PRiSE-EEG: A Prior-Guided Foundation Model with Depth-Stratified Experts for Cross-Paradigm EEG Representation Learning

Anonymous authors
Paper under double-blind review

Abstract

EEG foundation models aim to learn reusable representations across heterogeneous paradigms, yet existing approaches often use uniform adaptation mechanisms and are typically reported under separate downstream fine-tuning protocols. In this work, we first analyze dense EEG Transformers from two complementary perspectives. Gradient similarity across six downstream datasets reveals substantial optimization conflicts among EEG paradigms, while CKA analysis on mixed-paradigm batches shows a consistent depth-wise transition: shallow layers preserve stronger cross-paradigm similarity, whereas deeper layers become increasingly specialized. Motivated by these findings, we propose **PRiSE-EEG**, a prior-guided EEG foundation model with CKA-calibrated Depth-Stratified Experts. PRiSE-EEG forms continuous multi-channel EEG patches using weak static cortical and network priors and dynamic short-time channel interactions, then allocates shared and specialized experts across MoE Transformer blocks according to a sigmoid mapping from layer-wise CKA sharedness. This design preserves common EEG regularities in early blocks while assigning more specialized capacity to later task-specific transformations. Experiments on 12 public EEG benchmarks show strong performance. Compact ablations further show that CKA-derived expert allocation improves over dense Transformers, uniform MoE, and manually fixed shared-specific expert ratios. Code is available in an anonymous repository.

1 Introduction

Electroencephalography (EEG) signals differ substantially in montage, temporal statistics, dominant rhythms, label semantics, and task-relevant spatial patterns. Classical models for EEG suffer from subject-, session-, and dataset-specific variability (Lawhern et al., 2018; Song et al., 2022; Demir et al., 2021). EEG foundation models (EEG-FM) aim to learn reusable representations across heterogeneous paradigms, yet cross-paradigm transfer remains difficult. Therefore, an EEG-FM must preserve general EEG regularities while still adapting to task-specific evidence. Existing models usually address this problem either by improving pretraining and tokenization, or by adding adaptation mechanisms during fine-tuning. Recent works such as BENDR (Kostas et al., 2021), BIOT (Yang et al., 2023), EEGPT (Wang et al., 2024), CBraMod (Wang et al., 2025), CSBrain (Zhou et al., 2025) improve scalable self-supervised pretraining and heterogeneous modeling. Discrete-token approaches, including LaBraM (Jiang et al., 2024), CodeBrain (Ma et al., 2025) and TFM-Tokenizer (Pradeepkumar et al., 2025), further show that how EEG is converted into discrete tokens is critical for representation quality. In parallel, EEGMoE (Gao et al., 2026) demonstrates that domain-specific experts can help decouple heterogeneous EEG domains in pre-training stage. However, two issues remain insufficiently explored. First, existing EEG-FMs rarely model cross-task relationships, although heterogeneous paradigms may induce different update directions. Second, spatial structure beyond electrode coordinates or montage embeddings remains underused for patch formation.

Although sensor-space EEG does not provide direct source-level measurements, it still preserves structured spatial regularities that are useful for representation learning. Electrophysiological studies show that scalp recordings contain coarse network-level organization (Brookes et al., 2011; Hipp et al., 2012; Xavier et al., 2025), while EEG microstates and lag-sensitive coupling analyses suggest that multi-channel interactions can change within sub-100 ms windows (Vicente et al., 2008; Michel

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

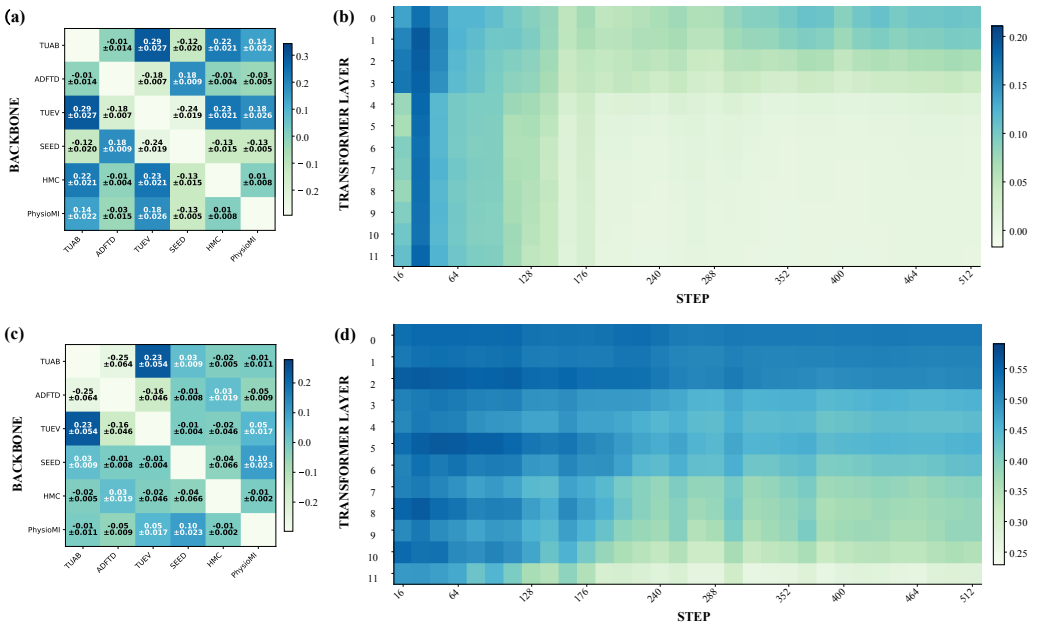


Figure 1: Pairwise cosine similarity of downstream fine-tuning gradients across datasets for LaBraM (a) and CSBrain (c), respectively. Lower similarities indicate task-dependent optimization conflicts. Layer-wise sample CKA across pretraining steps for LaBraM (b) and CSBrain (d), respectively. Shallow layers show stronger cross-paradigm alignment, whereas deeper layers become more specialized, motivating CKA-calibrated depth-stratified expert allocation.

& Koenig, 2018; Van de Ville et al., 2010; Mehra et al., 2025). These findings suggest that more sensor-space prior and short-window dynamics can provide weak guidance for EEG learning.

We explore the task-relation issue afterwards by examining whether heterogeneous EEG tasks impose compatible updates at fine-tuning stage, as shown in Fig. 1. We train dense Transformer baselines on six downstream datasets and compute pairwise gradient similarities. The resulting heatmaps show clear optimization-direction discrepancies across tasks, suggesting that fully shared dense parameters may face conflicting update preferences across paradigms (Sener & Koltun, 2018). This motivates conditional capacity through MoE, where different inputs can activate different expert parameters. However, a uniform MoE allocation still ignores where shared and specialized capacity should appear across model depth. We therefore examine the representation geometry of a dense pilot EEG Transformer using Centered Kernel Alignment (CKA) (Kornblith et al., 2019). Given mixed-paradigm batches, we compute block-wise CKA across predefined pretraining checkpoints, with within-paradigm and between-paradigm similarities averaged separately. The analysis shows a consistent depth-wise pattern: early blocks retain higher cross-paradigm similarity, while later blocks become increasingly task- or paradigm-specific. This suggests that EEG representations follow a general-to-specific transition across the pretraining Transformer stack, and that shared and specialized experts should not be allocated uniformly across all MoE blocks.

Guided by these empirical observations, we propose PRiSE-EEG, a Prior-guided EEG foundation model with depth-Stratified Experts. PRiSE-EEG forms continuous multi-channel EEG patches by combining weak static region and network priors with dynamic short-window channel interactions. On top of these patches, PRiSE-EEG introduces a CKA-calibrated expert allocation rule. For each MoE Transformer block, we estimate the ratio between between-paradigm and within-paradigm CKA, and map it to the shared expert proportion. Blocks with higher sharedness receive more shared experts, while blocks with lower sharedness allocate more capacity to specialized experts.

We evaluate PRiSE-EEG on 12 public EEG benchmarks covering sleep staging, motor imagery, emotion recognition, and other cognitive or clinical tasks. In evaluations across multiple EEG paradigms, PRiSE-EEG outperforms dense Transformer baselines, uniform Mixture-of-Expert(MoE) variants, and fixed-ratio shared-specific expert allocations. These results show that expert specialization in

EEG foundation models should not only be input-adaptive, but also depth-stratified according to the representation structure learned by dense pilot models. Our contributions are summarized as follows:

- We characterize two issues in cross-paradigm EEG learning, including conflicting task updates across layers and underused weak sensor-space structure.
- We propose CKA-calibrated Depth-Stratified Experts, which convert block-wise CKA sharedness into shared expert proportions through a simple sigmoid allocation rule.
- We introduce PRiSE-EEG, combining prior-guided continuous tokenization with depth-stratified expert allocation, and validate it on 12 EEG benchmarks.

