdera, Isaac Pedisich, and Christoph T. Weidemann. Penn electrophysiology of encoding and retrieval study (peers). OpenNeuro dataset, 2023.

[61] Emmanuel K. Kalunga, Sylvain Chevallier, Quentin Barthélemy, Karim Djouani, Eric Monacelli, and Yskandar Hamam. Online SSVEP-based BCI using Riemannian geometry. *Neurocomputing*, 191:55–68, 2016.

[62] Zoltan Kekecs, Kyra Girán, Vanda Vizkievicz, Anna Lutoskin, and Yeganeh Farahzadi. The effects of sham hypnosis techniques. OpenNeuro dataset, 2023.

[63] 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.

[64] Hassan Aqeel Khan, Rahat Ul Ain, Awais Mehmood Kamboh, Hammad Tanveer Butt, Saima Shafait, Wasim Alamgir, Didier Stricker, and Faisal Shafait. The NMT scalp EEG dataset: An open-source annotated dataset of healthy and pathological EEG recordings for predictive modeling. *Frontiers in neuroscience*, 15:755817, 2022.

[65] Andy Jeesu Kim, Mara Mather, Santiago Morales, and Joshua Senior. Young adult resting state and auditory oddball task EEG data. OpenNeuro dataset, 2025.

[66] Louis Korczowski, Martine Cederhout, Anton Andreev, Grégoire Cattan, Pedro Luiz Coelho Rodrigues, Violette Gautheret, and Marco Congedo. Brain invaders calibration-less P300-based BCI with modulation of flash duration dataset (bi2015a). Zenodo dataset, 2019.

[67] Louis Korczowski, Martine Cederhout, Anton Andreev, Grégoire Cattan, Pedro Luiz Coelho Rodrigues, Violette Gautheret, and Marco Congedo. Brain Invaders Cooperative versus Competitive: Multi-User P300-based Brain-Computer Interface Dataset (BI2015b). Zenodo dataset, 2019.

[68] Louis Korczowski, Ekaterina Ostaschenko, Anton Andreev, Grégoire Cattan, Pedro Luiz Coelho Rodrigues, Violette Gautheret, and Marco Congedo. Brain invaders calibration-less P300-based BCI using dry EEG electrodes dataset (bi2014a). Zenodo dataset, 2019.

[69] Louis Korczowski, Ekaterina Ostaschenko, Anton Andreev, Grégoire Cattan, Pedro Luiz Coelho Rodrigues, Violette Gautheret, and Marco Congedo. Brain invaders solo versus collaboration: Multi-user P300-based brain-computer interface dataset (bi2014b). Zenodo dataset, 2019.

[70] 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.

[71] Min-Ho Lee, O-Yeon Kwon, Yong-Jeong Kim, Hong-Kyung Kim, Young-Eun Lee, John Williamson, Siamac Fazli, and Seong-Whan Lee. EEG dataset and OpenBMI toolbox for three BCI paradigms: An investigation into BCI illiteracy. *GigaScience*, 8(5):giz002, 2019.

[72] Robert Leeb, Felix Lee, Claudia Keinrath, Reinhold Scherer, Horst Bischof, and Gert Pfurtscheller. Brain–computer communication: motivation, aim, and impact of exploring a virtual apartment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 15(4):473–482, 2007.

[73] 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.

[74] Weilong Li and Hua Fan. Perceiveimagine. OpenNeuro dataset, 2024.

[75] Zi-Yi Li, Wei-Long Zheng, Jiwen Xu, Yong Lu, and Bao-Liang Lu. Gram: A large general EEG model for raw data classification and restoration. *IEEE Transactions on Affective Computing*, 2025.

[76] Otavio G Lins, Terence W Picton, Patrick Berg, and Michael Scherg. Ocular artifacts in EEG and event-related potentials I: Scalp topography. *Brain topography*, 6(1):51–63, 1993.

[77] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2019.

[78] Benjamin Lowe, Jonathan Robinson, Naohide Yamamoto, Hinze Hogendoorn, and Patrick Johnston. Visual attribute-specific contextual trajectory paradigm. OpenNeuro dataset, 2023.

[79] Jingying Ma, Feng Wu, Qika Lin, Yucheng Xing, Chenyu Liu, Ziyu Jia, and Mengling Feng. CodeBrain: Bridging decoupled tokenizer and multi-scale architecture for EEG foundation model. In *International Conference on Learning Representations*, volume 2026, pages 83526–83574, 2026.

[80] Szymon Mąka, Marta Chrustowicz, and Łukasz Okruszek. Discrepancy between self-report and neurophysiological markers of socio-affective responses in lonely individuals. OpenNeuro dataset, 2025.

[81] Szymon Mąka, Marta Chrustowicz, and Łukasz Okruszek. Targeted neuromodulation of the left dorsolateral prefrontal cortex alleviates altered affective response evaluation in lonely individuals. OpenNeuro dataset, 2025.

[82] Dominique Makowski, An-Shu Te, Stephanie Kirk, and Zi Liang Ngoi. Fakefaceemo_data. OpenNeuro dataset, 2023.

[83] Christopher J Markiewicz, Krzysztof J Gorgolewski, Franklin Feingold, Ross Blair, Yaroslav O Halchenko, Eric Miller, Nell Hardcastle, Joe Wexler, Oscar Esteban, Mathias Goncavles, et al. The OpenNeuro resource for sharing of neuroscience data. *eLife*, 10:e71774, 2021.

[84] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[85] Donia Metwalli, Eslam Ahmed, Antony Emil, Yousef A. Radwan, Mariam Barakat, Anas Ahmed, Amro Omar, and Sahar Selim. ArEEG: Arabic inner speech EEG dataset. OpenNeuro dataset, 2024.

[86] Denise Moerel, Tijl Grootswagers, Amanda Robinson, Sophia Shatek, Alexandra Woolgar, Thomas Carlson, and Anina Rich. The time-course of feature-based attention effects dissociated from temporal expectation and target-related processes. OpenNeuro dataset, 2022.

[87] Navid Mohammadi Foumani, Geoffrey Mackellar, Soheila Ghane, Saad Irtza, Nam Nguyen, and Mahsa Salehi. EEG2Rep: enhancing self-supervised EEG representation through informative masked inputs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5544–5555, 2024.

[88] Shayan Motamedi-Fakhr, Mohamed Moshrefi-Torbati, Martyn Hill, Catherine M Hill, and Paul R White. Signal processing techniques applied to human sleep EEG signals—a review. *Biomedical Signal Processing and Control*, 10:21–33, 2014.

[89] Wajid Mumtaz. MDD patients and healthy controls EEG data (new). figshare dataset, 11 2016.

[90] Christa Neuper, Michael Wörtz, and Gert Pfurtscheller. ERD/ERS patterns reflecting sensorimotor activation and deactivation. In Christa Neuper and Wolfgang Klimesch, editors, *Event-Related Dynamics of Brain Oscillations*, volume 159 of *Progress in Brain Research*, pages 211–222. Elsevier, 2006.

[91] Iyad Obeid and Joseph Picone. The temple university hospital EEG data corpus. *Frontiers in Neuroscience*, 10:196, 2016.

[92] Patrick Ofner, Andreas Schwarz, Joana Pereira, and Gernot R. Müller-Putz. Upper limb movements can be decoded from the time-domain of low-frequency EEG. *PLoS ONE*, 12(8):e0182578, 2017.

[93] Julie Onton and Scott Makeig. Imagined Emotion Study. OpenNeuro dataset, 2022.

[94] 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.

[95] Saarang Panchavati, Uddhav Panchavati, Corey Arnold, and William Speier. Laya: A LeJEPA approach to EEG via latent prediction over reconstruction. *arXiv preprint arXiv:2603.16281*, 2026.

[96] Tasos Papastylianou, Rodrigo Ramele, Luca Citi, Caterina Cinel, and Riccardo Poli. Pes - pandemic emergency scenario. OpenNeuro dataset, 2023.

[97] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976, 2019.

[98] Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *Proceedings of the European conference on computer vision (ECCV)*, pages 268–284, 2018.

[99] Yuri G. Pavlov. Semantic_conditioning. OpenNeuro dataset, 2024.

[100] 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)*, pages 3599–3602. IEEE, 2022.

[101] Wilder Penfield and Edwin Boldrey. Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain: A journal of neurology*, 60(4):389–443, 1937.

[102] François Perrin, Jacques Pernier, Olivier Bertrand, and Jean Francois Echallier. Spherical splines for scalp potential and current density mapping. *Electroencephalography and clinical neurophysiology*, 72(2):184–187, 1989.

[103] Mathias Perslev, Sune Darkner, Lykke Kempfner, Miki Nikolic, Poul Jørgen Jennum, and Christian Igel. U-Sleep: resilient high-frequency sleep staging. *NPJ digital medicine*, 4(1):72, 2021.

[104] G. Pfurtscheller, C. Brunner, A. Schlögl, and F. H. Lopes da Silva. Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks. *NeuroImage*, 31(1):153–159, 2006.

[105] G. Pfurtscheller and F.H. Lopes da Silva. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology*, 110(11):1842–1857, 1999.

[106] G. Pfurtscheller and C. Neuper. Motor imagery and direct brain-computer communication. *Proceedings of the IEEE*, 89(7):1123–1134, 2001.

[107] Huy Phan and Kaare Mikkelsen. Automatic sleep staging of EEG signals: recent development, challenges, and future directions. *Physiological measurement*, 43(4):04TR01, 2022.

[108] Aura Polo, Elmer León, Mariana Pino-Melgarejo, and Julie Viloria-Porto. Dataset of electroencephalograms of juvenile offenders. OpenNeuro dataset, 2025.

[109] Jathurshan Pradeepkumar, Xihao Piao, Zheng Chen, and Jimeng Sun. Tokenizing single-channel EEG with time-frequency motif learning. *arXiv preprint arXiv:2502.16060*, 2025.

[110] Jia Hong Puah, Sim Kuan Goh, Ziwei Zhang, Zixuan Ye, Chow Khuen Chan, Kheng Seang Lim, Si Lei Fong, Kok Sin Woon, and Cuntai Guan. EEGDM: EEG representation learning via generative diffusion model. *arXiv preprint arXiv:2508.14086*, 2025.

[111] M. Ramne, S. Damercheli, F. Apelgren, I. Pettersson, and E. Lendaro. High density resting state EEG of phantom limb pain and controls. OpenNeuro dataset, 2025.

[112] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[113] Petunia Reinke, Lisa Deneke, and Sebastian Ocklenburg. EEG-asymmetries dataset. OpenNeuro dataset, 2026.

[114] Matthew A. Reyna, Audrey Weigle, Qichen Li, Zuzana Koscova, Haoqi Sun, Shenghan Wen, Umakanth Katwa, Robert Thomas, Dennis Hwang, Lynn Marie Trotti, Emmanuel Mignot, Reza Sameni, Samaneh Nasiri, M. Brandon Westover, and Gari D. Clifford. Screening for Cognitive Impairment During Sleep Studies: The George B. Moody PhysioNet Challenge 2026. In *Computing in Cardiology 2026*, volume 53, pages 1–4, 2026.

[115] Maria J. Ribeiro and Miguel Castelo-Branco. EEG, ECG and pupil data from young and older adults: rest and auditory cued reaction time tasks. OpenNeuro dataset, 2021.

[116] Angela Riccio, Luca Simione, Francesca Schettini, Alessia Pizzimenti, Maurizio Inghilleri, Marta Olivetti Belardinelli, Donatella Mattia, and Febo Cincotti. Attention and P300-based BCI performance in people with amyotrophic lateral sclerosis. *Frontiers in Human Neuroscience*, 7:732, 2013.

[117] Alexander P. Rockhill, Nicko Jackson, Jobi George, Adam Aron, and Nicole C. Swann. UC San Diego Resting State EEG Data from Patients with Parkinson's Disease. OpenNeuro dataset, 2021.

[118] Pablo Rodríguez-San Esteban, Ana B. Chica, and José A. González-López. Neural representation of consciously seen and unseen information. OpenNeuro dataset, 2024.

[119] Joseph H. Rudoler, Matthew R. Dougherty, Brandon S. Katerman, James P. Bruska, Woohyeuk Chang, David J. Halpern, Nicholas B. Diamond, and Michael J. Kahana. Spatial memory and non-invasive closed-loop stimulus timing. OpenNeuro dataset, 2023.

[120] Anushri Saha, Vikash Minz, Sanjith Bonela, SR Sreeja, Ritwika Chowdhury, and Debasis Samanta. Classification of EEG signals for cognitive load estimation using deep learning architectures. In *International Conference on Intelligent Human Computer Interaction*, pages 59–68. Springer, 2018.

[121] 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.

[122] 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.

[123] Reinhold Scherer, Josef Faller, Elisabeth V. C. Friedrich, Eloy Opisso, Ursula Costa, Andrea Kübler, and Gernot R. Müller-Putz. Individually adapted imagery improves brain-computer interface performance in end-users with disability. *PLoS ONE*, 10(5):e0123727, 2015.

[124] 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.

[125] Martijn Schreuder, Thomas Rost, and Michael Tangermann. Listen, you are writing! Speeding up online spelling with a dynamic auditory BCI. *Frontiers in Neuroscience*, 5:112, 2011.

[126] Tong Shan, Madeline S. Cappelloni, and Ross K. Maddox. Subcortical responses to music and speech are alike while cortical responses diverge. OpenNeuro dataset, 2022.

[127] Sophia M. Shatek, Amanda K. Robinson, Tijl Grootswagers, and Thomas A. Carlson. Capacity for movement is an organisational principle in object representations: EEG data from Experiment 1. OpenNeuro dataset, 2021.

[128] Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

[129] Jaeyoung Shin, Alexander von Lühmann, Benjamin Blankertz, Do-Won Kim, Jichai Jeong, Han-Jeong Hwang, and Klaus-Robert Müller. Open access dataset for EEG+NIRS single-trial classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1735–1745, 2017.

[130] Seyed Yahya Shirazi, Alexandre Franco, Maurício Scopel Hoffmann, Nathalia B. Esper, Dung Truong, Arnaud Delorme, Michael Milham, and Scott Makeig. HBN-EEG: The FAIR implementation of the healthy brain network (HBN) electroencephalography dataset. *bioRxiv*, 2024.

[131] Elizabeth M. Siefert, Sindhuja Uppuluri, Jianing Mu, Marlie C. Tandoc, James W. Antony, and Anna C. Schapiro. Siefert2024. OpenNeuro dataset, 2024.