2 Related Works

2.1 EEG Foundation Models and Tokenization

Early convolutional, graph-based and attention-based models improve EEG spatial-temporal feature extraction but are commonly optimized for specific subjects or datasets (Lawhern et al., 2018; Song et al., 2022; Demir et al., 2021; Roy et al., 2019; Song et al., 2021). Recent EEG foundation models learn transferable representations through self-supervised pretraining on heterogeneous EEG corpora, including BENDR (Kostas et al., 2021), BIOT (Yang et al., 2023), EEGPT (Wang et al., 2024), CBraMod (Wang et al., 2025), CSBrain (Zhou et al., 2025), and REVE (Ouahidi et al., 2025). Tokenization has also become central: LaBraM (Jiang et al., 2024) learns vector-quantized neural codes, CodeBrain (Ma et al., 2025) decouples temporal and frequency tokens, and TFM-Tokenizer (Pradeepkumar et al., 2025) learns single-channel time-frequency motifs. MMM organizes EEG spatial information through a unified topology and a multi-level channel hierarchy (Yi et al., 2023). PRiSE-EEG combines learnable static spatial biases with sample-dependent short-time dynamic groups to form continuous multi-channel representations, and further uses layer-wise representation sharedness to allocate shared and specialized experts.

2.2 Conditional Experts and Depth-wise Representation Analysis

MoE models provide scalable conditional computation through sparsely gated experts (Shazeer et al., 2017; Fedus et al., 2022), and have also been used in neuroimaging for heterogeneous aging modeling (Eavani et al., 2016) and cross-subject fMRI decoding (Wei et al., 2025). In EEG representation learning, EEGMoE (Gao et al., 2026) introduces shared and specific expert groups to decouple domain-specific representations during self-supervised pretraining. Shared structure in MoE can also be configured through RSA-based task affinities (Vandenhende et al., 2020), learned task-expert specialization (Chen et al., 2023), compute-constrained architecture search (Jawahar et al., 2023), or GLU activation statistics in dense language models (Zhao et al., 2026). SVCCA (Raghu et al., 2017), PWCCA (Morcos et al., 2018), and CKA (Kornblith et al., 2019) compare neural representations across layers, models, and training stages, while studies in language, vision, and multimodal models show that different depths can encode different information (Tenney et al., 2019; Mischler et al., 2024; Zhang et al., 2025). PRiSE-EEG uses cross-paradigm CKA as an architureture calibration signal to set layer-wise shared-expert counts before training. This makes representation similarity a structural prior rather than only a post-hoc diagnostic.

2.3 Weak Sensor-Space Priors for EEG

EEG sensor-space signals do not directly provide source-level functional networks. Nevertheless, electrophysiological recordings still exhibit network-level organization (Brookes et al., 2011; Hipp et al., 2012; Xavier et al., 2025) and rapid short-window interactions relevant to dynamic brain states (Vicente et al., 2008; Michel & Koenig, 2018; Van de Ville et al., 2010; Mehra et al., 2025). PRiSE-EEG therefore treats static region-network groupings and short-window channel interactions as weak inductive biases rather than hard anatomical assignments. To avoid making these priors overly rigid, we add learnable bias terms that can relax or adjust the predefined sensor-group associations during training. This design preserves coarse spatial guidance while allowing data-driven correction for task-, dataset-, and montage-specific deviations.

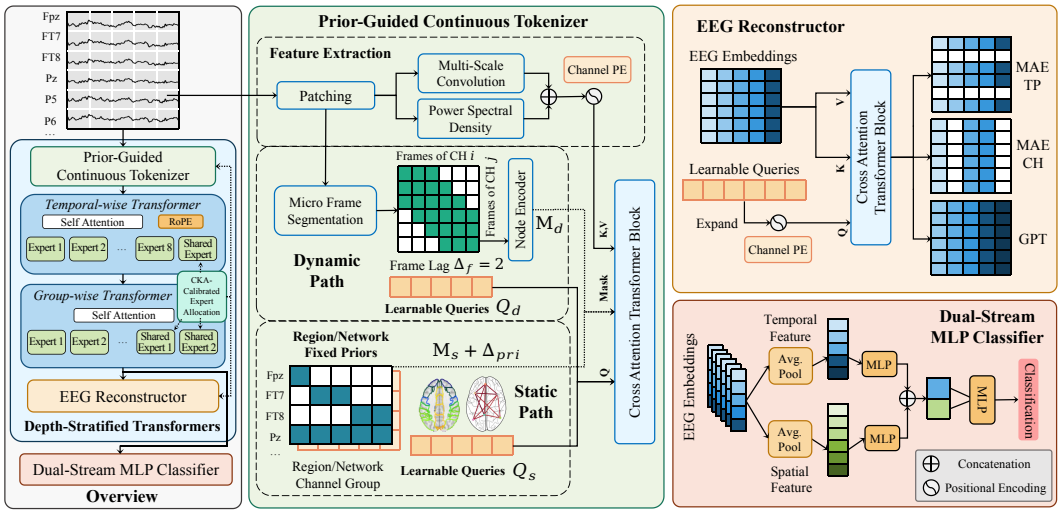


Figure 2: The architecture of PRiSE-EEG framework. The Prior-Guided Continuous Tokenizer exploits weak sensor-space prior to extract EEG features and CKA-Calibrated Depth-Stratified Transformers refine expert allocations. The EEG Reconstructor generates the original EEG signal with three pre-train tasks (MAE-TP, MAE-CH, and GPT), while Dual-Stream MLP Classifiers are applied on both spatial and temporal demension for downstream classification.

3 METHODS

3.1 OVERVIEW

PRiSE-EEG consists of a prior-guided continuous patch tokenizer and a CKA-calibrated depth-stratified expert transformer. Given an EEG batch $\mathbf{x}_p \in \mathbb{R}^{B\times T\times C\times P}$, where B is the batch size, T is the number of 1-second patches, C is the number of channels, and P is the number of sample points, the tokenizer maps $\mathbf{x}_p$ into continuous group representations $\mathbf{h}_0 \in \mathbb{R}^{B\times T\times G\times D_m}$. Here G is the number of static and dynamic groups, and D_m is the model dimension. PRiSE-EEG applies depth-stratified MoE blocks across the tokenizer, temporal/group encoder, and pretraining decoder, with shared-expert ratios determined by block-wise CKA sharedness from a dense pilot model. Unless otherwise specified, we report results of PRiSE-EEG-B. **More details are in Appx. B and D.6**.

3.2 PRIOR-GUIDED CONTINUOUS TOKENIZATION

We first embed each EEG patch into channel-level representations. Let $\mathbf{x} \in \mathbb{R}^{B\times T\times C\times D_m}$ denote the embedded patch features after temporal-spectral feature extraction and channel positional encoding.

$$\mathbf{x} = \mathrm{Embed}(\mathbf{x}_p) + \mathbf{E}_c, \tag{1}$$

where $\mathbf{E}_c$ is learnable positional encoding. The implementation of feature extraction, channel standardization, and positional encoding is described as Eq. 11 and 12 in Appendix B.2.1.

To inject static sensor-space priors, we define coarse region-network groups $\mathcal{P}$. Let $Q_s \in \mathbb{R}^{|\mathcal{P}|\times D_m}$ be learnable static group queries and $\mathbf{M}_s \in \mathbb{R}^{|\mathcal{P}|\times C}$ the corresponding static attention-bias matrix. Since these predefined groups are weak priors rather than hard anatomical assignments, we introduce a learnable relaxation term $\mathbf{\Delta}_{\mathrm{pri}} \in \mathbb{R}^{|\mathcal{P}|\times C}$. Static group attention is computed as

$$\mathrm{Attn}(Q_s, K, V) = \mathrm{Softmax}\left(\frac{Q_s K^\top}{\sqrt{d_k}} + \mathbf{M}_s + \mathbf{\Delta}_{\mathrm{pri}}\right) V, \tag{2}$$

where $K, V \in \mathbb{R}^{(BT)\times C\times D_m}$ are projected from $\mathbf{x}$, and d_k is the attention head dimension. Static groups provide stable coarse structure, while dynamic groups capture sample-dependent micro-frame interactions. We denote the dynamic bias construction as $\mathbf{M}_d = \mathrm{DynBias}(\mathbf{x}_p; \Delta_f, k_d)$, where Δ_f

is the maximum micro-frame lag and k_d is the number of selected dynamic groups. In practice, $\mathrm{DynBias}(\cdot)$ subdivides each patch into micro-frames, computes lag-limited channel affinities, and selects informative frame-pair groups. Let $Q_d \in \mathbb{R}^{k_d \times D_m}$ be the dynamic group queries. The final query and bias sets are $\mathbf{Q} = [Q_s; Q_d]$ and $\mathbf{M} = [\mathbf{M}_s + \mathbf{\Delta}_{\mathrm{pri}}; \mathbf{M}_d]$. The tokenizer output is

$$\mathbf{h}_0 = \mathrm{CrossAttn}(\mathbf{Q}, \mathbf{x}; \mathbf{M}) \in \mathbb{R}^{B \times T \times G \times D_m}, \tag{3}$$

where $G = |\mathcal{P}| + k_d$. Thus, PRiSE-EEG forms continuous multi-channel group tokens that handle montage heterogeneity while injecting priors. Details are described in Appx. B.2.2 and B.2.3.