[132] Sophie Smit, Almudena Ramírez-Haro, Manuel Varlet, Denise Moerel, Genevieve L. Quek, and Tijl Grootswagers. A comprehensive EEG dataset for investigating visual touch perception. OpenNeuro dataset, 2024.

[133] Shelagh JM Smith. EEG in the diagnosis, classification, and management of patients with epilepsy. *Journal of Neurology, Neurosurgery & Psychiatry*, 76(suppl 2):ii2–ii7, 2005.

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

[135] Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG Conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023.

[136] Jan Sosulski and Michael Tangermann. Electroencephalogram signals recorded from 13 healthy subjects during an auditory oddball paradigm under different stimulus onset asynchrony conditions. University of Freiburg dataset, 2019.

[137] James R. Stieger, Stephen A. Engel, and Bin He. Continuous sensorimotor rhythm based brain computer interface learning in a large population. *Scientific Data*, 8:98, 2021.

[138] Thomas J. Stoll, Nathan D. Vandjelovic, Melissa J. Polonenko, Nadja R. S. Li, Adrian K. C. Lee, and Ross K. Maddox. The auditory brainstem response to natural speech is not affected by selective attention. OpenNeuro dataset, 2025.

[139] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[140] Haoqi Sun, Wolfgang Ganglberger, Samaneh Nasiri, Aditya Gupta, Manohar Ghanta, Valdery Moura Junior, Sydney Cash, Katie Stone, Zhiyong Zhang, Gauri Ganjoo, Thijs E. Nassi, Ruoqi Wei, Erik-Jan Meulenbrugge, Rhoda Au, Gari Clifford, Lynn Marie Trotti, Dennis Hwang, Emmanuel Mignot, Umakanth Katwa, and M. Brandon Westover. The Human Sleep Project (version 2.0). Brain Data Science Platform, 2023.

[141] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 06–11 Aug 2017.

[142] 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.

[143] Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J Miller, Gernot R Müller-Putz, et al. Review of the BCI competition IV. *Frontiers in Neuroscience*, 6:55, 2012.

[144] Jack E. Taylor, Rasmus Sinn, Cosimo Iaia, and Christian J. Fiebach. Alphabetic decision task (arial light font). OpenNeuro dataset, 2024.

[145] Hanneke van Dijk, Guido A. van Wingen, Damiaan Denys, Sebastian Olbrich, Rosalinde van Ruth, and Martijn Arns. The two decades brainclinics research archive for insights in neurophysiology (TDBRAIN) database. *Scientific Data*, 9:333, 2022.

[146] Kaido Värbu, Naveed Muhammad, and Yar Muhammad. Past, present, and future of EEG-based BCI applications. *Sensors*, 22(9):3331, 2022.

[147] J. Veillette, S. Heald, B. Wittenbrink, and H. Nusbaum. EEG-neuroforecasting. OpenNeuro dataset, 2022.

[148] John Veillette, Pedro Lopes, and Howard Nusbaum. Illusion of agency over electrically-actuated movements. OpenNeuro dataset, 2023.

[149] Guagnyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. *Advances in Neural Information Processing Systems*, 37:39249–39280, 2024.

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

[151] Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Shijian Li, and Gang Pan. EEGMamba: An EEG foundation model with mamba. *Neural Networks*, 192:107816, 2025.

[152] Yulin Wang, Wei Duan, Lihong Ding, Debo Dong, and Xu Lei. A test-retest resting and cognitive state EEG dataset. OpenNeuro dataset, 2022.

[153] Xiaoxi Wei, A. Aldo Faisal, Moritz Grosse-Wentrup, Alexandre Gramfort, Sylvain Chevallier, Vinay Jayaram, Camille Jeunet, Stylianos Bakas, Siegfried Ludwig, Konstantinos Barmpas, Mehdi Bahri, Yannis Panagakis, Nikolaos Laskaris, Dimitrios A. Adamos, Stefanos Zafeiriou, William C. Duong, Stephen M. Gordon, Vernon J. Lawhern, Maciej Śliwowski, Vincent Rouanne, and Piotr Tempczyk. 2021 BEETL competition: Advancing transfer learning for subject independence & heterogenous EEG data sets. In *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, 2022.

[154] Chuqin Xiang, Xinrui Fan, Duo Bai, Ke Lv, and Xu Lei. A Resting-state EEG Dataset for Sleep Deprivation. OpenNeuro dataset, 2024.

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

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

[157] Weibo Yi, Shuang Qiu, Kun Wang, Hongzhi Qi, Lixin Zhang, Peng Zhou, Feng He, and Dong Ming. Evaluation of EEG oscillatory patterns and cognitive process during simple and compound limb motor imagery. *PLoS ONE*, 9(12):e114853, 2014.

[158] Carolina Zaehme, Isabelle Sander, and Klaus Gramann. Urban appraisal: Physiological recording during rating of different urban environments. OpenNeuro dataset, 2025.

[159] Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The national sleep research resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, 25(10):1351–1358, 2018.

[160] Haoming Zhang, Mingqi Zhao, Chen Wei, Dante Mantini, Zherui Li, and Quanying Liu. EEGdenoiseNet: a benchmark dataset for deep learning solutions of EEG denoising. *Journal of Neural Engineering*, 18(5):056057, oct 2021.

[161] Bangyan Zhou, Xiaopei Wu, Zhao Lv, Lei Zhang, and Xiaojin Guo. A fully automated trial selection method for optimization of motor imagery based brain-computer interface. *PLoS ONE*, 11(9):e0162657, 2016.

[162] 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. *Advances in Neural Information Processing Systems*, 38:96859–96904, 2025.

[163] Natalia Zhozhikashvili, Maria Protopova, Tatiana Shkurenko, Marie Arsalidou, Ilya Zakharov, Boris Kotchoubey, Sergey Malykh, and Yuri Pavlov. Sternberg difficult. OpenNeuro dataset, 2024.

[164] 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 Computing resources

To ensure reproducibility, we report the detailed computing environment, as follows:

- **Hardware:** HP Z8 Fury G5 Workstation equipped with an Intel Xeon w5-3425 CPU, 128 GB of RAM, and dual NVIDIA RTX 5000 Ada Generation GPUs.
- **Software & environment:** The environment is set up on Ubuntu 24.04.4 LTS and Python 3.10.20, supported by CUDA Toolkit 13.0 and cuDNN 9.15 for hardware acceleration. Model implementation and training are built upon PyTorch 2.10.0 and the Hugging Face ecosystem, utilizing `transformers` 5.3.0 and `accelerate` 1.13.0. Data preprocessing is performed using MNE-Python 1.11.0.
- **Training efficiency & cost:** To maximize training efficiency, we employed distributed data parallel (DDP) across both GPUs and utilized Bfloat16 (BF16) mixed-precision training. The total computational cost for pretraining our largest model is approximately 287 GPU hours.
- **Code & model availability:** Our codebase is provided through the repository linked in the paper, including training, evaluation, and preprocessing scripts.

B Pretraining corpus

Datasets are collected from multiple public repositories, including OpenNeuro [83], MOABB [6], PhysioNet [38], NSRR [159], TUH [91], Brainclinics (`https://brainclinics.com/resources`; see also [145]), and the Max Planck Institute [10]. Table 5 reports subject counts adjusted for known cross-dataset overlaps, and Table 6 lists per-dataset counts. To avoid double-counting, known subject-level overlaps are removed (e.g., overlaps among the Shin2017 series [129], Lee2019 series [71], and OpenNeuro datasets ds007169 [15]/ds007262 [14]). In addition, to prevent any evaluation leakage, subjects reserved for downstream tasks (TUAB and TUEV) are excluded from the TUH corpus. For HSP [140], we use the small training subset released for the PhysioNet Challenge 2026 [114]. Unless otherwise noted, reported hours reflect the pretraining corpus after these exclusions.

Table 5: Summary statistics of the pretraining corpus (aggregated from dataset-level metadata).

Grouping	# Subjects	Hours (h)	# Datasets
By category			
Clinical	21,916	65,387	11
Cognitive	6,248	12,915	55
BCI	884	1,362	40
By channel count			
3–32	22,728	66,051	40
33–64	1,963	2,114	42
> 64	4,357	11,499	24
By segment length			
10 s	3,414	1,967	46
30 s	4,236	11,378	19
60 s	21,398	66,319	41
Total	29,048	79,664	106

Table 6: Datasets used in the pretraining corpus.

Dataset	Repository	# Subjects	Hours (h)	Channels	License
Clinical					
TUH [91]	TUH	12,299	25,785	Mostly 30–41	NEDC data use agreement

Continued on next page

Dataset	Repository	# Subjects	Hours (h)	Channels	License
TDBRAIN [145]	Brainclinics	1,274	81.67	26	Brainclinics data use agreement
MESA [24]	NSRR	2,056	10,568	3	NSRR data use agreement
Siena [29, 30]	PhysioNet	14	141	29	CC BY 4.0
I-CARE [4, 3]	PhysioNet	607	22,278	19	CC BY-NC-SA 4.0
HSP (subset) [140, 114]	Kaggle	1,103	5,799	6	BDSP license 1.5.0
HMS [59]	Kaggle	1,950	237	8	CC BY-NC 4.0
NMT [64]	Zenodo	2,417	474	19	CC BY 4.0
ds002778 [117]	OpenNeuro	31	2.33	32	CC0
ds004902 [154]	OpenNeuro	71	17.62	61	CC0
ds007020 [55]	OpenNeuro	94	3.60	63	CC0
Cognitive					
MPI-Lemon [10]	Max Planck Institute	215	48.35	59	PDDL
THINGS2 [37]	FigShare	10	79.25	64	CC BY 4.0
ds002680 [27]	OpenNeuro	14	17.85	19	CC0
ds002721 [26]	OpenNeuro	31	25.17	19	CC0
ds003004 [93]	OpenNeuro	34	47.25	134	CC0
ds003690 [115]	OpenNeuro	75	45.13	59	CC0
ds003825 [44]	OpenNeuro	50	45.75	63	CC0
ds003885 [127]	OpenNeuro	24	26.80	128	CC0
ds003969 [28]	OpenNeuro	98	65.65	64	CC0
ds004043 [86]	OpenNeuro	20	18.03	63	CC0
ds004147 [46]	OpenNeuro	12	9.50	31	CC0
ds004148 [152]	OpenNeuro	60	60.00	64	CC0
ds004264 [48]	OpenNeuro	21	7.27	31	CC0
ds004315 [22]	OpenNeuro	50	21.04	60	CC0
ds004356 [126]	OpenNeuro	22	46.03	34	CC0
ds004395 [60]	OpenNeuro	364	8,687	128	CC0
ds004408 [31]	OpenNeuro	19	18.68	128	CC0
ds004477 [96]	OpenNeuro	9	13.48	79	CC0
ds004561 [148]	OpenNeuro	23	11.35	61	CC0
ds004572 [62]	OpenNeuro	52	47.37	58	CC0
ds004582 [82]	OpenNeuro	73	33.65	59	CC0
ds004603 [78]	OpenNeuro	37	30.61	64	CC0
ds004706 [119]	OpenNeuro	34	468	128	CC0
ds004816 [42]	OpenNeuro	20	10.50	63	CC0
ds004951 [49]	OpenNeuro	11	25.75	63	CC0
ds005095 [163]	OpenNeuro	48	16.48	63	CC0
ds005121 [131]	OpenNeuro	34	40.17	64	CC0
ds005189 [50]	OpenNeuro	30	19.28	61	CC0
ds005273 [118]	OpenNeuro	33	51.17	63	CC0
ds005410 [99]	OpenNeuro	81	22.94	63	CC0
ds005505 [130]	OpenNeuro	136	105	128	CC BY-SA 4.0
ds005506 [130]	OpenNeuro	150	115	128	CC BY-SA 4.0
ds005507 [130]	OpenNeuro	184	142	128	CC BY-SA 4.0
ds005508 [130]	OpenNeuro	324	137	128	CC BY-SA 4.0
ds005509 [130]	OpenNeuro	330	251	128	CC BY-SA 4.0
ds005510 [130]	OpenNeuro	134	92.20	128	CC BY-SA 4.0
ds005511 [130]	OpenNeuro	381	250	128	CC BY-SA 4.0
ds005512 [130]	OpenNeuro	257	164	128	CC BY-SA 4.0
ds005514 [130]	OpenNeuro	295	183	128	CC BY-SA 4.0
ds005515 [130]	OpenNeuro	533	148	128	CC BY-SA 4.0
ds005516 [130]	OpenNeuro	430	245	128	CC BY-SA 4.0
ds005594 [144]	OpenNeuro	16	12.17	64	CC0
ds005620 [11]	OpenNeuro	21	16.25	62	CC0
ds005662 [132]	OpenNeuro	80	79.85	64	CC0
ds006018 [54]	OpenNeuro	127	50.39	26	CC0
ds006434 [138]	OpenNeuro	66	120	32	CC0
ds006480 [65]	OpenNeuro	68	69.98	65	CC0

Continued on next page

Dataset	Repository	# Subjects	Hours (h)	Channels	License
ds006850 [158]	OpenNeuro	63	78.30	64	CC0
ds006861 [81]	OpenNeuro	120	100	32	CC0
ds006866 [80]	OpenNeuro	148	122	62	CC0
ds006921 [111]	OpenNeuro	38	16.01	63	CC0
ds006923 [108]	OpenNeuro	140	33.85	128	CC0
ds007172 [113]	OpenNeuro	100	47.89	32	CC0
ds007176 [51]	OpenNeuro	45	25.08	58	CC0
HBN_NC [130]	Child Mind Institute	458	251	128	CC BY-NC-SA 4.0
BCI					
ds003846 [35]	OpenNeuro	19	22.38	63	CC0
ds004152 [47]	OpenNeuro	21	11.27	31	CC0
ds004284 [147]	OpenNeuro	18	9.43	128	CC0
ds004324 [13]	OpenNeuro	26	19.18	28	CC0
ds004357 [41]	OpenNeuro	16	11.28	128	CC0
ds004817 [43]	OpenNeuro	20	11.88	63	CC0
ds005089 [1]	OpenNeuro	36	62.76	63	CC0
ds005262 [85]	OpenNeuro	12	24.62	8	CC0
ds005586 [16]	OpenNeuro	23	33.49	60	CC0
ds005697 [74]	OpenNeuro	51	70.67	64	CC0
ds007169 [15]	OpenNeuro	18	4.90	19	CC0
ds007262 [14]	OpenNeuro	18	4.43	19	CC0
AlexMI [12]	MOABB	8	1.09	16	CC BY-SA 4.0
Beetl2021_A [153]	MOABB	3	1.00	63	CC BY 4.0
BI2014a [68]	MOABB	64	12.32	16	CC BY 4.0
BI2014b [69]	MOABB	38	2.31	32	CC BY 4.0
BI2015a [66]	MOABB	43	11.48	32	CC BY 4.0
BI2015b [67]	MOABB	44	25.87	32	CC BY 4.0
BNCI2003_004 [32]	MOABB	5	3.97	118	CC BY 4.0
BNCI2014_004 [72]	MOABB	9	26.19	3	CC BY-ND 4.0
BNCI2014_008 [116]	MOABB	8	3.11	8	CC BY-NC-ND 4.0
BNCI2014_009 [5]	MOABB	10	1.58	16	CC BY-NC-ND 4.0
BNCI2015_001 [34]	MOABB	12	16.66	13	CC BY-NC-ND 4.0
BNCI2015_003 [125]	MOABB	10	0.91	8	CC BY-NC-ND 4.0
BNCI2015_004 [123]	MOABB	9	13.72	30	CC BY-NC-ND 4.0
Cho2017 [25]	MOABB	52	20.46	64	CC BY 4.0
Dreyer2023 [33]	MOABB	87	68.12	27	CC BY 4.0
EPFLP300 [52]	MOABB	8	2.79	32	Not specified
Kalunga2016 [61]	MOABB	12	2.56	8	CC BY 4.0
Lee2019ERP [71]	MOABB	54	57.85	62	GPL-3.0
Lee2019MI [71]	MOABB	54	42.81	62	GPL-3.0
Lee2019SSVEP [71]	MOABB	54	40.45	62	GPL-3.0
Ofner2017 [92]	MOABB	15	13.33	61	CC BY 4.0
Schirrmeister2017 [124]	MOABB	14	26.44	128	CC BY 4.0
Shin2017A [129]	MOABB	29	14.41	30	GPL-3.0
Shin2017B [129]	MOABB	29	14.43	30	GPL-3.0
Sosulski2019 [136]	MOABB	13	18.30	31	CC BY-SA 4.0
Stieger2021 [137]	MOABB	62	615	60	CC BY-NC 4.0
Weibo2014 [157]	MOABB	10	12.44	60	CC0
Zhou2016 [161]	MOABB	4	6.24	14	CC BY 4.0

C Pretraining details

C.1 Detailed hyperparameters

In this section, we present the detailed hyperparameters for pretraining our models. Table 7 summarizes the universal settings shared across all model sizes, including the core architecture, optimization schedule, and augmentation parameters. Table 8 details the specific architectural dimensions and learning rates scaled for the 18M, 48M, 194M, and 396M models.

Table 7: Hyperparameters for the SPERA model pretraining.

Hyperparameter	Value
EEG input & tokenization	
Sampling rate	200 Hz
Patch size	1.0 sec
Masking & student augmentation	
Temporal/spatial mask probability	0.80 / 0.80
Temporal mask ratio (min/max)	0.15 / 0.35
Spatial mask ratio (min/max)	0.10 / 0.30
Amplitude gain (min/max)	0.80 / 1.20
Gaussian noise std (min/max)	0.00 / 0.03
Coordinate jitter (prob/std)	0.50 / 0.05
Architecture	
MLP type / ratio	SwiGLU / 4.0
Predictor layers	2
Encoder architecture	Hybrid (full attn every 4th layer)
Normalization	Layer normalization
Dropout (residual/attention)	0.0
Temporal embedding	RoPE ($\theta = 10000$)
Legendre degree (L)	8
Legendre anchor points (M)	32
Optimization	
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$)
Weight decay	0.05
Total steps	47,000
Input tokens per update (nominal)	393,216
- Context / target tokens	262,144 / 131,072
LR scheduler	Warmup-cosine annealing
Warmup ratio	0.05
EMA schedule	Cosine annealing ($0.996 \rightarrow 0.9999$)
RSReg	
Weight schedule	Delayed linear ramp-decay
Weight bounds	$0 \rightarrow 0.02 \rightarrow 0.002$
Schedule steps	5% (ramp start), 10% (peak), 50% (decay start)
Projection dimension	128
Temperature (τ_z, τ_s)	0.1, 0.1

Table 8: Scale-specific architecture and optimization hyperparameters. Attention and predictor head counts and hidden dimensions are identical within each model size. Learning rates are scaled according to the model size.

Model size	Params	Layers	Hidden dim	Attention heads	Peak LR
Tiny	18M	8	384	4	4.0×10^{-4}
Small	48M	12	512	8	3.5×10^{-4}
Base	194M	16	896	14	3.0×10^{-4}
Large	396M	25	1024	16	2.5×10^{-4}

C.2 Pretraining token accounting and loss curve

We train SPERA for 47,000 optimization steps, corresponding to approximately two epochs over the pretraining corpus. Optimization steps are triggered adaptively: microbatches accumulate until the cumulative target-token count reaches the per-update budget of 131,072 tokens (Table 7), so the number of microbatches per update varies.

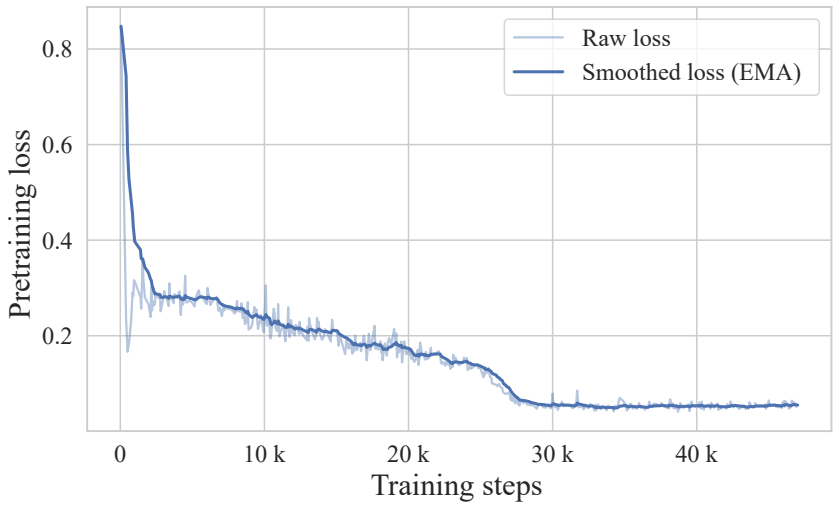

Figure 3: Pretraining loss curve of the SPERA-Base (194M) model. The model is trained for approximately 47,000 steps (∼2 epochs of the pretraining corpus). The raw per-step loss (light) is shown together with an exponential moving average for readability (solid); this plot-level smoothing is distinct from the EMA used to update the teacher encoder.

Figure 3 shows the pretraining loss curve of the SPERA-Base (194M) model. Each epoch realizes ∼9.45B input tokens (∼6.25B context, ∼3.20B target), or ∼98.4% of the ∼9.60B nominal token count derived from corpus statistics (Table 5). The remaining ∼1.6% corresponds to rare (C, P_t) shape buckets that do not accumulate a full microbatch within the dataloader's pending window and are dropped by the shape-aware batcher. The residual is small because the unit that is dropped is a rare shape bucket rather than a dataset: even our smallest dataset (∼0.91 h, 8 channels) produces ∼26K tokens, exceeding the per-microbatch budget. The resulting loss curve (Figure 3) converges smoothly and stabilizes around 0.05 by the end of training.

D Detailed downstream task results

In this section, we describe each downstream dataset, its preprocessing procedure, and the full experimental results. For all datasets, we follow the preprocessing pipelines of prior works [155, 57, 150, 94] to ensure direct comparison with baseline models. While we report balanced accuracy as the primary metric in the main text, we additionally report task-specific metrics commonly used in prior EEG foundation model evaluations [155, 57, 150, 94]. For multi-class classification tasks, we report Cohen's κ and weighted F1-score in addition to balanced accuracy. For binary classification tasks, we report AUC-PR and AUROC in addition to balanced accuracy. SPERA results are reported as mean $\pm$ standard deviation over five random seeds; baseline statistics follow the evaluation protocols of the source publications. Each downstream task table includes only baseline models whose original studies report results on that task. A dash indicates that the corresponding model does not report that dataset or metric. Unless otherwise specified, all results are obtained using the SPERA-Base (194M) model.

D.1 Imagined speech classification

BCIC2020-3 [56] is a 5-class imagined speech dataset. Fifteen subjects are instructed to silently imagine pronouncing one of five phrases ("hello", "help me", "stop", "thank you", and "yes") without articulation. EEG signals are recorded from 64 channels at 256 Hz, and resampled to 200 Hz. Each trial consists of a 3-second segment, yielding 6,000 samples in total with predefined train/validation/test splits of 60/10/10 trials per class.

As shown in Table 9, SPERA achieves 0.6504 balanced accuracy, with Cohen's κ of 0.5630 and weighted F1 of 0.6498.

Table 9: The results of different methods on imagined speech classification (BCIC2020-3, 5-class).

Methods	Balanced accuracy	Cohen's κ	Weighted F1
EEGNet	0.4413 ± 0.0096	0.3016 ± 0.0123	0.4413 ± 0.0102
EEGConformer	0.4506 ± 0.0133	0.3133 ± 0.0183	0.4488 ± 0.0154
SPaRCNet	0.4426 ± 0.0156	0.3033 ± 0.0233	0.4420 ± 0.0108
ContraWR	0.4257 ± 0.0162	0.3078 ± 0.0218	0.4407 ± 0.0182
CNN-Transformer	0.4533 ± 0.0092	0.3166 ± 0.0118	0.4506 ± 0.0127
FFCL	0.4678 ± 0.0197	0.3301 ± 0.0359	0.4689 ± 0.0205
ST-Transformer	0.4126 ± 0.0122	0.2941 ± 0.0159	0.4247 ± 0.0138
BIOT	0.4920 ± 0.0086	0.3650 ± 0.0176	0.4917 ± 0.0079
LaBraM-Base	0.5060 ± 0.0155	0.3800 ± 0.0242	0.5054 ± 0.0205
CBraMod	0.5373 ± 0.0108	0.4216 ± 0.0163	0.5383 ± 0.0096
REVE-Base	0.5635 ± 0.0123	0.4543 ± 0.0154	0.5633 ± 0.0124
CSBrain	0.6004 ± 0.0187	0.5006 ± 0.0233	0.6003 ± 0.0192
CodeBrain	0.6101 ± 0.0052	0.5127 ± 0.0065	0.6101 ± 0.0053
EEGMamba	0.5305 ± 0.0109	0.4183 ± 0.0157	0.5347 ± 0.0082
SPERA	$\mathbf{0.6504 \pm 0.0122}$	$\mathbf{0.5630 \pm 0.0152}$	$\mathbf{0.6498 \pm 0.0129}$

D.2 Motor imagery classification

BCIC-IV-2a [143] is a 4-class motor imagery dataset comprising nine subjects performing left hand, right hand, both feet, and tongue imagery. EEG signals are recorded across two sessions using 22 electrodes at 250 Hz. Each trial is cropped to the 2–6 s segment, bandpass filtered to 0.5–99.5 Hz, and resampled to 200 Hz, yielding 5,184 4-second samples. Following prior work [94], Euclidean alignment is additionally applied. We adopt a subject-independent split, with subjects 1–5, 6–7, and 8–9 used for training, validation, and test, respectively.

PhysioNet-MI [38, 122] is a 4-class motor imagery dataset recorded from 64 channels at 160 Hz, covering left fist, right fist, both fists, and both feet imagery. We extract 4-second segments yielding 9,837 samples, and apply a subject-independent split with subjects 1–70, 71–89, and 90–109 for training, validation, and test, respectively. All recordings are loaded at 128 Hz with a 64 Hz low-pass filter and subsequently resampled to 200 Hz.

The detailed experimental results for both motor imagery tasks are summarized in Table 10. SPERA improves over REVE-Base on both BCIC-IV-2a (+0.0420) and PhysioNet-MI (+0.0246) in balanced accuracy, with similar trends in the other two metrics. The improvement is consistent across the two datasets despite their differing channel counts (22 vs. 64) and acquisition protocols.

Table 10: The results of different methods on motor imagery classification.

Methods	BCIC-IV-2a, 4-class			PhysioNet-MI, 4-class		
	Balanced accuracy	Cohen's κ	Weighted F1	Balanced accuracy	Cohen's κ	Weighted F1
EEGNet	0.4482 ± 0.0094	0.2693 ± 0.0120	0.4226 ± 0.0108	0.5814 ± 0.0125	0.4468 ± 0.0199	0.5796 ± 0.0115
EEGConformer	0.4696 ± 0.0106	0.2924 ± 0.0141	0.4533 ± 0.0128	0.6049 ± 0.0104	0.4736 ± 0.0171	0.6062 ± 0.0095
SPaRCNet	0.4635 ± 0.0117	0.2847 ± 0.0147	0.4432 ± 0.0120	0.5932 ± 0.0152	0.4564 ± 0.0234	0.5937 ± 0.0147
ContraWR	0.4678 ± 0.0125	0.2905 ± 0.0160	0.4413 ± 0.0142	0.5892 ± 0.0133	0.4527 ± 0.0248	0.5918 ± 0.0116
CNN-Transformer	0.4600 ± 0.0108	0.2800 ± 0.0148	0.4460 ± 0.0114	0.6053 ± 0.0118	0.4725 ± 0.0223	0.6041 ± 0.0105
FFCL	0.4470 ± 0.0143	0.2627 ± 0.0176	0.4238 ± 0.0139	0.5726 ± 0.0092	0.4323 ± 0.0182	0.5701 ± 0.0079
ST-Transformer	0.4575 ± 0.0145	0.2733 ± 0.0190	0.4471 ± 0.0142	0.6035 ± 0.0081	0.4712 ± 0.0199	0.6053 ± 0.0075
BIOT	0.4748 ± 0.0093	0.2997 ± 0.0134	0.4607 ± 0.0125	0.6153 ± 0.0154	0.4875 ± 0.0272	0.6158 ± 0.0197
LaBraM-Base	0.4869 ± 0.0085	0.3159 ± 0.0154	0.4728 ± 0.0105	0.6173 ± 0.0122	0.4912 ± 0.0192	0.6177 ± 0.0141
CBraMod	0.5138 ± 0.0066	0.3518 ± 0.0094	0.4984 ± 0.0085	0.6417 ± 0.0091	0.5222 ± 0.0169	0.6427 ± 0.0100
REVE-Base	0.6396 ± 0.0095	0.5194 ± 0.0126	0.6339 ± 0.0110	0.6480 ± 0.0140	0.5306 ± 0.0187	0.6484 ± 0.0170
CSBrain	0.5657 ± 0.0071	0.4209 ± 0.0093	0.5637 ± 0.0087	0.6304 ± 0.0090	0.5071 ± 0.0120	0.6308 ± 0.0095
EEGMamba	–	–	–	0.6323 ± 0.0084	0.5131 ± 0.0124	0.6342 ± 0.0098
SPERA	$\mathbf{0.6816 \pm 0.0147}$	$\mathbf{0.5755 \pm 0.0197}$	$\mathbf{0.6817 \pm 0.0155}$	$\mathbf{0.6726 \pm 0.0072}$	$\mathbf{0.5633 \pm 0.0096}$	$\mathbf{0.6731 \pm 0.0080}$

D.3 Emotion recognition

FACED [23] is a 9-class emotion recognition dataset comprising 123 subjects recorded using 32 channels at 250 Hz, covering amusement, inspiration, joy, tenderness, anger, fear, disgust, sadness, and neutral. EEG signals are resampled to 200 Hz and segmented into 10-second samples, yielding 10,332 samples in total. Subjects 1–80, 81–100, and 101–123 are used for training, validation, and test, respectively.