3.3 CKA-Calibrated Depth-Stratified Expert MoE Transformer

The CKA-calibrated design is applied to all MoE Transformer layers, where each feed-forward block is replaced by a depth-stratified expert layer. Let $\mathbf{h}_l \in \mathbb{R}^{B \times T \times G \times D_m}$ to be input of l-th layer.

Interleaved temporal and group layers. The encoder uses L_E alternating Transformer layers: $L_E/2$ temporal-wise layers and $L_E/2$ group-wise layers (Bertasius et al., 2021). Temporal-wise attention models dependencies across the patch dimension T within each group, while group-wise attention models dependencies across the prior-guided group dimension G within each time patch. Each layer, regardless of its attention type, is followed by an independent depth-stratified MoE FFN:

$$\tilde{\mathbf{h}}_l = \mathrm{Attn}_l^{a_l}(\mathbf{h}_l), \qquad \mathbf{h}_{l+1} = \tilde{\mathbf{h}}_l + \mathrm{DSE}_l(\mathrm{LN}(\tilde{\mathbf{h}}_l)), \tag{4}$$

where $a_l \in \{\mathrm{temp}, \mathrm{group}\}$ specifies the layer type. This factorized design separates temporal and group-wise interactions while coupling them through alternating layers.

CKA-Calibrated Depth-Stratified Expert Allocation. Before training PRiSE-EEG, we run a dense pilot version with the same non-MoE layout and compute CKA on paradigm batches at selected checkpoints $\mathcal{T}$. For layer l in pretraining, let $c_{l,t}^{\mathrm{intra}}$ and $c_{l,t}^{\mathrm{inter}}$ denote the within-paradigm and between-paradigm CKA averages at checkpoint t. We define the layer sharedness as

$$s_{l,t} = \mathrm{clip}\left(\frac{c_{l,t}^{\mathrm{inter}}}{c_{l,t}^{\mathrm{intra}} + \epsilon}, 0, 1\right), \qquad \bar{s}_l = \frac{1}{|\mathcal{T}|}\sum_{t \in \mathcal{T}} s_{l,t}, \tag{5}$$

where ϵ is a small constant. We convert $\bar{s}_l$ into the shared expert proportion by

$$\rho_l^{\mathrm{sh}} = \rho_{\min} + (\rho_{\max} - \rho_{\min})\sigma\left(\frac{\bar{s}_l - \tau}{T_{\mathrm{cka}}}\right), \tag{6}$$

where $\rho_{\min}$ and $\rho_{\max}$ bound the shared-expert proportion, τ is the sharedness threshold, and T_{cka} is the temperature. Each layer has $E_l = 9$ experts: one fixed shared expert and eight allocated by CKA. Both E_l^{sh} and ρ_l^{sh} include the fixed expert. We set

$$E_l^{\mathrm{sh}} = \mathrm{round}(E_l \rho_l^{\mathrm{sh}}), \qquad E_l^{\mathrm{sp}} = E_l - E_l^{\mathrm{sh}}. \tag{7}$$

Let $\mathcal{E}_l^{\mathrm{sh}}$ denote all shared experts and $\mathcal{F}_{l,e}$ denote expert e. The expert block is

$$\mathrm{DSE}_l(\mathbf{h}_l) = \sum_{e \in \mathcal{E}_l^{\mathrm{sh}}} \mathcal{F}_{l,e}(\mathbf{h}_l) + \sum_{e \in \mathcal{K}_l(\mathbf{h}_l)} p_{l,e}(\mathbf{h}_l)\mathcal{F}_{l,e}(\mathbf{h}_l). \tag{8}$$

All shared experts process every token; $\mathcal{K}_l(\mathbf{h}_l)$ selects the top-2 specialized experts, and $p_{l,e}$ is the corresponding routing weight. More details of MoE layer are in Appx. B.3. The calibrated allocation is fixed during training; specialized routing remains token-dependent.

3.4 Stage-Specific Decoder

PRiSE-EEG uses different lightweight output modules for pretraining and downstream adaptation.

Pretraining Stage. An EEG reconstructor maps the encoder context $\mathbf{h}_{enc} \in \mathbb{R}^{B \times T \times G \times D_m}$ back to the channel space with learnable channel queries and Transformer decoding layers. It reconstructs both waveform and spectral targets in $\mathbf{x}_{\mathrm{ref}}$, supporting masked temporal-patch reconstruction, masked channel reconstruction, and future-patch prediction.

Fine-tuning Stage. We attach task-specific dual-stream MLP heads. One branch averages over group tokens to obtain temporal features, while the other averages over temporal patches to obtain group-level features. The two features are concatenated, normalized, and mapped to task logits. This head preserves both temporal and group-wise structure with lower cost than fully flattened classifiers.

3.5 PRETRAINING AND DOWNSTREAM ADAPTATION

During pretraining, PRiSE-EEG optimizes 3 self-supervised and 1 expert balancing loss functions:

$$\mathcal{L}_{pt} = \frac{1}{3}\left(\mathcal{L}_{\text{MAE-TP}} + \mathcal{L}_{\text{MAE-CH}} + \mathcal{L}_{\text{GPT}}\right) + \beta\mathcal{L}_{\text{Aux}}. \tag{9}$$

where $\mathcal{L}_{\text{MAE-TP}}$ reconstructs masked temporal patches, $\mathcal{L}_{\text{MAE-CH}}$ reconstructs masked channels, $\mathcal{L}_{\text{GPT}}$ predicts future patches, $\mathcal{L}_{\text{aux}}$ is the expert load-balancing loss, and β is its coefficient.

For downstream adaptation, PRiSE-EEG uses unified multi-dataset fine-tuning by default. One shared backbone is optimized across K downstream datasets with task-specific heads:

$$\mathcal{L}_{\text{ft}} = \sum_{k=1}^{K} \mathcal{L}_{\text{cls}}^{k} + \beta\mathcal{L}_{\text{aux}}, \tag{10}$$

where $\mathcal{L}_{\text{cls}}^{k}$ is the weighted cross-entropy loss for dataset k, and $\mathcal{L}_{\text{aux}}$ balances expert usage. Detailed adaptation protocols are given in Appendix B.5.1 and B.5.2.

4 EXPERIMENTS

4.1 DATASETS

Pre-training Datasets. To learn robust and generalizable representations, PRiSE-EEG is pre-trained on a diverse corpus combining 17 public datasets, totaling over 28,000 hours of EEG signals. Detailed information about pre-training datasets is in Appx. A.

Fine-tuning Datasets. We conduct experiments across **12 EEG datasets** spanning **10 task categories**, including: (1) Seizure Detection: Siena (Detti et al., 2020); (2) Emotion Recognition: SEED (Zheng & Lu, 2015), SEED-V (Liu et al., 2022a); (3) Motor Imagery: PhysioMI (Schalk et al., 2004), Mimul-11 (Jeong et al., 2020); (4) Mental Stress: Workload (Zyma et al., 2019); (5) Sleep Staging: HMC (Alvarez-Estevez & Rijsman, 2021); (6) Anomalous Event Detection: TUEV (Harati et al., 2015); (7) Abnormal Classification: TUAB (López et al., 2015); (8) Visual Target Detection: Things-EEG-2 (Gifford et al., 2022); (9) Alzheimer's Disease Identification: ADFTD (Miltiadous et al., 2023); and (10) Slowing Event Classification: TUSL (von Weltin et al., 2017) (details in Appx. A).

4.2 EXPERIMENTAL SETUP

Data Preprocessing. We utilize an MNE-based (Gramfort et al., 2013) pipeline to resample, filter, and normalize EEG signals, facilitating reproducible and scalable feature extraction.

Evaluation Metrics. To account for class imbalance in downstream tasks, we follow the protocol

Table 1: Performance comparison on 12 EEG datasets (Part 1): single-task baselines versus multi-task PRiSE-EEG. Best performance are highlighted in bolded. Statistically significant improvements over the second-best model are indicated by * (two-sample t-test: $p < 0.05$).

Dataset	Metrics	EEGNet	Conformer	BENDR	BIOT	LaBraM	EEGPT	CBraMod	CSBrain	REVE	PRiSE-EEG
TUSL	B-Acc	44.94±1.50	54.26±0.62	49.60±2.02	63.96±1.56	64.22±1.26	77.54±1.95	70.91±1.12	66.67±0.21	65.00±0.79	**79.82 ±1.95***
	W-F1	43.95±1.09	46.73±1.58	37.80±2.42	61.51±1.90	51.41±5.59	73.63±2.70	67.51±2.98	60.90±2.34	58.70±1.98	**77.34 ±2.30***
	Kappa	23.06±0.88	25.98±1.85	21.52±3.80	44.19±2.10	37.97±2.61	65.12±3.48	52.18±3.79	43.63±1.08	40.50±1.65	**65.45 ±2.31**
Siena Scalp	B-Acc	81.47±3.08	83.27±1.10	78.22±2.07	66.55±0.51	66.36±1.08	81.63±1.53	83.96±0.55	74.11±0.60	70.65±1.52	**86.89±2.31***
	AUROC	89.83±1.97	88.78±1.15	93.40±1.03	70.64±1.88	77.39±1.80	**93.71±1.54**	93.15±1.49	89.07±2.39	83.56±1.28	89.12±2.30
	AUC-PR	99.78±0.05	99.72±0.02	99.82±0.03	99.05±0.13	99.50±0.07	99.85±0.03	99.84±0.04	99.72±0.11	99.64±0.03	**99.85±0.03**
SEED	B-Acc	51.40±0.42	55.79±0.50	60.92±0.53	64.57±1.00	64.52±0.53	67.53±0.29	70.86±1.80	71.92±0.39	70.29±0.78	**73.92±0.31***
	W-F1	51.62±0.84	54.27±0.41	59.98±0.94	64.01±0.80	64.60±0.35	67.77±0.05	70.09±2.06	71.84±0.20	69.62±0.72	**73.29±0.22***
	Kappa	27.43±0.79	33.09±0.56	41.55±0.79	47.03±1.50	47.10±0.79	51.92±0.61	56.46±2.70	58.02±0.59	55.59±1.18	**61.02±0.45***
SEED-V	B-Acc	29.61±0.28	35.37±0.23	22.49±0.19	30.48±0.24	39.60±1.09	38.33±0.69	40.98±1.58	38.02±0.23	40.84±0.98	**47.40±0.25***
	W-F1	27.49±1.20	34.87±0.38	22.22±1.95	31.05±0.30	39.55±0.85	36.96±0.89	41.18±1.46	38.07±1.03	40.58±2.00	**48.39±0.42***
	Kappa	10.06±0.39	17.72±0.25	3.86±0.03	14.07±0.28	24.17±1.20	22.08±1.06	25.58±1.87	22.48±0.42	25.59±1.48	**34.84±0.45***
PhysioMI	B-Acc	57.25±0.73	55.23±0.19	48.01±0.29	61.53±1.54	61.73±0.89	62.92±0.32	61.74±0.91	63.04±0.74	64.80±0.53	**65.55±0.38***
	W-F1	56.68±0.97	40.58±0.50	46.57±3.06	61.58±1.97	61.77±1.16	62.72±0.34	61.79±1.69	63.08±0.81	64.84±0.94	**65.83±0.56***
	Kappa	42.91±0.95	46.50±0.38	30.66±0.39	48.75±2.72	52.22±1.19	53.12±1.63	52.22±1.00	53.80±0.76	53.04±0.85	**54.78±0.69***
Mimul-11	B-Acc	45.92±0.59	49.25±0.81	49.68±0.94	41.73±0.28	48.57±1.04	47.09±0.66	50.51±0.63	42.67±0.77	45.00±0.62	**51.33±0.24***
	W-F1	49.29±1.15	46.51±1.70	58.49±1.22	50.59±0.56	57.95±0.96	56.60±0.95	58.09±0.61	51.93±0.87	54.63±0.71	**58.82±0.23***
	Kappa	20.28±1.62	21.98±1.05	32.74±3.52	15.96±1.08	31.91±2.54	26.23±2.89	**33.63±1.80**	18.13±2.05	23.10±1.31	31.49±0.98

Table 2: Performance comparison on 12 EEG datasets (Part 2): single-task baselines versus multi-task PRiSE-EEG. Best performance are highlighted in bolded. Statistically significant improvements over the second-best model are indicated by * (two-sample t-test: $p < 0.05$).

Dataset	Metrics	EEGNet	Conformer	BENDR	BIOT	LaBraM	EEGPT	CBraMod	CSBrain	REVE	PRiSE-EEG
Workload	B-Acc	66.74±1.92	67.47±0.30	62.67±0.99	63.67±0.59	57.93±0.86	69.31±1.69	74.23±1.36	71.27±0.75	69.53±2.65	**82.47±0.30***
	AUROC	77.48±1.26	76.94±0.42	71.38±3.02	75.09±1.42	78.29±0.48	74.70±0.98	82.53±0.75	79.63±1.20	77.07±4.26	**89.21±0.19***
	AUC-PR	49.98±1.50	50.30±1.14	50.24±3.75	69.09±2.25	60.15±0.55	55.93±1.92	61.48±1.90	55.10±2.21	64.67±3.20	**70.75±0.47***
TUEV	B-Acc	58.15±2.31	58.00±0.63	67.81±0.61	59.43±0.70	65.85±0.77	61.31±1.36	67.64±0.47	71.20±1.34	63.24±0.73	**75.28±0.51***
	W-F1	74.77±1.39	73.02±0.25	81.19±1.54	77.42±2.03	78.22±2.36	80.00±0.91	81.31±0.54	85.07±0.85	84.97±0.33	**88.25±0.51***
	Kappa	59.55±2.91	56.69±0.21	66.04±2.78	63.12±3.05	64.49±3.54	66.87±1.35	67.22±1.02	74.67±1.44	75.12±0.76	**80.90±0.62***
TUAB	B-Acc	79.87±0.46	81.78±0.53	81.67±0.70	81.23±0.42	80.74±0.43	81.75±0.65	80.89±0.36	78.97±0.17	80.53±0.12	**82.22±0.12***
	AUROC	88.78±0.25	89.58±0.49	88.08±0.45	88.36±0.27	86.72±1.48	89.60±1.05	87.81±0.45	85.33±1.51	87.43±1.04	**89.93±0.31**
	AUC-PR	87.86±0.67	90.60±0.53	88.18±1.41	88.91±0.33	87.22±1.49	90.00±1.56	88.51±0.53	82.73±4.69	86.00±2.20	**90.11±0.15**
Things-EEG-2	B-Acc	51.27±0.29	60.30±0.50	63.76±1.57	51.87±0.09	51.48±0.21	60.83±0.82	60.25±1.19	57.20±0.29	61.03±0.90	**65.15±0.48***
	AUROC	62.10±0.10	68.42±1.39	77.54±0.70	66.17±0.73	52.78±0.23	73.53±1.56	75.11±1.92	62.67±0.53	66.77±0.52	**86.63±0.35***
	AUC-PR	16.82±0.22	22.31±1.37	40.89±1.20	18.47±0.53	12.39±0.05	31.91±3.21	34.19±1.92	21.53±0.42	22.87±1.72	**42.82±1.19***
HMC	B-Acc	58.31±0.95	66.68±0.94	70.25±1.00	71.33±0.50	70.68±0.89	71.93±0.41	71.48±0.40	72.18±0.24	71.13±0.26	**74.33±0.21***
	W-F1	58.37±1.87	64.73±1.35	72.30±1.35	72.62±0.33	71.73±1.50	72.79±0.64	72.76±0.55	74.16±0.27	73.10±0.22	**77.31±0.20***
	Kappa	48.29±1.94	57.15±1.05	64.51±1.50	64.94±0.36	64.04±1.30	64.97±0.53	64.86±0.41	66.43±0.16	65.20±0.29	**70.69±0.19***
ADFTD	B-Acc	37.07±1.24	36.68±0.45	41.83±0.37	42.31±1.25	37.86±2.07	50.81±2.28	52.20±0.93	48.96±2.65	44.38±0.75	**58.75±0.66***
	W-F1	30.65±6.15	33.09±1.46	43.10±0.51	44.01±2.37	35.31±2.91	50.10±1.52	55.29±1.50	50.17±3.76	50.55±0.43	**64.12±0.63***
	Kappa	17.14±15.40	5.43±0.83	16.35±0.68	16.78±2.11	6.08±2.92	25.68±0.99	31.38±1.82	24.96±4.13	23.17±0.83	**39.41±0.79***

in Jiang et al. (2025) and employ Balanced Accuracy, Cohen's Kappa, and Weighted F1 for multi-class classification, and Balanced Accuracy, AUC-PR, and AUROC for binary classification. Model optimization dynamics and consistency are assessed using cosine similarity and expert router statistics.
Baselines. We compare PRiSE-EEG against two supervised benchmarks (EEGNet (Lawhern et al., 2018), EEGConformer (Song et al., 2022)) and seven EEG foundation models (BENDR (Kostas et al., 2021), BIOT (Yang et al., 2023), LaBraM (Jiang et al., 2024), REVE (Ouahidi et al., 2025), CBraMod (Wang et al., 2025), EEGPT (Wang et al., 2024), and CSBrain (Zhou et al., 2025)). **Experimental details are in Appx. D**.