Detailed results for emotion recognition are shown in Table 11. SPERA and CSBrain achieve nearly identical balanced accuracy (0.5754 vs. 0.5752). CodeBrain reports higher values on all three metrics; FACED is the only task among the nine where a recent model leads SPERA on every metric.

Table 11: The results of different methods on emotion recognition (FACED, 9-class).

Methods	Balanced accuracy	Cohen's κ	Weighted F1
EEGNet	0.4090 ± 0.0122	0.3342 ± 0.0251	0.4124 ± 0.0141
EEGConformer	0.4559 ± 0.0125	0.3858 ± 0.0186	0.4514 ± 0.0107
SPaRCNet	0.4673 ± 0.0155	0.3978 ± 0.0289	0.4729 ± 0.0133
ContraWR	0.4887 ± 0.0078	0.4231 ± 0.0151	0.4884 ± 0.0074
CNN-Transformer	0.4697 ± 0.0132	0.4017 ± 0.0168	0.4720 ± 0.0125
FFCL	0.4673 ± 0.0158	0.3987 ± 0.0383	0.4699 ± 0.0145
ST-Transformer	0.4810 ± 0.0079	0.4137 ± 0.0133	0.4795 ± 0.0096
BIOT	0.5118 ± 0.0118	0.4476 ± 0.0254	0.5136 ± 0.0112
LaBraM-Base	0.5273 ± 0.0107	0.4698 ± 0.0188	0.5288 ± 0.0102
CBraMod	0.5509 ± 0.0089	0.5041 ± 0.0122	0.5618 ± 0.0093
REVE-Base	0.5646 ± 0.0164	0.5080 ± 0.0191	0.5659 ± 0.0172
CSBrain	0.5752 ± 0.0042	0.5204 ± 0.0036	0.5796 ± 0.0031
CodeBrain	$\mathbf{0.5941 \pm 0.0098}$	$\mathbf{0.5406 \pm 0.0084}$	$\mathbf{0.5953 \pm 0.0113}$
EEGMamba	0.5646 ± 0.0096	0.5134 ± 0.0127	0.5673 ± 0.0076
SPERA	0.5754 ± 0.0175	0.5122 ± 0.0286	0.5711 ± 0.0138

D.4 Mental stress detection

Mental arithmetic (MAT) [164] is a 2-class mental stress dataset comprising 36 subjects recorded using 20 channels at 500 Hz, distinguishing resting state from active mental arithmetic. EEG signals are bandpass filtered (0.5–45 Hz), resampled to 200 Hz, and segmented into 1,707 5-second samples. Subjects 1–28, 29–32, and 33–36 are used for training, validation, and test, respectively.

The experimental results are shown in Table 12. SPERA exceeds REVE-Base by $+0.0118$ in balanced accuracy and $+0.0454$ in AUROC, while its AUC-PR is lower.

D.5 Sleep staging

ISRUC [63] is a 5-class sleep staging dataset comprising 100 subjects recorded using 6 channels at 200 Hz. EEG signals are segmented into 30-second epochs and annotated according to the AASM standard (Wake, N1, N2, N3, REM). Subjects 1–80, 81–90, and 91–100 are used for training, validation, and test, respectively. Subject 8 was excluded from the final dataset due to missing requisite EEG channels (e.g., F3, F4), resulting in a total of 88,140 samples. Sleep staging is commonly framed as a sequence-to-sequence classification problem with a sequence length of 20 consecutive 30-second epochs to model sleep stage transitions [107]. Following the protocol of CBraMod [150], we adopt this standard setting and additionally report single-segment classification results to assess the quality of our pretrained representations without sequential context. We further include two widely recognized sleep-staging-specific baselines, DeepSleepNet [142] and U-Sleep [103], in order to compare against classical methods established in the sleep staging literature.

Full classification results are summarized in Table 13. The single-segment variant alone (0.8093) exceeds the strongest baseline (CSBrain, 0.7925) in balanced accuracy, indicating that the gain does not require the sequence-to-sequence formulation. The sequence-to-sequence variant further improves performance to 0.8137. These results indicate that SPERA remains effective on ISRUC,

Table 12: The results of different methods on mental stress detection (MAT, 2-class).

Methods	Balanced accuracy	AUC-PR	AUROC
EEGNet	0.6770 ± 0.0116	0.5763 ± 0.0102	0.7321 ± 0.0108
EEGConformer	0.6805 ± 0.0123	0.5829 ± 0.0134	0.7424 ± 0.0128
SPaRCNet	0.6879 ± 0.0107	0.5825 ± 0.0193	0.7418 ± 0.0132
ContraWR	0.6631 ± 0.0097	0.5787 ± 0.0164	0.7332 ± 0.0082
CNN-Transformer	0.6779 ± 0.0268	0.5777 ± 0.0285	0.7258 ± 0.0336
FFCL	0.6798 ± 0.0142	0.5786 ± 0.0266	0.7330 ± 0.0198
ST-Transformer	0.6631 ± 0.0173	0.5672 ± 0.0259	0.7132 ± 0.0174
BIOT	0.6875 ± 0.0186	0.6004 ± 0.0195	0.7536 ± 0.0144
LaBraM-Base	0.6909 ± 0.0125	0.5999 ± 0.0155	0.7721 ± 0.0093
CBraMod	0.7256 ± 0.0132	0.6267 ± 0.0099	0.7905 ± 0.0073
REVE-Base	0.7660 ± 0.0355	$\mathbf{0.7470 \pm 0.0807}$	0.8450 ± 0.0514
CSBrain	0.7558 ± 0.0106	0.6696 ± 0.0221	0.8478 ± 0.0297
CodeBrain	0.7514 ± 0.0203	0.7177 ± 0.0421	0.8707 ± 0.0209
SPERA	$\mathbf{0.7778 \pm 0.0216}$	0.7291 ± 0.0226	$\mathbf{0.8904 \pm 0.0196}$

Table 13: The results of different methods on sleep staging (ISRUC, 5-class).

Methods	Balanced accuracy	Cohen's κ	Weighted F1
EEGNet	0.7154 ± 0.0121	0.7040 ± 0.0173	0.7513 ± 0.0124
EEGConformer	0.7400 ± 0.0133	0.7143 ± 0.0162	0.7634 ± 0.0150
SPaRCNet	0.7487 ± 0.0075	0.7097 ± 0.0132	0.7624 ± 0.0092
ContraWR	0.7402 ± 0.0126	0.7178 ± 0.0156	0.7610 ± 0.0137
CNN-Transformer	0.7363 ± 0.0087	0.7129 ± 0.0121	0.7719 ± 0.0105
FFCL	0.7277 ± 0.0182	0.7016 ± 0.0291	0.7614 ± 0.0197
ST-Transformer	0.7381 ± 0.0205	0.7013 ± 0.0352	0.7681 ± 0.0175
DeepSleepNet	0.7419 ± 0.0144	0.7036 ± 0.0241	0.7643 ± 0.0122
U-Sleep	0.7586 ± 0.0116	0.7209 ± 0.0143	0.7805 ± 0.0105
BIOT	0.7527 ± 0.0121	0.7192 ± 0.0231	0.7790 ± 0.0146
LaBraM-Base	0.7633 ± 0.0102	0.7231 ± 0.0182	0.7810 ± 0.0133
CBraMod	0.7865 ± 0.0110	0.7442 ± 0.0152	0.8011 ± 0.0099
REVE-Base	0.7819 ± 0.0078	0.7500 ± 0.0156	0.8005 ± 0.0135
CSBrain	0.7925 ± 0.0030	0.7406 ± 0.0102	0.7990 ± 0.0091
CodeBrain	0.7835 ± 0.0033	0.7476 ± 0.0040	0.8020 ± 0.0018
EEGMamba	0.7899 ± 0.0093	0.7486 ± 0.0122	0.8054 ± 0.0076
SPERA (single-segment)	0.8093 ± 0.0108	0.7542 ± 0.0187	0.8115 ± 0.0150
SPERA (sequence-to-sequence)	$\mathbf{0.8137 \pm 0.0075}$	$\mathbf{0.7662 \pm 0.0121}$	$\mathbf{0.8217 \pm 0.0097}$

which has the fewest channels (6) and the longest segment duration (30 s) among the nine downstream datasets.

D.6 Mental disorder diagnosis

Mumtaz [89] is a 2-class depression detection dataset comprising 34 patients with major depressive disorder (MDD) and 30 normal controls, recorded using 19 channels at 256 Hz. Only eyes-open and eyes-closed resting-state sessions are used; MDD subject 8 is excluded because only a P300 recording is available. EEG signals are bandpass filtered (0.3–75 Hz), notch filtered (50 Hz), resampled to 200 Hz, and segmented into 5-second samples. Following the publicly released preprocessing implementation of prior EEG foundation model studies [150, 94][1], we apply the released train/validation/test partition. The resulting split contains 22/5/6 MDD and 21/4/5 control

[1]The textual descriptions in [150, 94] report subject counts of 24/5/5 MDD and 19/4/5 controls, whereas applying the released preprocessing code to the resting-state recordings yields 22/5/6 and 21/4/5, respectively. We follow the released implementation, which determines the data partition used in the experiments.

Table 14: The results of different methods on mental disorder diagnosis (Mumtaz, 2-class).

Methods	Balanced accuracy	AUC-PR	AUROC
EEGNet	0.9232 ± 0.0104	0.9626 ± 0.0095	0.9639 ± 0.0093
EEGConformer	0.9308 ± 0.0117	0.9684 ± 0.0105	0.9702 ± 0.0101
SPaRCNet	0.9316 ± 0.0095	0.9754 ± 0.0065	0.9781 ± 0.0083
ContraWR	0.9195 ± 0.0115	0.9589 ± 0.0102	0.9621 ± 0.0092
CNN-Transformer	0.9305 ± 0.0068	0.9757 ± 0.0074	0.9742 ± 0.0055
FFCL	0.9314 ± 0.0038	0.9717 ± 0.0021	0.9753 ± 0.0033
ST-Transformer	0.9135 ± 0.0103	0.9578 ± 0.0086	0.9594 ± 0.0059
BIOT	0.9358 ± 0.0052	0.9736 ± 0.0034	0.9758 ± 0.0042
LaBraM-Base	0.9409 ± 0.0079	0.9798 ± 0.0093	0.9782 ± 0.0057
CBraMod	0.9560 ± 0.0056	0.9923 ± 0.0032	0.9921 ± 0.0025
REVE-Base	$\mathbf{0.9644 \pm 0.0097}$	0.9961 ± 0.0013	0.9957 ± 0.0015
CSBrain	0.9643 ± 0.0155	0.9936 ± 0.0034	0.9956 ± 0.0020
SPERA	0.9637 ± 0.0018	$\mathbf{0.9973 \pm 0.0007}$	$\mathbf{0.9975 \pm 0.0010}$

subjects (33 MDD and 30 controls in total), yielding a total of 7,143 5-second samples consistent with prior work.

The experimental results are summarized in Table 14. SPERA and REVE-Base are comparable in balanced accuracy (0.9637 vs. 0.9644), with SPERA achieving the highest AUC-PR and AUROC. SPERA also exhibits the smallest across-seed standard deviation in balanced accuracy ($\pm$ 0.0018) among all compared methods.

D.7 EEG abnormality detection

TUAB [91] is a 2-class EEG abnormality detection dataset recorded using 23 channels at 250 Hz, with 16 common bipolar montage channels used for analysis. EEG signals are resampled to 200 Hz and segmented into 409,455 10-second samples. The original train/test split is adopted, with the training set further divided into training and validation subsets (80%:20%). TUAB is excluded from our pretraining corpus, preventing any evaluation leakage.

Table 15: The results of different methods on EEG abnormality detection (TUAB, 2-class).