4.3 Main Results

Tab. 1 and 2 report the results of **PRiSE-EEG** (**basic version: 41.06M parameters with 18.05M activated**) across diverse benchmarks. Baselines are trained or fine-tuned separately per task, whereas PRiSE-EEG uses unified multi-task fine-tuning. PRiSE-EEG results are mean ± standard deviation over 10 runs. Best results are in **bold**, and statistically significant improvements over the second-best model are marked by *. **PRiSE-EEG** achieves strong performance on most tasks. It shows large gains on disease-related and cognitive tasks, including ADFTD, SEED-V, and Workload, improving primary metrics by **6-8%** over existing EEG-FMs. Consistent **2-5%** gains are observed on TUEV, TUSL, PhysioMI, SEED, and HMC. For tasks such as TUAB and Mimul-11, where the gain in B-Acc is approximately 1%, **PRiSE-EEG** displays lower variance with enhanced stability. On Things-EEG-2, it achieves a **>9%** AUROC gain, indicating stronger fine-grained visual target discrimination.

Compared with supervised baselines (*e.g.*, EEGNet and Conformer), **PRiSE-EEG** benefits from self-supervised pretraining on large EEG corpora, reducing overfitting to task-specific labeled data. It also outperforms existing EEG-FMs, which we attribute to prior-guided continuous tokenization and the CKA-calibrated depth-stratified expert transformer. **Under single-task fine-tuning, the pretrained model surpasses existing EEG-FMs on most tasks** (Tab. 9), showing that the gains do not rely solely on unified multi-dataset fine-tuning. Overall, these results support the effectiveness of combining weak sensor-space priors with CKA-calibrated depth-aware expert allocation for robust cross-paradigm EEG representation learning. **Additional results on fine-tuning strategies, scaling, computational costs and representation-analysis are in Appx. E and G, H**.

4.4 Ablation Study

Fig. 3 reports ablations on four representative datasets, covering prior-guided tokenization, CKA-calibrated depth-stratified experts, the classification head and unified fine-tuning.

Prior-guided continuous tokenization. Combining static and dynamic paths achieves higher Balanced Accuracy than either path alone on all four tasks, showing that both coarse sensor-space priors and sample-dependent short-window interactions are useful for EEG patch formation. The

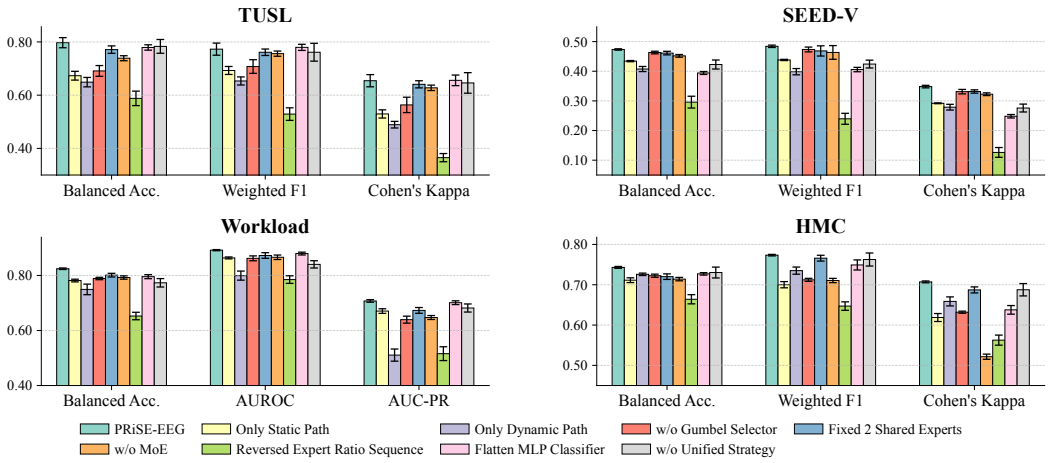


Figure 3: Ablations of tokenization, expert allocation, classifier, and unified fine-tuning. "Fixed 2 shared experts" selects two shared experts from the expert pool, giving three shared experts in total.

static path is more effective on TUSL, SEED-V, and Workload, whereas the dynamic path performs better on HMC. Removing Gumbel selection lowers Balanced Accuracy by 1.09-10.71%, with the largest drop on TUSL.

CKA-calibrated depth-stratified experts. Replacing MoE with the standard dense FFN lowers Balanced Accuracy by 2.21-5.94%. Depth-stratified allocation improves over the fixed-sharing baseline by 1.28-2.70%, while reversing the allocation order causes larger losses of 7.93-21.02%. These results support both the tested expert structure and its depth-dependent allocation rather than uniform sharing. Detailed results and computational costs are provided in Appx. E.3 and H.

Classifier and fine-tuning strategy. The dual-stream head achieves comparable or stronger performance than the flattened MLP while reducing head parameters by over 90%, suggesting that separate temporal and group pooling preserves useful structure without parameter explosion. The drop without unified fine-tuning further supports joint adaptation within this configuration, without establishing which component accounts for most of the overall gain.

4.5 Representation and Routing Analysis

We analyze PRiSE-EEG from three perspectives: prior-group attribution, CKA-calibrated expert allocation, and the relation between gradient geometry and expert routing.

Attribution of prior groups. We treat static and dynamic attention biases in the tokenizer as attribution targets and compute Integrated Gradients (Sundararajan et al., 2017) of the target logits with respect to each group bias using a zero baseline. Absolute attributions are accumulated across heads and temporal positions to identify influential prior groups. As shown in Tab. 3, PRiSE-

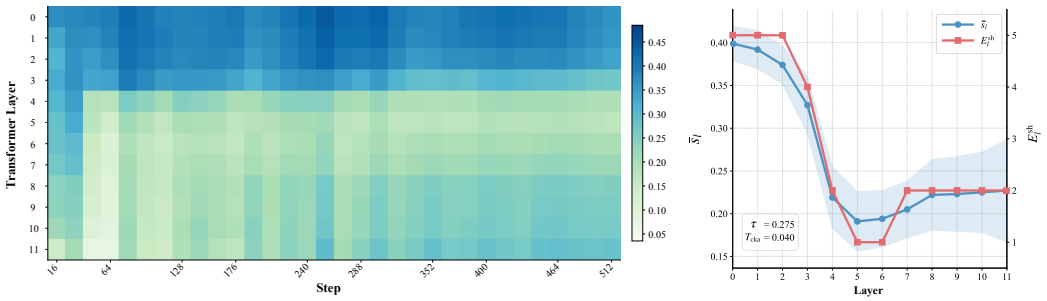


Figure 4: Layer-wise CKA sharedness over steps 16-512 (left) and total shared-expert counts (right), including the fixed expert; $\tau = 0.275$, $T_{\text{cka}} = 0.040$.

Table 3: The significant attribution groups for static and dynamic attention biases via Integrated Gradient.

Path	SEED		THINGS-EEG-2	
	Group Name	Focus Region	Group Name	Focus Region
Static	Frontal	-	Occipital	-
	Right Temporal	-	Visual Network	-
	Salience Network	-	Dorsal Attention	-
Dynamic	Lag-0	Frontal Temporal	Lag-0	Occipital
	Lag-1	Frontal Temporal	Lag-1	Occipital Parietal
	Lag-2	Frontal Temporal Parietal	Lag-2	Temporal Frontal

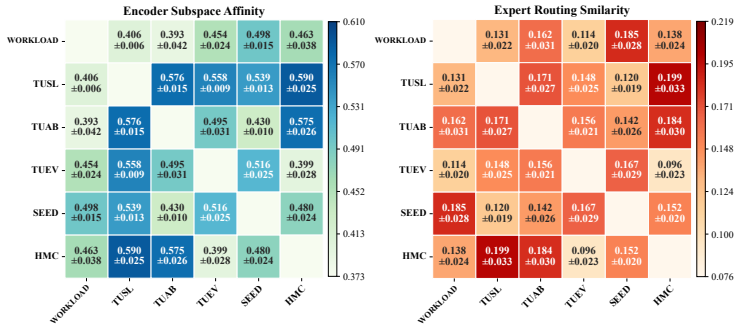


Figure 5: Gradient subspace affinity (rank=5) and similarity of routed expert distribution among different datasets.

EEG consistently assigns high attribution to frontal and temporal regions for emotion-related tasks, while occipital and parietal groups become more prominent in visual or decision-related tasks. Dynamic groups further reveal lag-dependent shifts of attended regions, suggesting that the tokenizer uses both stable spatial priors and sample-dependent short-window interactions. Additional Grad-CAM (Selvaraju et al., 2017) visualizations are provided in Appx. F.

CKA-calibrated shared expert allocation. Fig. 4 visualizes the layer-wise CKA sharedness used to configure the depth-stratified expert transformer. The CKA map shows a clear depth-wise transition: shallow layers maintain higher cross-paradigm alignment, whereas middle and deeper layers become more specialized. Accordingly, the sigmoid mapping assigns more shared experts to early layers and more specialized capacity to later layers. CKA guides layer-wise sharing; fixed-count and reversed-order ablations assess the value of depth-dependent allocation.

Gradient subspace and expert routing. Fig. 5 compares gradient subspace affinity with expert routing similarity across datasets. Encoder gradients are collected for 512 optimization steps under dataset-specific heads, projected to a lower-dimensional space, and summarized by rank-k PCA bases. For each dataset pair, subspace affinity measures the overlap between dominant update directions, while routing similarity is computed from cumulative expert usage. Datasets with more similar gradient subspaces tend to induce more similar routing patterns, such as SEED-TUEV and HMC-TUAB. This suggests that the depth-stratified experts do not route samples arbitrarily, but reflect dataset-level optimization geometry. Detailed computation method is described in Appx. C.

5 Conclusion

This work presents **PRiSE-EEG**, a prior-guided EEG foundation model for cross-paradigm representation learning. PRiSE-EEG forms continuous multi-channel patches with weak sensor-space priors and short-window interactions, and allocates shared and specialized expert capacity across layers using CKA-calibrated sharedness. Evaluated on 12 public benchmarks spanning 10 EEG task categories, PRiSE-EEG achieves strong performance in evaluations across multiple EEG paradigms. Further ablations and analyses show that prior-guided tokenization, depth-stratified expert allocation, and unified fine-tuning jointly improve cross-paradigm adaptation. Overall, PRiSE-EEG provides a compact framework for balancing shared EEG regularities with task-specific specialization.

AI USE STATEMENT

In this work, generative AI tools were used primarily to improve the clarity, grammar, and readability of the manuscript. They were not used for research ideation, methodology design, implementation, data processing, experimentation, or the analysis and interpretation of results. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the final manuscript.

ETHICS STATEMENT

This study uses existing EEG datasets made available for research, with no new data collection or participant recruitment. We follow the datasets' access conditions and ethical protocols, do not attempt participant re-identification, and report no participant-level information. EEG recordings contain sensitive physiological information, and differences in populations and acquisition settings may introduce dataset-specific biases. Our results do not establish clinical validity across populations. PRiSE-EEG is intended for research use; deployment in clinical or other high-stakes settings requires independent validation, appropriate data governance, and qualified professional oversight.

REPRODUCIBILITY STATEMENT

Code for reproducing the experiments is available at `https://anonymous.4open.science/r/PRiSE-EEG-A66F`. Section 3 describes the architecture, expert allocation, and training objectives. Appendices A, B, and D detail the datasets, implementation, preprocessing, data splits, training settings, and evaluation protocols. Appendix C describes the gradient and routing analyses, and Appendix E reports additional results and ablations.

REFERENCES

Mashail Alsolamy and Anas Fattouh. Emotion estimation from eeg signals during listening to quran using psd features. In *CSIT*, pp. 1–5, 2016.

Diego Alvarez-Estevez and Roselyne M. Rijsman. Inter-database validation of a deep learning approach for automatic sleep scoring. *PLOS ONE*, 16(8):1–27, 08 2021.

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*, pp. 6836–6846, 2021.

Christina Artemenko. Developmental fronto-parietal shift of brain activation during mental arithmetic across the lifespan: A registered report protocol. *PLOS ONE*, 16(8):1–13, 08 2021.

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*, pp. 813–824. PMLR, 2021.

Benjamin Blankertz, Guido Dornhege, Matthias Krauledat, Klaus-Robert Müller, and Gabriel Curio. The non-invasive berlin brain–computer interface: Fast acquisition of effective performance in untrained subjects. *NeuroImage*, 37(2):539–550, 2007.

O Boucher, CH Bastien, G Muckle, D Saint-Amour, SW Jacobson, and JL Jacobson. Behavioural correlates of the p3b event-related potential in school-age children. *International Journal of Psychophysiology*, 76(3):148–157, 2010.

Matthew J Brookes, Mark Woolrich, Henry Luckhoo, Darren Price, Joanne R Hale, Mary C Stephenson, Gareth R Barnes, Stephen M Smith, and Peter G Morris. Investigating the electrophysiological basis of resting state networks using magnetoencephalography. *Proceedings of the National Academy of Sciences*, 108(40):16783–16788, 2011.

Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. *Theoretical Computer Science*, 312(1):3–15, 2004. ISSN 0304-3975. doi: 10.1016/S0304-3975(03)00400-6. Automata, Languages and Programming.

Zitian Chen, Yikang Shen, Mingyu Ding, Zhenfang Chen, Hengshuang Zhao, Erik Learned-Miller, and Chuang Gan. Mod-Squad: Designing mixtures of experts as modular multi-task learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Andac Demir, Toshiaki Koike-Akino, Ye Wang, Masaki Haruna, and Deniz Erdogmus. Eeg-gnn: Graph neural networks for classification of electroencephalogram (eeg) signals. In *EMBC*, pp. 1061–1067, 2021.

Paolo Detti et al. Siena scalp eeg database. *physionet*, 10:493, 2020.

Harini Eavani, Meng Kang Hsieh, Yang An, Guray Erus, Lori Beason-Held, Susan Resnick, and Christos Davatzikos. Capturing heterogeneous group differences using mixture-of-experts: Application to a study of aging. *Neuroimage*, 125:498–514, 2016.

Lingzhong Fan, Hai Li, Junjie Zhuo, Yu Zhang, Jiaojian Wang, Liangfu Chen, Zhengyi Yang, Congying Chu, Sangma Xie, Angela R Laird, et al. The human brainnetome atlas: a new brain atlas based on connectional architecture. *Cerebral cortex*, 26(8):3508–3526, 2016.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Xuange Gao, Danli Wang, and Yanyan Zhao. Eegmoe: A domain-decoupled mixture-of-experts model for self-supervised eeg representation learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2026.

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.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti Hämäläinen. Meg and eeg data analysis with mne-python. *Frontiers in Neuroscience*, 7, 2013.

Tijl Grootswagers, Ivy Zhou, Amanda K Robinson, Martin N Hebart, and Thomas A Carlson. Human eeg recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Scientific Data*, 9(1):3, 2022.

A. Harati, M. Golmohammadi, S. Lopez, I. Obeid, and J. Picone. Improved eeg event classification using differential energy. In *SPMB*, pp. 1–4, 2015.

Joerg F Hipp, David J Hawellek, Maurizio Corbetta, Markus Siegel, and Andreas K Engel. Large-scale cortical correlation structure of spontaneous oscillatory activity. *Nature neuroscience*, 15(6): 884–890, 2012.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparametrization with gumble-softmax. In *International Conference on Learning Representations (ICLR 2017)*, 2017.

Ganesh Jawahar, Subhabrata Mukherjee, Xiaodong Liu, Young Jin Kim, Muhammad Abdul-Mageed, Laks V. S. Lakshmanan, Ahmed Hassan Awadallah, Sebastien Bubeck, and Jianfeng Gao. AutoMoE: Heterogeneous mixture-of-experts with adaptive computation for efficient neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 9116–9132, 2023.