Methods	Balanced accuracy	AUC-PR	AUROC
EEGNet	0.7642 ± 0.0036	0.8299 ± 0.0043	0.8412 ± 0.0030
EEGConformer	0.7758 ± 0.0049	0.8427 ± 0.0054	0.8445 ± 0.0038
SPaRCNet	0.7896 ± 0.0018	0.8414 ± 0.0018	0.8676 ± 0.0011
ContraWR	0.7746 ± 0.0041	0.8421 ± 0.0104	0.8456 ± 0.0074
CNN-Transformer	0.7777 ± 0.0022	0.8433 ± 0.0039	0.8461 ± 0.0013
FFCL	0.7848 ± 0.0038	0.8448 ± 0.0065	0.8569 ± 0.0051
ST-Transformer	0.7966 ± 0.0023	0.8521 ± 0.0026	0.8707 ± 0.0019
EEG2Rep	–	–	0.8843 ± 0.0309
BIOT	0.7959 ± 0.0057	0.8792 ± 0.0023	0.8815 ± 0.0043
LaBraM-Base	0.8140 ± 0.0019	0.8965 ± 0.0016	0.9022 ± 0.0008
LaBraM-Large	0.8226 ± 0.0015	0.9130 ± 0.0005	0.9127 ± 0.0005
LaBraM-Huge	0.8258 ± 0.0011	0.9204 ± 0.0011	0.9162 ± 0.0016
CBraMod	0.8289 ± 0.0022	0.9258 ± 0.0008	0.9227 ± 0.0011
REVE-Base	0.8315 ± 0.0014	0.9281 ± 0.0009	0.9245 ± 0.0013
CSBrain	0.8172 ± 0.0043	0.9005 ± 0.0066	0.8957 ± 0.0046
CodeBrain	0.8294 ± 0.0013	0.9100 ± 0.0006	0.9030 ± 0.0009
LuMamba	0.8099 ± 0.0022	0.8918 ± 0.0032	0.8825 ± 0.0038
SPERA	$\mathbf{0.8488 \pm 0.0035}$	$\mathbf{0.9287 \pm 0.0045}$	$\mathbf{0.9253 \pm 0.0046}$

As shown in Table 15, SPERA improves balanced accuracy by +0.0173 over REVE-Base, with smaller gains in AUC-PR and AUROC.

D.8 Event type classification

TUEV [91] is a 6-class EEG event classification dataset recorded using 23 channels at 250 Hz, covering spike and sharp wave, generalized periodic epileptiform discharges, periodic lateralized epileptiform discharges, eye movement, artifact, and background. EEG signals from 16 common bipolar montage channels are resampled to 200 Hz and segmented into 112,491 5-second samples. The original train/test split is adopted, with the training set further divided into training and validation subsets (80%:20%). As with TUAB, TUEV is excluded from our pretraining corpus.

The experimental results are summarized in Table 16. SPERA achieves the highest balanced accuracy among the evaluated methods (0.7684). EEGDM reports higher Cohen's κ and weighted F1 than SPERA, while GRAM reports the highest weighted F1.

Table 16: The results of different methods on event type classification (TUEV, 6-class).

Methods	Balanced accuracy	Cohen's κ	Weighted F1
EEGNet	0.3876 ± 0.0143	0.3577 ± 0.0155	0.6539 ± 0.0120
EEGConformer	0.4074 ± 0.0164	0.3967 ± 0.0195	0.6983 ± 0.0152
SPaRCNet	0.4161 ± 0.0262	0.4233 ± 0.0181	0.7024 ± 0.0104
ContraWR	0.4384 ± 0.0349	0.3912 ± 0.0237	0.6893 ± 0.0136
CNN-Transformer	0.4087 ± 0.0161	0.3815 ± 0.0134	0.6854 ± 0.0290
FFCL	0.3979 ± 0.0104	0.3732 ± 0.0188	0.6783 ± 0.0120
ST-Transformer	0.3984 ± 0.0228	0.3765 ± 0.0306	0.6823 ± 0.0190
EEG2Rep	0.5295 ± 0.0158	–	0.7508 ± 0.0121
EEGDM	0.7557 ± 0.0147	$\mathbf{0.7423 \pm 0.0136}$	0.8688 ± 0.0066
BIOT	0.5281 ± 0.0225	0.5273 ± 0.0249	0.7492 ± 0.0080
LaBraM-Base	0.6409 ± 0.0065	0.6637 ± 0.0093	0.8312 ± 0.0052
LaBraM-Large	0.6581 ± 0.0156	0.6622 ± 0.0136	0.8315 ± 0.0040
LaBraM-Huge	0.6616 ± 0.0170	0.6745 ± 0.0195	0.8329 ± 0.0086
CBraMod	0.6671 ± 0.0107	0.6772 ± 0.0096	0.8342 ± 0.0064
REVE-Base	0.6759 ± 0.0229	0.6783 ± 0.0199	0.8451 ± 0.0129
CSBrain	0.6903 ± 0.0059	0.6833 ± 0.0047	0.8333 ± 0.0057
CodeBrain	0.6428 ± 0.0062	0.6912 ± 0.0101	0.8362 ± 0.0048
GRAM	0.7487 ± 0.0151	0.7130 ± 0.0333	$\mathbf{0.8824 \pm 0.0158}$
SPERA	$\mathbf{0.7684 \pm 0.0283}$	0.7320 ± 0.0366	0.8619 ± 0.0178

D.9 Linear probing results

Table 17 reports the full per-task LP results across all nine downstream datasets. The largest gains over baselines are observed on BCIC2020-3 and TUEV, where SPERA-Large surpasses REVE-Large by over 0.1 in balanced accuracy. On BCIC-IV-2a, PhysioNet-MI, MAT, and Mumtaz, SPERA models reach competitive but not superior performance. Mumtaz in particular shows limited headroom, with most foundation models exceeding 0.9 balanced accuracy. While most tasks improve with model size, PhysioNet-MI, FACED, and MAT show non-monotonic scaling under LP.

Table 18 additionally compares SPERA-Base under LP with recent models on TUAB and TUEV, the tasks from our evaluation suite that overlap with their reported evaluations. SPERA-Base attains the highest balanced accuracy on both tasks. On TUEV, however, EEGPT reports a higher Cohen's κ and weighted F1.

D.10 Per-task significance

We evaluate the statistical significance of SPERA's task-level performance differences across the nine downstream tasks. For consistency across tasks, the comparator pool is restricted to baselines with balanced-accuracy results reported on all nine tasks (Table 2); methods with partial task coverage are

Table 17: LP results on nine downstream tasks. The results are reported as mean balanced accuracy ± standard deviation over five random seeds. The best performance for each task is highlighted in bold. Baseline model results are extracted from [94]. The standard deviations for the average performance of the baseline models are omitted as they are not reported in the original paper.

Methods	Params	BCIC2020-3	BCIC-IV-2a	PhysioNet-MI	FACED	MAT
CBraMod	4M	0.374 ± 0.007	0.376 ± 0.006	0.531 ± 0.015	0.261 ± 0.013	0.605 ± 0.020
REVE-Base	69M	0.390 ± 0.017	0.517 ± 0.015	0.510 ± 0.012	0.422 ± 0.028	$\mathbf{0.740 \pm 0.073}$
REVE-Large	408M	0.378 ± 0.021	$\mathbf{0.603 \pm 0.011}$	$\mathbf{0.617 \pm 0.000}$	0.469 ± 0.007	0.665 ± 0.103
SPERA-Tiny	18M	0.324 ± 0.024	0.497 ± 0.009	0.532 ± 0.007	0.343 ± 0.009	0.689 ± 0.014
SPERA-Small	48M	0.407 ± 0.015	0.497 ± 0.013	0.523 ± 0.011	0.363 ± 0.017	0.630 ± 0.022
SPERA-Base	194M	0.502 ± 0.011	0.560 ± 0.010	0.551 ± 0.009	$\mathbf{0.478 \pm 0.011}$	0.672 ± 0.014
SPERA-Large	396M	$\mathbf{0.593 \pm 0.019}$	0.590 ± 0.014	0.583 ± 0.008	0.469 ± 0.010	0.693 ± 0.016

Methods	Params	ISRUC	Mumtaz	TUAB	TUEV	Average
CBraMod	4M	0.430 ± 0.043	0.907 ± 0.027	0.500 ± 0.000	0.482 ± 0.037	0.496
REVE-Base	69M	0.662 ± 0.030	0.931 ± 0.021	0.809 ± 0.004	0.508 ± 0.073	0.610
REVE-Large	408M	0.758 ± 0.001	$\mathbf{0.980 \pm 0.009}$	0.809 ± 0.004	0.550 ± 0.014	0.648
SPERA-Tiny	18M	0.687 ± 0.027	0.885 ± 0.029	0.784 ± 0.003	0.517 ± 0.007	0.584 ± 0.006
SPERA-Small	48M	0.721 ± 0.024	0.929 ± 0.011	0.787 ± 0.006	0.558 ± 0.010	0.601 ± 0.005
SPERA-Base	194M	0.772 ± 0.014	0.944 ± 0.015	0.821 ± 0.002	0.651 ± 0.008	0.661 ± 0.003
SPERA-Large	396M	$\mathbf{0.775 \pm 0.019}$	0.957 ± 0.032	$\mathbf{0.842 \pm 0.003}$	$\mathbf{0.653 \pm 0.008}$	$\mathbf{0.684 \pm 0.005}$

Table 18: LP comparison with recent models on TUAB and TUEV. For the binary TUAB task, we report AUC-PR and AUROC; for the 6-class TUEV task, we report Cohen's κ and weighted F1. Baseline numbers are taken from the original publications. Laya evaluates TUAB under its own LP protocol, in which reproduced baseline scores differ substantially from those in Table 17.

Methods	TUAB, 2-class			TUEV, 6-class		
	Balanced accuracy	AUC-PR	AUROC	Balanced accuracy	Cohen's κ	Weighted F1
EEGPT	0.798 ± 0.003	–	0.872 ± 0.005	0.623 ± 0.011	$\mathbf{0.635 \pm 0.013}$	$\mathbf{0.819 \pm 0.006}$
EEG2Rep	–	–	0.832 ± 0.033	0.433 ± 0.031	–	0.700 ± 0.032
Laya	0.798 ± 0.007	–	–	–	–	–
SPERA-Base	$\mathbf{0.821 \pm 0.002}$	$\mathbf{0.898 \pm 0.001}$	$\mathbf{0.901 \pm 0.002}$	$\mathbf{0.651 \pm 0.008}$	0.580 ± 0.015	0.784 ± 0.008

Table 19: Per-task significance under LPFT. The comparator is the highest-balanced-accuracy baseline among methods reporting results on all nine downstream tasks; methods with partial task coverage are excluded. Δ denotes the balanced-accuracy difference relative to the comparator, brackets show 95% confidence intervals, and bold indicates $p_{\mathrm{Holm}} < 0.05$.

Task	Comparator	SPERA-Base (194M)			SPERA-Large (396M)		
		Δ [95% CI]	p	p_{Holm}	Δ [95% CI]	p	p_{Holm}
BCIC2020-3	CSBrain	$+0.0500$ [$+0.026$, $+0.074$]	0.002	**0.011**	$+0.0879$ [$+0.063$, $+0.112$]	< 0.001	**< 0.001**
TUEV	CSBrain	$+0.0781$ [$+0.043$, $+0.113$]	0.003	**0.014**	$+0.0966$ [$+0.061$, $+0.132$]	0.001	**0.008**
BCIC-IV-2a	REVE-Base	$+0.0420$ [$+0.023$, $+0.061$]	0.001	**0.009**	$+0.0627$ [$+0.044$, $+0.081$]	< 0.001	**< 0.001**
ISRUC	CSBrain	$+0.0212$ [$+0.012$, $+0.030$]	0.002	**0.011**	$+0.0217$ [$+0.009$, $+0.035$]	0.008	**0.033**
TUAB	REVE-Base	$+0.0173$ [$+0.013$, $+0.022$]	< 0.001	**0.001**	$+0.0293$ [$+0.026$, $+0.033$]	< 0.001	**< 0.001**
PhysioNet-MI	REVE-Base	$+0.0246$ [$+0.007$, $+0.042$]	0.013	0.052	$+0.0337$ [$+0.016$, $+0.051$]	0.003	**0.014**
MAT	REVE-Base	$+0.0118$ [-0.033, $+0.056$]	0.547	1.000	$+0.0361$ [-0.008, $+0.080$]	0.093	0.279
FACED	CSBrain	$+0.0002$ [-0.021, $+0.022$]	0.981	1.000	$+0.0167$ [-0.006, $+0.040$]	0.115	0.279
Mumtaz	REVE-Base	-0.0007 [-0.013, $+0.011$]	0.881	1.000	$+0.0058$ [-0.006, $+0.018$]	0.254	0.279

excluded. For each task, SPERA is compared with the highest-performing baseline in this pool using Welch's t-test over five random seeds, with Holm correction applied across the nine tasks separately for each SPERA model. The tests are two-sided, and the 95% confidence intervals are pointwise and are not adjusted for multiple comparisons. The tests are unpaired because baseline results are taken from the original publications and per-seed values are unavailable.

Table 19 reports the difference in balanced accuracy Δ with its 95% confidence interval, alongside raw and Holm-corrected p-values. For SPERA-Base, five tasks remain significant after correction: BCIC2020-3, TUEV, BCIC-IV-2a, ISRUC, and TUAB. PhysioNet-MI is borderline ($p_{\mathrm{Holm}} = 0.052$). The differences on FACED, MAT, and Mumtaz did not reach statistical significance after correction.

Mumtaz shows limited headroom, with generally high balanced accuracy across foundation models and model scales. For MAT, the estimated performance difference is small relative to the across-seed variability, and the confidence interval includes zero, whereas on FACED, SPERA-Base and CSBrain achieve nearly identical balanced accuracy (0.5754 vs. 0.5752).

SPERA-Large shows a larger Δ than SPERA-Base on every task: the same five tasks remain significant, and PhysioNet-MI also reaches significance after correction. FACED, MAT, and Mumtaz remain non-significant. We report the main-text claim for SPERA-Base, the default model used throughout the paper, and include SPERA-Large here for completeness.

E Additional ablation experiments

E.1 Pretraining objective and corpus

The ablations in Table 3 vary architectural components while holding the primary prediction objective and corpus fixed. Here we vary those two factors instead, to separate the contribution of latent-space prediction from that of the pretraining data. Both variants use the SPERA-Small backbone and follow the ablation protocol of Section 4.5: LP on PhysioNet-MI, ISRUC, and TUAB. Each configuration is evaluated over five random seeds.

For the reconstruction variant, we replace the latent prediction objective with input-space reconstruction. The predictor outputs are decoded back to raw patches by a linear head and the model is trained against the masked input patches, so that the target is computed from the raw signal. All other components, including the Legendre prior, the hybrid backbone, and the relational spectral regularizer, are unchanged.

For the REVE corpus variant, we restrict pretraining to dataset sources shared with REVE [94] to assess the effect of this corpus restriction on SPERA's downstream performance. The proxy comprises 81 datasets from our corpus that are also listed in REVE's pretraining corpus, totaling 61,191 h from 22,994 subjects, comparable in duration to the approximately 60,000 h reported for REVE. The 25 excluded datasets account for 18,473 h and contribute 6,054 additional subjects; 16,367 h are sleep recordings from MESA and the HSP subset, and the remaining 2,106 h come from 23 datasets.