Ji-Hoon Jeong, Jeong-Hyun Cho, Kyung-Hwan Shim, Byoung-Hee Kwon, Byeong-Hoo Lee, Do-Yeun Lee, Dae-Hyeok Lee, and Seong-Whan Lee. Multimodal signal dataset for 11 intuitive movement tasks from single upper extremity during multiple recording sessions. *GigaScience*, 9 (10):giaa098, 2020.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. In *ICLR*, 2024.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. Neurolm: A universal multi-task foundation model for bridging the gap between language and eeg signals. In *ICLR*, 2025.

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). Technical report, GIPSA-lab, July 2019.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *FHN*, 15: 653–659, 2021.

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), 2018.

Dongdong Li, Zhongliang Zeng, Zhe Wang, and Hai Yang. Estformer: Transformer utilising spatiotemporal dependencies for electroencephalogram super-resolution. *Knowledge-Based Systems*, 317:113345, 2025.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Bingxiu Liu, Jifeng Guo, CL Philip Chen, Xia Wu, and Tong Zhang. Fine-grained interpretability for eeg emotion recognition: Concat-aided grad-cam and systematic brain functional network. *IEEE Transactions on Affective Computing*, 15(2):671–684, 2023.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE TCDS*, 14(2):715–729, 2022a.

Wei Liu, Wei-Long Zheng, Ziyi Li, Si-Yuan Wu, Lu Gan, and Bao-Liang Lu. Identifying similarities and differences in emotion recognition with eeg and eye movements among chinese, german, and french people. *Journal of Neural Engineering*, 19(2):026012, 2022b.

Matthew D. Luciw, Ewa Jarocka, and Benoni B. Edin. Multi-channel eeg recordings during 3,936 grasp and lift trials with varying weight and friction. *Scientific Data*, 1(1):140047, Nov 2014.

S. López, G. Suarez, D. Jungreis, I. Obeid, and J. Picone. Automated identification of abnormal adult eegs. In *SPMB*, pp. 1–5, 2015.

Jingying Ma, Feng Wu, Qika Lin, Yucheng Xing, Chenyu Liu, Ziyu Jia, and Mengling Feng. Codebrain: Towards decoupled interpretability and multi-scale architecture for eeg foundation model. *arXiv preprint arXiv:2506.09110*, 2025.

Perrin Margaux, Maby Emmanuel, Daligault Sébastien, Bertrand Olivier, and Mattout Jérémie. Objective and subjective evaluation of online error correction during p300-based spelling. *Advances in Human-Computer Interaction*, 2012(1):578295, 2012.

Chirag Mehra, Ahmad Beyh, Petroula Laiou, Pilar Garces, Emily JH Jones, Luke Mason, Jan Buitelaar, Mark H Johnson, Declan Murphy, Eva Loth, et al. Zero-phase-delay synchrony between interacting neural populations: implications for functional connectivity-derived biomarkers. *Imaging Neuroscience*, 3:IMAG–a, 2025.

Christoph M Michel and Thomas Koenig. Eeg microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: a review. *Neuroimage*, 180:577–593, 2018.

Andreas Miltiadous, Katerina D Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Nikolaos Grigoriadis, Dimitrios G Tsalikakis, Pantelis Angelidis, Markos G Tsipouras, Euripidis Glavas, Nikolaos Giannakeas, et al. A dataset of scalp eeg recordings of alzheimer's disease, frontotemporal dementia and healthy subjects from routine eeg. *Data*, 8(6):95, 2023.

Gavin Mischler, Yinghao Aaron Li, Stephan Bickel, Ashesh D Mehta, and Nima Mesgarani. Contextual feature extraction hierarchies converge in large language models and the brain. *Nature Machine Intelligence*, 6(12):1467–1477, 2024.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in neural information processing systems*, 31, 2018.

Raffaele Nardone, Luca Sebastianelli, Viviana Versace, Leopold Saltuari, Piergiorgio Lochner, Vanessa Frey, Stefan Golaszewski, Francesco Brigo, Eugen Trinka, and Yvonne Höller. Usefulness of eeg techniques in distinguishing frontotemporal dementia from alzheimer's disease and other dementias. *Disease markers*, 2018(1):6581490, 2018.

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

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*, abs/2510.21585, 2025.

John Polich. Updating p300: an integrative theory of p3a and p3b. *Clinical neurophysiology*, 118 (10):2128–2148, 2007.

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

Pavel Prado, Vicente Medel, Raul Gonzalez-Gomez, Agustín Sainz-Ballesteros, Victor Vidal, Hernando Santamaría-García, Sebastian Moguilner, Jhony Mejia, Andrea Slachevsky, Maria Isabel Behrens, et al. The brainlat project, a multimodal neuroimaging dataset of neurodegeneration from underrepresented backgrounds. *Scientific Data*, 10(1):889, 2023.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Christopher Rorden. Mricrogl: voxel-based visualization for neuroimaging. *Nature methods*, 22(8): 1613–1614, 2025.

Subhrajit Roy, Isabell Kiral-Kornek, and Stefan Harrer. Chrononet: A deep recurrent neural network for abnormal eeg identification. In David Riaño, Szymon Wilk, and Annette ten Teije (eds.), *Artificial Intelligence in Medicine*, pp. 47–56, 2019.

Arman Savran, Koray Çiftçi, Guillaume Chanel, Javier Mota, Luong Viet, Bulent Sankur, Lale Akarun, Alice Caplier, and Michèle Rombaut. Emotion detection in the loop from brain signals and facial images. In *Summer Workshop on Multimodal Interfaces (eINTERFACE 2006)*, pp. 69–80, 2006.

G. Schalk, D.J. McFarland, T. Hinterberger, N. Birbaumer, and J.R. Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE TBE*, 51(6):1034–1043, 2004.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Seyed Yahya Shirazi, Alexandre Franco, Maurício Scopel Hoffmann, Nathalia B Esper, Dung Truong, Arnaud Delorme, Michael P Milham, and Scott Makeig. Hbn-eeg: The fair implementation of the healthy brain network (hbn) electroencephalography dataset. *bioRxiv*, pp. 2024–10, 2024.

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

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *TNSRE*, 31:710–719, 2022.

Sebastian Stober, Avital Sternin, Adrian M Owen, and Jessica A Grahn. Towards music imagery information retrieval: Introducing the openmiir dataset of eeg recordings from music perception and imagination. In *ISMIR*, pp. 763–769, 2015.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4593–4601, 2019.

Mastaneh Torkamani-Azar, Sumeyra Demir Kanik, Serap Aydin, and Mujdat Cetin. Prediction of reaction time and vigilance variability from spatio-spectral features of resting-state eeg in a long sustained attention task. *IEEE JBHI*, 24(9):2550–2558, 2020.

Logan T. Trujillo. Mental effort and information-processing costs are inversely related to global brain free energy during visual categorization. *Frontiers in Neuroscience*, 13, 2019.

Logan T. Trujillo, Candice T. Stanfield, and Ruben D. Vela. The effect of electroencephalogram (eeg) reference choice on information-theoretic measures of the complexity and integration of eeg signals. *Frontiers in Neuroscience*, 11, 2017.

Dimitri Van de Ville, Juliane Britz, and Christoph M Michel. Eeg microstate sequences in healthy humans at rest reveal scale-free dynamics. *Proceedings of the National Academy of Sciences*, 107 (42):18179–18184, 2010.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Simon Vandenhende, Stamatios Georgoulis, Bert De Brabandere, and Luc Van Gool. Branched multi-task networks: Deciding what layers to share. In *Proceedings of the British Machine Vision Conference*, 2020.

Mariska J Vansteensel, Martin G Bleichner, Zac V Freudenburg, Dora Hermes, Erik J Aarnoutse, Frans SS Leijten, Cyrille H Ferrier, Johan Martijn Jansma, and Nick F Ramsey. Spatiotemporal characteristics of electrocortical brain activity during mental calculation. *Human Brain Mapping*, 35(12), 2014.

Raul Vicente, Leonardo L Gollo, Claudio R Mirasso, Ingo Fischer, and Gordon Pipa. Dynamical relaying can yield zero time lag neuronal synchrony despite long conduction delays. *Proceedings of the National Academy of Sciences*, 105(44):17157–17162, 2008.

Eva von Weltin, Tameem Ahsan, Vinit Shah, Dawer Jamshed, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. Electroencephalographic slowing: A primary source of error in automatic seizure detection. In *2017 IEEE signal processing in medicine and biology symposium (SPMB)*, pp. 1–5. IEEE, 2017.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. In *NeurIPS*, pp. 39249–39280, 2024.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. In *ICLR*, 2025.