Replacing latent prediction with input-space reconstruction lowers average balanced accuracy by 0.022 with the backbone, corpus, and token budget held fixed, and the drop is consistent across all three tasks. Pretraining on the REVE corpus proxy lowers the average by 0.005. Specifically, PhysioNet-MI and TUAB are nearly identical, while ISRUC is lower by 0.016. This is consistent with the composition of the excluded data, 89% of which is sleep recordings, and ISRUC is the only sleep-staging task among the three.

Table 20: Ablation of the pretraining objective and corpus. All variants use the SPERA-Small backbone under the ablation protocol of Section 4.5. Results are mean balanced accuracy $\pm$ standard deviation over five random seeds. *Reconstruction head* replaces latent prediction with input-space reconstruction; *REVE corpus* pretrains on the 81-dataset proxy for the corpus used by REVE [94].

Variants	PhysioNet-MI	ISRUC	TUAB	Average
SPERA-Small (latent prediction)	$\mathbf{0.523 \pm 0.011}$	$\mathbf{0.721 \pm 0.024}$	0.787 ± 0.006	$\mathbf{0.677 \pm 0.011}$
with reconstruction head	0.499 ± 0.015	0.706 ± 0.019	0.761 ± 0.011	0.655 ± 0.013
with REVE corpus	0.522 ± 0.013	0.705 ± 0.021	$\mathbf{0.788 \pm 0.008}$	0.672 ± 0.010

E.2 Temporal positional encoding

SPERA applies rotary position embeddings (RoPE) along the temporal patch axis, while electrode position is supplied by the Legendre prior along the spatial axis. We ablate the temporal encoding by replacing RoPE with a standard sinusoidal absolute position embedding added to the patch embeddings, keeping the spatial encoding, the backbone, and the pretraining objective fixed. The variant follows the ablation protocol of Section 4.5. Note that the substitution affects the temporal axis only: in the factorized spatial pass RoPE is already disabled, and the geometry encodings continue to provide positional information there.

Table 21: Ablation of the temporal positional encoding. Both variants use the SPERA-Small backbone under the ablation protocol of Section 4.5. Results are mean balanced accuracy ± standard deviation over five random seeds; the best result in each column is in bold. The spatial encoding is unchanged in both variants.

Variants	PhysioNet-MI	ISRUC	TUAB	Average
RoPE (SPERA-Small)	**0.523 ± 0.011**	**0.721 ± 0.024**	**0.787 ± 0.006**	**0.677 ± 0.011**
Sinusoidal absolute encoding	0.502 ± 0.013	0.699 ± 0.020	0.764 ± 0.010	0.655 ± 0.011

The result is summarized in Table 21. RoPE is stronger on all three tasks, by 0.022 on average, supporting its use as the temporal positional encoding in the present architecture. This result is also consistent with prior reports that relative positional schemes can generalize across varying sequence lengths [139]. The three tasks in this ablation use 4-s, 30-s, and 10-s segments for PhysioNet-MI, ISRUC, and TUAB, respectively, corresponding to 4, 30, and 10 temporal tokens per channel with 1-s patches.

E.3 Masking strategy

To assess the impact of our masking strategy, we conduct an ablation comparing two strategies: block masking and semantic subsequence preserving (SSP) masking. We evaluate both strategies across three downstream tasks at five levels of expected masking ratio under LP using SPERA-Small. The total masking ratio is determined by independent temporal and spatial masking processes, each applied with probability $p = 0.8$. To prevent the trivial case where no tokens are masked, temporal masking is applied as a fallback. The expected total masking ratio ρ is computed as the expected union of the masked regions, assuming independence between temporal and spatial masking. We define five difficulty levels ($\rho \approx 0.27, 0.34, 0.45, 0.65, 0.75$); detailed hyperparameter ranges for each level are summarized in Table 22.

Table 22: Masking ratio hyperparameters per difficulty level. The expected masking ratio ρ accounts for the combined application of spatial and temporal masking, including the fallback condition.

Level	Spatial range	Spatial mean	Temporal range	Temporal mean	Expected ratio (ρ)
0	0.05 – 0.25	0.15	0.10 – 0.30	0.20	**0.27**
1	0.10 – 0.30	0.20	0.15 – 0.35	0.25	**0.34**
2	0.18 – 0.32	0.25	0.30 – 0.44	0.37	**0.45**
3	0.27 – 0.43	0.35	0.52 – 0.68	0.60	**0.65**
4	0.36 – 0.54	0.45	0.62 – 0.79	0.71	**0.75**

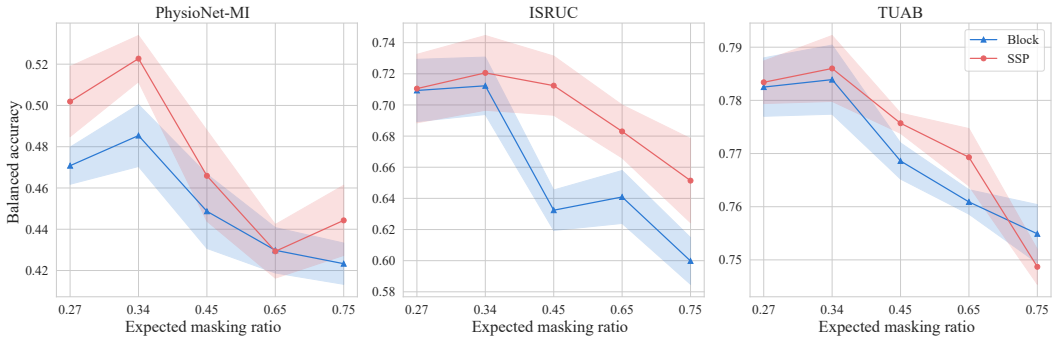

Figure 4: Masking strategy analysis across five difficulty levels on PhysioNet-MI, ISRUC, and TUAB. The blue line represents the performance of block masking, while the red line represents the performance of SSP masking. The shaded areas indicate one standard deviation across five random seeds.

As shown in Figure 4, SSP masking generally outperforms block masking, with the best performance on each task achieved under SSP at $\rho \approx 0.34$. This suggests that preserving semantic subsequences is beneficial for learning transferable EEG representations.

E.4 Additional metrics for component ablations

Table 3 reports balanced accuracy only. Table 23 adds Cohen's κ and weighted F1 for the multi-class tasks and AUC-PR and AUROC for binary TUAB, on the same runs. Balanced accuracy is repeated so that the comparison can be read within a single table. On each task the two leading variants are the same under every metric.

Table 23: Additional metrics for the component ablations of Table 3. Each cell reports balanced accuracy followed by two task-appropriate metrics: Cohen's κ and weighted F1 for the multi-class tasks, AUC-PR and AUROC for binary TUAB. The best result per column is in **bold**; the second-best is underlined.

Variant	PhysioNet-MI (BAcc / κ / wF1)	ISRUC (BAcc / κ / wF1)	TUAB (BAcc / AUC-PR / AUROC)
Encoder architecture			
Full attention only	0.4450 / 0.2600 / 0.4492	0.7097 / 0.6314 / 0.7131	0.7749 / 0.8538 / 0.8541
Factorized attention only	0.4833 / 0.3111 / 0.4781	0.7153 / 0.6378 / 0.7187	0.7865 / 0.8656 / 0.8650
Hybrid + late full	0.4871 / 0.3161 / 0.4820	0.6973 / 0.6194 / 0.7009	0.7860 / 0.8651 / 0.8645
RSReg			
No RSReg	0.5182 / 0.3576 / 0.5138	0.6994 / 0.6198 / 0.7029	0.7759 / 0.8548 / 0.8550
Electrode geometry encoding			
Coordinate embedding only	0.4821 / 0.3095 / 0.4769	0.7009 / 0.6215 / 0.7044	0.7744 / 0.8533 / 0.8536
Legendre bias only	0.5193 / 0.3591 / 0.5149	**0.7320 / 0.6566 / 0.7352**	**0.7929 / 0.8722 / 0.8710**
Legendre anchor encoding only	0.5071 / 0.3428 / 0.5025	0.7137 / 0.6360 / 0.7171	0.7855 / 0.8646 / 0.8640
Learnable pairwise MLP	0.4463 / 0.2617 / 0.4403	0.6679 / 0.5842 / 0.6826	0.7671 / 0.8458 / 0.8468
REVE spatial encoding	0.4921 / 0.3228 / 0.4862	0.6910 / 0.6159 / 0.6943	0.7607 / 0.8412 / 0.8420
Gaussian (geodesic)	0.4512 / 0.2683 / 0.4489	0.6977 / 0.6189 / 0.7010	0.7705 / 0.8501 / 0.8511
SPERA-Small	**0.5227 / 0.3636 / 0.5184**	0.7206 / 0.6438 / 0.7239	0.7869 / 0.8660 / 0.8653

F Analysis of the Legendre spherical prior

F.1 Geometry encoding ablations

Building on the electrode geometry ablation in Section 4.5, we evaluate a broader set of alternative geometry encodings. Everything outside the geometry encoding is held fixed. *Coordinate embedding only* removes the pairwise prior and retains the linear embedding of 3D electrode coordinates. *Anisotropic SH* replaces the rotation-invariant Legendre expansion with a general spherical-harmonic (SH) prior that permits direction-dependent interactions and contains the isotropic Legendre family as a special case. *Gaussian (geodesic)* replaces the expansion with a radial basis function (RBF) kernel of geodesic electrode distance, with learnable width and amplitude. *REVE 4D encoding* ports the coupled spatio-temporal encoding of REVE [94] as published, which also replaces our temporal RoPE. Since that conflates spatial and temporal handling, we derive *REVE spatial encoding*, which imports only the 3D coordinate component and leaves RoPE in place, isolating the geometry comparison.

The results are summarized in Table 24. The Legendre prior is the strongest configuration on all three tasks, and REVE's encoding provides no transfer advantage in our backbone. Replacing REVE's joint 4D encoding with its spatial component while retaining RoPE improves performance on all three tasks, with the largest increase on ISRUC (0.593 to 0.691). At matched temporal handling, REVE's Fourier encoding is comparable to our linear coordinate embedding, and the improvement comes from adding the prior on top. The geodesic Gaussian fixes the kernel to a monotone decay in distance, whereas our expansion learns its profile end-to-end; the learned kernel is not monotone (Appendix F.2). The anisotropic SH prior also underperforms the isotropic Legendre prior (0.658 vs. 0.677 average), providing no empirical benefit from the added directional flexibility under the present setting.

Table 24: Geometry encoding ablations. All variants use the SPERA-Small backbone under the ablation protocol of Section 4.5. Results are mean balanced accuracy ± standard deviation over five random seeds; the best result in each column is in bold.

Geometry encoding	**PhysioNet-MI**	**ISRUC**	**TUAB**	**Average**
Anisotropic SH	0.501 ± 0.013	0.706 ± 0.017	0.767 ± 0.008	0.658 ± 0.013
Coordinate embedding only	0.482 ± 0.015	0.701 ± 0.027	0.774 ± 0.006	0.653 ± 0.010
REVE spatial encoding	0.492 ± 0.017	0.691 ± 0.021	0.761 ± 0.009	0.648 ± 0.009
Gaussian (geodesic)	0.451 ± 0.016	0.698 ± 0.023	0.771 ± 0.009	0.640 ± 0.010
REVE 4D encoding	0.445 ± 0.015	0.593 ± 0.028	0.712 ± 0.010	0.583 ± 0.017
Legendre prior (SPERA-Small)	$\mathbf{0.523 \pm 0.011}$	$\mathbf{0.721 \pm 0.024}$	$\mathbf{0.787 \pm 0.006}$	$\mathbf{0.677 \pm 0.011}$

F.2 Learned coefficients and angular bias profile

To examine what spatial structure is learned by the pairwise Legendre bias, we analyze the coefficients of the pretrained SPERA-Base model before downstream fine-tuning. The pairwise bias of Eq. (2) can be written as a rotation-invariant angular bias function

$$K(\theta) = \sum_{\ell=0}^{L} w_\ell P_\ell(\cos\theta), \tag{7}$$

where θ denotes the angular separation between two electrodes. Because $P_0(\cos\theta) = 1$, the degree-0 term contributes only an angle-independent offset. We therefore exclude it when analyzing the relative shape of the learned spatial kernel.

As shown in Table 25, the learned coefficients are concentrated at low Legendre degrees. Degrees 1–3 account for 93.1% of the summed squared coefficients over degrees 1–8, indicating that the learned bias is dominated by broad, smoothly varying angular structure. Evaluating $K(\theta)$ over angular separations from 0° to 180° using the learned coefficients shows that the bias reaches its maximum at approximately 20.0° and its minimum at 180°, assigning relatively higher bias to nearby electrode pairs than to widely separated or antipodal pairs. The nonzero location of the maximum also shows that this locality-oriented profile is not a simple monotonic decay with angular distance.

Table 25: Learned pairwise Legendre coefficients of the pretrained SPERA-Base model. Degree 0 is omitted because it contributes only an angle-independent offset.

Legendre degree	1	2	3	4	5	6	7	8
Coefficient	0.1332	0.0831	0.0415	0.0074	−0.0003	−0.0205	−0.0089	−0.0376

F.3 Decomposition of the anchor-based attention

Unlike the pairwise Legendre bias of Eq. (2), the anchor-based formulation injects the geometry feature $\boldsymbol{\phi}_c$ into the queries and keys. Expanding Eq. (4) therefore yields four score components:

$$\begin{aligned} s_{ij} &= \frac{1}{\sqrt{d}} \left(\mathbf{q}_i + \Gamma_Q \boldsymbol{\phi}_{c_i}\right)^\top \left(\mathbf{k}_j + \Gamma_K \boldsymbol{\phi}_{c_j}\right) \\ &= \frac{1}{\sqrt{d}} \Big(\underbrace{\mathbf{q}_i^\top \mathbf{k}_j}_{\text{content–content}} + \underbrace{\mathbf{q}_i^\top \Gamma_K \boldsymbol{\phi}_{c_j} + (\Gamma_Q \boldsymbol{\phi}_{c_i})^\top \mathbf{k}_j}_{\text{content–geometry}} + \underbrace{(\Gamma_Q \boldsymbol{\phi}_{c_i})^\top \Gamma_K \boldsymbol{\phi}_{c_j}}_{\text{geometry–geometry}} \Big). \end{aligned} \tag{8}$$

The geometry–geometry term captures pairwise interaction between the Legendre anchor encodings, while the two cross terms couple token content with electrode geometry. These cross terms have no counterpart in Eq. (2), where geometry enters only as an additive pairwise score bias.

We measure the contribution of each term to the pre-softmax logit using its root-mean-square magnitude over randomly sampled token pairs and attention heads in each full-attention block. The

two content–geometry cross terms together exceed the geometry–geometry term in every full-attention block of both SPERA-Base and SPERA-Large, by factors of 1.3 to 13. Content–content remains the largest single term except in the earliest full-attention block, where the two cross terms combined exceed it in both models.

F.4 Pairwise-bias and anchor-based implementations

The two implementations differ in how geometry enters attention and therefore in computational cost. In the pairwise Legendre bias form, the Legendre values $P_\ell(\mathbf{e}_i^\top \mathbf{e}_j)$ are computed once per forward pass and reused, so the overhead does not arise from evaluating the polynomials. Instead, adding the pairwise Legendre bias to the attention scores requires an $N \times N$ bias matrix and prevents the fused attention path used in our implementation. We test whether this overhead can be reduced using PyTorch's FlexAttention, whose `score_mod` applies the pairwise Legendre bias inside the fused kernel without materializing the full score matrix.

Table 26 compares three whole-model configurations of SPERA-Small. Throughput increases from 52,263 to 66,326 tok/s with FlexAttention, but remains about 29% below that of the anchor-based design. Since the anchor form costs only 0.005 average balanced accuracy relative to the pairwise bias (Table 3), we retain it in the full-attention blocks.

Table 26: Training throughput for three whole-model configurations of SPERA-Small, measured under the setting of Table 4. Only the treatment of the prior in full-attention blocks differs; factorized layers use the pairwise Legendre bias throughout.

Prior in full-attention blocks	**Throughput (tok/s)**
Pairwise Legendre bias	52,263
Pairwise Legendre bias with FlexAttention	66,326
Legendre anchor encoding (SPERA)	92,866

F.5 Extension to intracranial recordings

For scalp EEG, electrodes are distributed over the scalp surface, and we therefore represent their locations using unit-normalized directions. For iEEG and sEEG, where electrodes occupy the brain volume, radial position also becomes informative. The construction can be extended to the solid ball $\mathbb{B}^3$ while retaining the same rotation-invariant angular structure. On the solid ball, separable solutions of the Laplacian eigenproblem consist of a radial part given by spherical Bessel functions j_ℓ and an angular spherical-harmonic part. Using the corresponding rotation-invariant angular form yields

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{\ell,n} w_{\ell n}\, j_\ell(\kappa_n r_i)\, j_\ell(\kappa_n r_j)\, P_\ell(\mathbf{e}_i^\top \mathbf{e}_j), \tag{9}$$

where r_i and $\mathbf{e}_i$ are the radius and unit direction of electrode i, n indexes radial modes, κ_n denotes the radial wavenumber, and $w_{\ell n}$ are learnable coefficients. The angular Legendre structure is retained and radius enters through a learnable radial basis; the surface prior used in this paper is the special case of a shared radius.

Two practical obstacles remain: intracranial electrode positions require coregistration to a common anatomical space, such as the Montreal Neurological Institute (MNI) template, before a shared radial basis is meaningful, and the high-gamma content that carries much of the information in intracranial recordings lies outside our 0.5–75 Hz preprocessing pipeline, so the tokenizer would need revision. We leave this extension to future work.

F.6 Spatial attribution analysis

We examine whether the model's electrode-level attributions align with known EEG neurophysiological patterns. We use integrated gradients (IG) [141] on the fine-tuned SPERA-Base model, quantifying each electrode's contribution to the classification. We focus on motor imagery (MI), for which the relevant cortical regions and class-dependent spatial organization are well established [106].

IG is computed against a zero baseline on the test split. Per-class IG maps are strongly correlated (mean pairwise r = 0.68 on PhysioNet-MI, and 0.66 on BCIC-IV-2a), being dominated by a shared

Table 27: Class-relative attribution over the sensorimotor strip for BCIC-IV-2a and PhysioNet-MI. Each value is the class-relative IG after the two-step normalization. LI = (SM-right − SM-left) / (SM-right + SM-left); MedIdx = SM-mid / mean (SM-left, SM-right).

Dataset	Class	SM-left	SM-right	SM-mid	LI	MedIdx
BCIC-IV-2a	Left hand	0.953	1.095	0.959	+0.069	0.937
	Right hand	1.081	0.991	0.872	−0.044	0.841
	Feet	0.971	0.936	1.133	−0.018	1.187
	Tongue	0.994	0.978	1.036	−0.008	1.051
PhysioNet-MI	Left fist	0.942	1.123	0.980	+0.088	0.949
	Right fist	1.015	0.992	0.803	−0.012	0.800
	Both fists	0.988	0.915	1.088	−0.038	1.144
	Both feet	1.056	0.970	1.129	−0.042	1.115

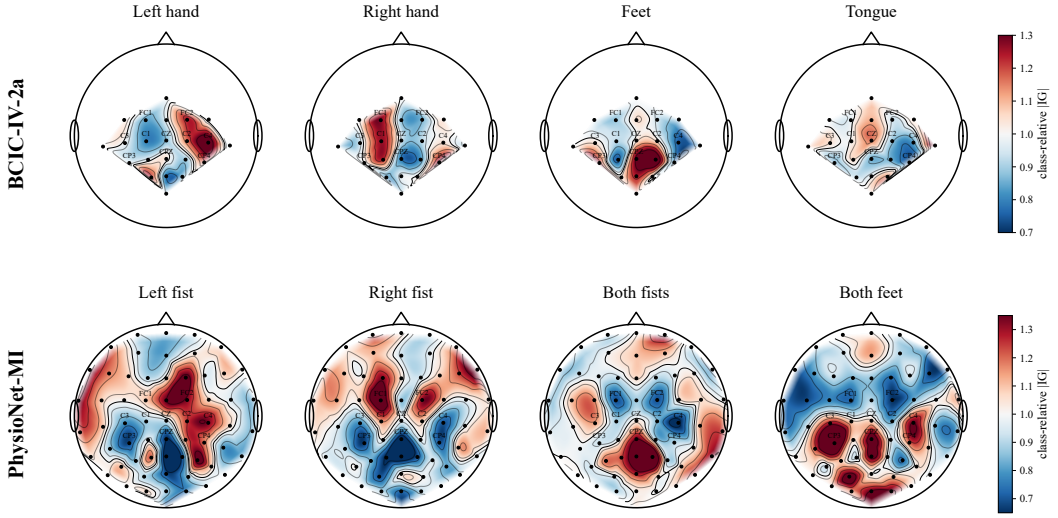

Figure 5: Class-relative attribution topographic maps on BCIC-IV-2a and PhysioNet-MI. Maps are computed separately for each class after two-step class-relative normalization.

channel-saliency profile that concentrates on the sensorimotor electrodes: mean absolute IG is 3.5–4.5 at C3, C4, CP3, and CP4 on PhysioNet-MI against 1.7–2.4 elsewhere, and 25–41 at CP3, CP4, FC3, and FC4 on BCIC-IV-2a against 15.7–17.5 in frontal and parietal regions. This shared profile is consistent with reliance on signals recorded over sensorimotor scalp regions. To highlight class-relative spatial differences, we apply a two-step normalization: each class map is divided by its mean over channels, and each channel is then divided by its mean over classes. In these normalized maps, 1.0 corresponds to the across-class mean for each channel. We summarize the class-relative maps over the sensorimotor strip (all FC*, C*, and CP* electrodes), split into left, right, and midline groups, using a lateralization index LI = (right − left) / (right + left) and a mediality index MedIdx = midline / mean (left, right).

On BCIC-IV-2a, the two unimanual classes follow the expected pattern (Table 27): LI is positive for left-hand and negative for right-hand imagery, and Figure 5 shows the corresponding increase over the sensorimotor region contralateral to the imagined hand, consistent with the established lateralization of hand motor imagery [104, 90]. The feet class has the highest mediality and a correspondingly central pattern around the midline, reflecting the medial cortical representation of the lower limbs [104, 101]. The tongue class has the smallest absolute LI of the four and is broadly distributed across the sensorimotor strip without the unilateral concentration seen for the hand classes, in line with reports of distributed parietofrontal involvement during tongue motor imagery [36].

PhysioNet-MI shows the same organization, though more weakly for the right-fist class: attribution shifts toward the right hemisphere for the left fist and only marginally toward the left hemisphere for

the right fist. Both bilateral classes have MedIdx above 1 while both unimanual classes fall below it, and the topographic maps place their attribution around central and midline sensorimotor sites.

G Few-shot experiments

To evaluate the data efficiency of SPERA, we conduct few-shot LP experiments on BCIC-IV-2a, PhysioNet-MI, and ISRUC (motor imagery and sleep staging). For each dataset, the encoder is frozen and only a linear classifier is trained using a limited subset of the original training split. We evaluate six training data ratios: 1%, 2%, 5%, 10%, 25%, and 50%. The subsets are sampled in a class-balanced manner based on the smallest class size in the training set. All experiments are repeated over five random seeds, and we report the mean balanced accuracy with standard deviation.

Across all three datasets, performance improves steadily as the training ratio increases (Figure 6). On BCIC-IV-2a, SPERA-Base surpasses the REVE-Base reference using 10% of the training data. PhysioNet-MI shows a more gradual trend, with the 50% setting closing the gap to the full-data SPERA-Base reference. On ISRUC, SPERA-Base exceeds the REVE-Base reference with only 2% of the training data and approaches the full-data SPERA-Base reference at higher ratios. These results indicate that SPERA provides data-efficient EEG representations.

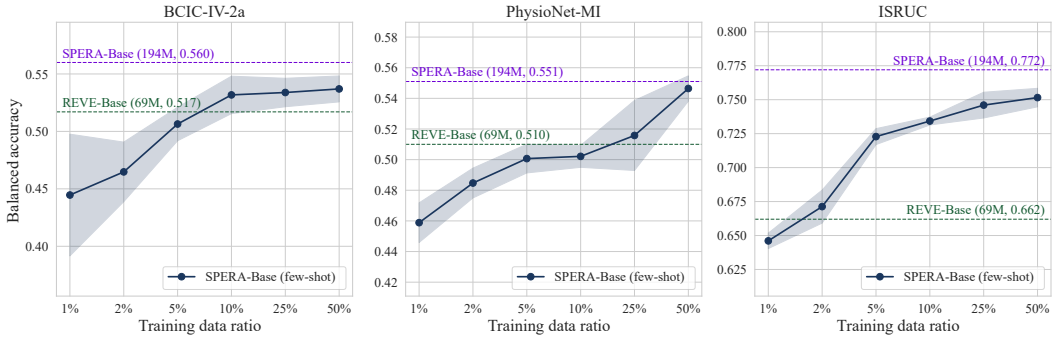

Figure 6: Few-shot LP results of SPERA-Base on BCIC-IV-2a, PhysioNet-MI, and ISRUC. The blue curve shows the mean balanced accuracy of SPERA-Base under limited labeled training data, with the shaded area indicating one standard deviation across five random seeds. Dashed horizontal lines indicate the full-data LP results of SPERA-Base and REVE-Base.

H Noise robustness analysis

We evaluate noise robustness under LP using SPERA-Small, CBraMod, and REVE-Base. To isolate the effect of the pretraining objective, we additionally include the matched reconstruction variant from Appendix E.1, which differs only in the prediction target. Since CBraMod and REVE do not report noise robustness results, we re-evaluate their publicly released checkpoints together with the noise-free condition under the same protocol. For each checkpoint, the encoder is frozen and the linear classifier and feature-normalization statistics are fitted only on unperturbed training features and then kept fixed across test conditions. We consider additive Gaussian noise and injected ocular artifacts separately.

H.1 Gaussian noise

Gaussian noise robustness is evaluated on TUAB, ISRUC, and PhysioNet-MI. Additive Gaussian white noise is applied only to the test inputs under five conditions: noise-free, 10 dB, 5 dB, 0 dB, and -5 dB. Noise variance is determined independently for each channel from its mean squared signal amplitude to match the target signal-to-noise ratio (SNR).

As shown in Figure 7, SPERA-Small generally maintains higher mean balanced accuracy than the comparison models across the evaluated noise conditions. The matched JEPA and reconstruction variants show broadly similar degradation patterns, while JEPA maintains higher mean accuracy in

nearly all evaluated conditions. Thus, the accuracy advantage of latent prediction generally persists under additive Gaussian corruption.

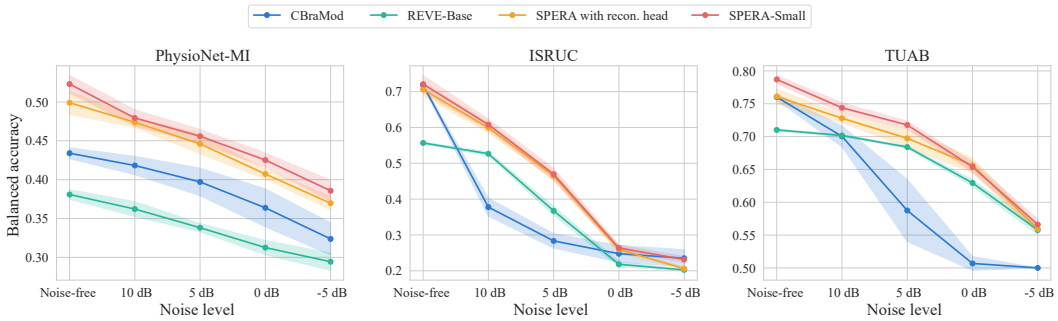

Figure 7: Gaussian noise robustness on PhysioNet-MI, ISRUC, and TUAB under LP. Frozen encoders with linear classifiers trained on unperturbed data are evaluated on test inputs with additive Gaussian noise at 10, 5, 0, and -5 dB. Shaded regions denote one standard deviation across five random seeds.

H.2 Physiological ocular artifacts

We further evaluate robustness to ocular artifacts on PhysioNet-MI and TUAB using electrooculographic (EOG) waveforms from EEGdenoiseNet [160]; ISRUC is excluded because the ocular artifact injection protocol is not appropriate for sleep EEG. Ocular artifacts are projected onto each montage using the empirical propagation factors reported by Lins et al. [76]. For TUAB, bipolar-channel factors are obtained by subtracting those of the constituent electrodes; for PhysioNet-MI, factors are estimated by three-dimensional thin-plate spline interpolation with a linear term, enforcing symmetry for the vertical field and antisymmetry for the horizontal field. For each trial, independently sampled vertical and horizontal EOG waveforms are normalized, projected using these spatial fields, and combined using the reported horizontal-to-vertical saccade amplitude ratio. A shared intermittent envelope leaves at least 30% of each trial without added artifacts, after which the combined artifact is scaled to the target SNR based on trial-level mean EEG and artifact power across channels and time.

Figure 8 shows task-dependent robustness patterns under injected ocular artifacts. On TUAB, SPERA-Small achieves the highest mean balanced accuracy across all conditions, whereas CBraMod performs best under the injected-artifact conditions on PhysioNet-MI. In the matched objective comparison,

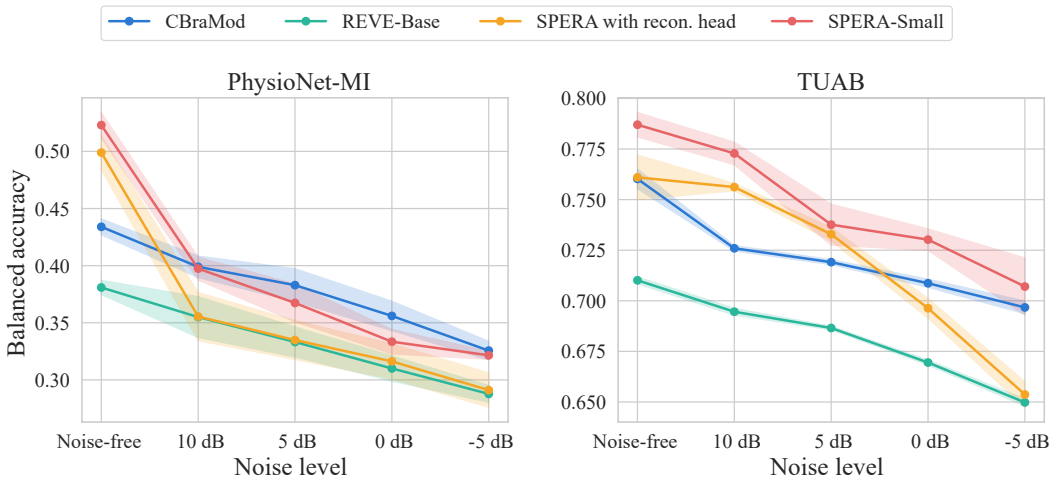

Figure 8: Ocular-artifact robustness on PhysioNet-MI and TUAB under LP. Frozen encoders with linear classifiers trained on unperturbed data are evaluated on test inputs with injected EEGdenoiseNet EOG artifacts at 10, 5, 0, and -5 dB. Shaded regions denote one standard deviation across five random seeds.

JEPA remains above reconstruction under every condition on both datasets. From noise-free to -5 dB, the decrease is smaller for JEPA on PhysioNet-MI (0.201 vs. 0.208) and TUAB (0.080 vs. 0.107), with the larger difference observed on TUAB.

I Representation analysis

I.1 UMAP visualization of learned representations

To qualitatively examine the structure of learned representations, we apply uniform manifold approximation and projection (UMAP) [84] to PhysioNet-MI, ISRUC, and TUAB. For each dataset, we visualize three representation levels: (i) raw EEG samples, (ii) features from the frozen pretrained encoder, and (iii) features after task-specific fine-tuning. The results are shown in Figure 9.

The first column shows that raw EEG samples exhibit no clear class-wise structure across all three datasets. The second column reveals that the frozen pretrained features already capture task-related structure to varying degrees. On ISRUC, sleep stages exhibit clearly distinguishable clusters even without task-specific supervision, with Wake, N3, and REM occupying separate areas of the embedding space. TUAB shows a weaker but visible separation between normal and abnormal recordings, while PhysioNet-MI exhibits more limited class-level structure at this stage. The third column shows that, after fine-tuning, all three datasets form clearer and more separable clusters.

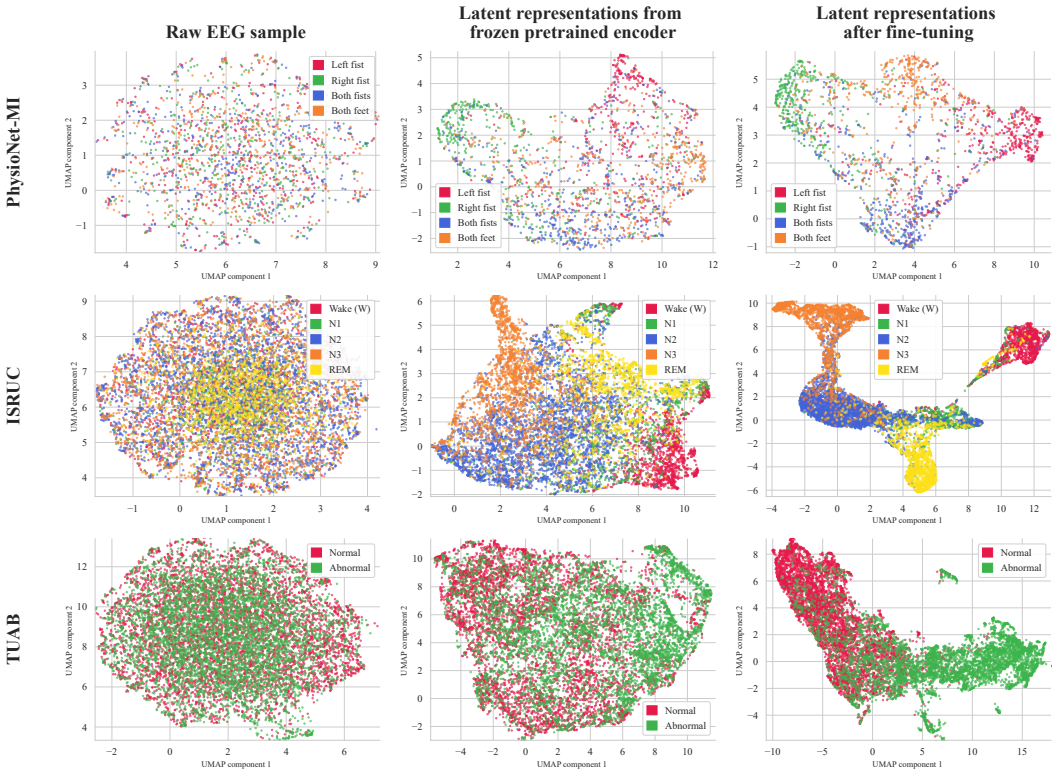

Figure 9: UMAP visualization of representations from the SPERA-Base model on three downstream tasks. Columns correspond to raw EEG, frozen pretrained features, and fine-tuned features. Rows correspond to PhysioNet-MI, ISRUC, and TUAB.

I.2 Linear accessibility and encoder adaptation

In terms of average balanced accuracy across the nine tasks, SPERA-Base leads REVE-Base by 0.051 under LP and by 0.035 under LPFT (Tables 2 and 17). At the aggregate level, SPERA-Base remains ahead under both protocols, although several task-level differences vary with model scale and adaptation. Under LP, REVE-Base leads SPERA-Base on MAT, while REVE-Large leads

SPERA-Large on BCIC-IV-2a, PhysioNet-MI, and Mumtaz. BCIC-IV-2a and PhysioNet-MI are motor-imagery tasks that involve frequency-specific sensorimotor activity [105], while MAT has been used to study spectral changes during mental arithmetic [164], motivating us to examine the linear accessibility of band-power information.

We regress channel-wise relative power in five frequency bands (delta to gamma) from frozen encoder representations using a linear readout on per-channel token features. Table 28 reports held-out macro Pearson's r, averaged over channel–band targets. Because SPERA and REVE differ in architecture, tokenization, corpus, and pretraining objective, we additionally evaluate the matched reconstruction variant of SPERA-Small from Appendix E.1.

Table 28: Band-power decodability from frozen representations. Values are held-out macro Pearson's r, averaged over channel–band targets.

Method	**BCIC-IV-2a**	**PhysioNet-MI**	**MAT**	**Average**
SPERA-Small	0.395	0.482	0.305	0.394
Reconstruction head	0.592	0.640	0.483	0.572
REVE-Base	0.735	0.680	0.589	0.668

Across all three tasks, band power is more linearly decodable from the matched reconstruction variant than from SPERA-Small, with REVE-Base showing the highest values. On PhysioNet-MI, however, the same objective swap increases band-power decodability from 0.482 to 0.640 while lowering LP accuracy from 0.523 to 0.499 (Appendix E.1), indicating that greater band-power accessibility alone does not explain the downstream accuracy difference.

To further investigate the LP–LPFT divergence, we test how much of the performance gap can be recovered by limited encoder adaptation. Unfreezing only the last of SPERA-Base's 16 encoder blocks after LP recovers 49%, 71%, and 61% of the LP–LPFT gain on BCIC-IV-2a (0.560 to 0.619), PhysioNet-MI (0.551 to 0.637), and MAT (0.672 to 0.736), respectively. These results suggest that a substantial portion of the LP–LPFT gap can be recovered through limited encoder adaptation.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and Section 1 describe SPERA as an EEG representation-learning framework based on latent-space prediction, with a Legendre-polynomial spherical prior for electrode geometry, a hybrid attention backbone, and a relational spectral regularizer. The empirical claims regarding downstream performance, scaling and parameter efficiency, and robustness are supported by Sections 4.3–4.5 and Appendices D–I.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 5 discusses limitations regarding (i) scaling beyond 396M parameters, (ii) the lack of systematic disentanglement of corpus size, model capacity, and architectural priors in the observed scaling behavior, (iii) the surface-electrode assumption of the spherical prior, (iv) evaluation restricted to classification over fixed-duration EEG segments of 3–30 seconds, and (v) incomplete demographic metadata for comprehensive fairness analysis.

   Guidelines:

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

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

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

   Justification: The paper presents an empirical contribution and does not include theoretical results such as theorems, propositions, or lemmas requiring formal proof.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Section 3 specifies the model architecture, training objectives, masking strategy, and augmentation, while Sections 4.1–4.2 describe the pretraining corpus, preprocessing, and downstream evaluation. Appendices A–D provide computing resources, corpus information, hyperparameters, and dataset-specific evaluation details, and Appendices E–I document the protocols for the additional ablations and analyses.

   Guidelines:

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

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

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

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

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

   Answer: [Yes]

   Justification: Our codebase is provided as a link in the paper (`https://github.com/sees4321/SPERA_EEG`), including training, evaluation, and preprocessing scripts together with a README detailing the environment setup and exact commands to reproduce the main results. All pretraining and downstream datasets are publicly available, with sources and licenses listed in Appendix B.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Section 4.1 specifies the optimizer and pretraining setup, and Section 4.2 describes SPERA's LP and LPFT protocols, validation-based hyperparameter selection, and baseline evaluation settings. Appendix C.1 provides the full hyperparameters across all model sizes, while Appendix D documents the dataset-specific preprocessing and train/validation/test splits, including implementation-specific details.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: SPERA experiments are repeated over five random seeds and reported as mean $\pm$ one standard deviation; for the aggregate results, variability is computed across the corresponding seed-wise averages. Appendix D.10 additionally reports unpaired Welch's t-tests with 95% confidence intervals and Holm correction across the nine tasks separately for each SPERA model, while figure captions specify shaded one-standard-deviation bands where applicable.

   Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Appendix A details the computing environment, including the hardware (HP Z8 workstation with two NVIDIA RTX 5000 Ada GPUs), software stack (PyTorch 2.10.0, CUDA 13.0), and training efficiency strategies (DDP and BF16 mixed precision). Pretraining the largest model (SPERA-Large, 396M) takes approximately 287 GPU hours. Section 4.5 and Table 4 additionally report throughput and peak GPU memory for the attention and Legendre prior variants. Additional experiments and ablations were conducted primarily with SPERA-Small (48M), including the matched objective and corpus controls, and therefore required substantially less compute per run than the largest-model pretraining.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and confirm that this research conforms to it in all respects.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: SPERA can support clinical applications such as sleep staging, abnormal-event detection, and mental disorder screening, as well as assistive BCI applications including motor imagery and imagined speech decoding, particularly in settings with limited labeled data. On the negative side, EEG-based models could be misused for unconsented brain-state inference or surveillance, and premature clinical deployment without additional validation could lead to incorrect diagnostic decisions. Fairness considerations also arise from potential demographic imbalance in the pretraining corpus, as noted in Section 5. We therefore caution against direct clinical or consumer deployment without further validation, fairness audits, and appropriate human oversight.

    Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: The released model is a representation encoder for EEG signals and does not generate content; it does not pose the high-risk misuse concerns associated with generative language or image models. The pretraining corpus consists of EEG recordings from established public repositories (e.g., OpenNeuro, MOABB, PhysioNet, NSRR, TUH) under their respective data use agreements, with no scraped or unverified content (Appendix B).

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: All baseline models, prior EEG foundation models, pretraining datasets, and downstream datasets are properly cited at first mention (Sections 4.1 and 4.2; Appendix B). Appendix B (Table 6) lists each of the 106 pretraining datasets together with its source platform and license (e.g., CC0, CC BY 4.0, CC BY-NC-SA 4.0, GPL-3.0, and the NEDC, NSRR, Brainclinics, and BDSP data use agreements). The software stack is reported with version numbers in Appendix A (PyTorch 2.10.0, transformers 5.3.0, accelerate 1.13.0, MNE-Python 1.11.0). All assets are used in compliance with their respective licenses and terms of use.

Guidelines:

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

13. **New assets**

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

Answer: [Yes]

Justification: We release SPERA, a new pretrained EEG foundation model, together with training, evaluation, and preprocessing code. The architecture, training procedure, hyperparameters, and computing environment are documented in Section 3, Section 4.1, and Appendices A–C.1.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

    Answer: [N/A]

    Justification: This work does not involve any crowdsourcing or new data collection from human subjects. All EEG data are obtained from existing public repositories under their respective data use agreements (Appendix B).

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: This work does not involve new research with human subjects. All EEG recordings come from existing public datasets whose original collection was approved by the relevant institutional review boards or equivalent oversight bodies, as documented in their respective source publications (Appendix B).

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: LLMs were not used as a component of the core methodology of this research.

Guidelines:

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