Yuxiang Wei, Yanteng Zhang, Xi Xiao, Tianyang Wang, Xiao Wang, and Vince D Calhoun. Morebrain: Routed mixture of experts for interpretable and generalizable cross-subject fmri visual decoding. *arXiv preprint arXiv:2505.15946*, 2025.

Marta Xavier, Inês Esteves, João Jorge, Rodolfo Abreu, Anne-Lise Giraud, Sepideh Sadaghiani, Jonathan Wirsich, and Patrícia Figueiredo. Consistency of resting-state correlations between fmri networks and eeg band power. *Imaging Neuroscience*, 3:IMAG–a, 2025.

Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. In *NeurIPS*, pp. 78240–78260, 2023.

Ke Yi, Yansen Wang, Kan Ren, and Dongsheng Li. Learning topology-agnostic EEG representations with geometry-aware modeling. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19781–19791, 2025.

Zihan Zhang, Xiao Ding, Yu Bao, Yi Zhao, Xia Liang, Bing Qin, and Ting Liu. Chisco: An eeg-based bci dataset for decoding of imagined speech. *Scientific Data*, 11(1):1265, 2024.

Ziyu Zhao, Tong Zhu, Zhi Zhang, Tiantian Fan, Jinluan Yang, Kun Kuang, Zhongyu Wei, Fei Wu, and Yu Cheng. ExpertWeaver: Unlocking the inherent MoE in dense LLMs with GLU activation patterns. *arXiv preprint arXiv:2602.15521*, 2026.

W. Zheng, W. Liu, Y. Lu, B. Lu, and A. Cichocki. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE TC*, pp. 1–13, 2018.

Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE TAMD*, 7(3):162–175, 2015.

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

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

This supplementary material presents additional details, analyses, and experiments to support our primary findings. The document is structured as follows:

- Appendix A describes the datasets utilized for pre-training and fine-tuning.
- Appendix B elaborates on the specific implementation details of the model.
- Appendix C details the method to analyze the gradient and routing dynamics.
- Appendix D details the complete experimental configurations.
- Appendix E presents supplementary experimental results and ablation studies evaluating the impact of different configurations.
- Appendix F contains additional visualizations of classification results and saliency topomaps to further illustrate the model's interpretability.
- Appendix G provides an empirical investigation into the scaling behavior governing EEG representations.
- Appendix I concludes with a discussion of current limitations and outlines potential avenues for future research.

A Datasets Description

A.1 Pre-training Datasets

The pre-training corpus for **PRiSE-EEG** comprises **17** diverse, publicly accessible EEG datasets. This multi-site corpus spans various experimental protocols, including clinical monitoring (*e.g.*, HBN, TUEG, Siena), cognitive assessments (*e.g.*, SEED series, Motor Imagery), and brain-computer interface (BCI) tasks (*e.g.*, BCIC, P300). Comprehensive specifications for each dataset, including specific citations, recording durations, and functional descriptions, are summarized in Tab. 4. To prevent pretraining data leakage, all TUEG recordings whose subject identifiers overlap with TUAB, TUEV, or TUSL evaluation subjects are removed from the pre-training corpus.

SEED-family datasets. The pretraining corpus includes SEED-IV, SEED-GER, and SEED-FRA, while SEED and SEED-V are used as downstream emotion-recognition datasets (Liu et al., 2022a). SEED and SEED-IV were collected from different participants (Yi et al., 2023). SEED-GER and SEED-FRA extend the SEED family to German and French populations, with eight participants in each dataset (Liu et al., 2022b). The downstream SEED and SEED-V results follow the subject-dependent trial-splitting protocols described in Appx. D.

A.2 Fine-tuning Datasets

The detailed information on the evaluation datasets are listed in Tab. 5: (1) Seizure Detection: **Siena** (Detti et al., 2020) (binary classification with seizure or healthy); (2) Emotion Recognition: **SEED** (Zheng & Lu, 2015) (3-class classification with sad, neutral or happy), **SEED-V** (Liu et al., 2022a) (5-class classification with disgust, fear, sad, neutral or happy); (3) Motor Imagery: **PhysioMI** (Schalk et al., 2004) (4-class classification with left fist, right fist, both fists or feet), **Mimul-11** (Jeong et al., 2020) (3-class classification with reaching, grasping or twisting); (4) Mental Stress: **Workload** (Zyma et al., 2019) (binary-class classification with arithmetic calculation or resting); (5) Sleep Staging: **HMC** (Alvarez-Estevez & Rijsman, 2021) (5-class classification with wake, REM, N1, N2 or N3); (6) Anomalous Event Detection: **TUEV** (Harati et al., 2015) (6-class classification with spike and slow wave, generalized periodic epileptiform discharge, periodic lateralized epileptiform dischage, eye movement artifact or background); (7) Abnormal Classification: **TUAB** (López et al., 2015) (binary-class classification with abnormal or normal); (8) Visual Target Detection: **Things-EEG-2** (Gifford et al., 2022) (binary-class classification with target or non-target); (9) Alzheimer 's Disease Identification: **ADFTD** (Miltiadous et al., 2023) (3-class classification with Alzheimer's Disease, Frontotemporal Dementia or healthy); (10) Slowing Event Classification: **TUSL** (von Weltin et al., 2017) (3-class classification with seizure, slow wave or background).

Table 4: Detailed information about pre-training datasets. Window durations are in seconds.

Dataset	Category	#Channel	Duration	#Train	#Valid	Description
TUEG (Obeid & Picone, 2016)	Clinical Recordings	19-22	60	1515391	75436	A rich archive of clinical EEG recordings collected at Temple University Hospital.
HBN (Shirazi et al., 2024)	Multiple Categories	70	15	395605	42376	EEG data and behavioral responses collected during EEG experiments from more than 3000 participants (5-21) performing six distinct tasks in the HBN project.
SEED-IV (Zheng et al., 2018)	Emotion Recognition	60	10	13678	977	A emotion EEG dataset that 15 subjects watch 72 film clips which will induce happiness, sadness, fear or neutral emotions.
SEED-GER (Liu et al., 2022b)	Emotion Recognition	60	10	8440	919	A emotion dataset eight German subjects watch 20 film clips (positive, neutral, negative) as stimuli.
SEED-FRA (Liu et al., 2022b)	Emotion Recognition	60	10	6846	978	Eight French subjects watch 21 film clips in French (positive, neutral, negative) as stimuli.
BCIC-1A (Blankertz et al., 2007)	Motor Imagery	43	8	3155	535	EEG recordings for motor imagery tasks, where subjects imagined moving either their left hand, right hand, or foot.
Emobrain (Savran et al., 2006)	Emotion Recognition	54	10	1370	405	Multimodal emotion detection dataset using brain signals (EEG, fNIRS) from 5 male subjects.
Grasp and Lift (Luciw et al., 2014)	Motor Execution	32	5	7003	1390	Grasp and lift action from 12 subjects in total, 10 series of trials for each subject.
Inria BCI P300 (Margaux et al., 2012)	ERP	56	5	13647	7901	A P300-based spelling dataset including 26 subjects.
Resting State (Trujillo et al., 2017)	Resting	64	10	981	100	A dataset comprising 22 subjects for a resting eyes closed and eyes open.
Raw EEG Data (Trujillo, 2019)	Visual Stimulus	64	30	4806	319	A dataset recorded during reported Information-Integration categorization and reported multidimensional Rule-Based categorization tasks.
SPIS Resting State (Torkamani-Azar et al., 2020)	Resting	64	10	270	30	A resting-state EEG from 10 subjects contains 2.5 minutes of eyes-open and 2.5 minutes of eyes-closed.
Target Versus Non-Target (Korczowski et al., 2019)	ERP	32	15	3403	333	Dataset contains EEG recordings of 50 subjects playing to a visual P300 Brain-Computer Interface (BCI) video game.
BrainLat (Prado et al., 2023)	Alzheimer's Disease	48	10	4164	912	The Latin American Brain Health Institute dataset comprises multimodal data of 780 participants from Latin American with neurodegenerative diseases.
Chisco (Zhang et al., 2024)	Imagined Speech	62	30	9941	2061	Chinese Imagined Speech Corpus (Chisco), including over 20,000 sentences of high-density EEG recordings of imagined speech from healthy adults.
Things EEG 1 (Grootswagers et al., 2022)	Visual Stimulus	62	20	6568	1749	The THINGS-EEG dataset provides neuroimaging recordings to a systematic collection of objects and concepts.
OpenMIIR (Stober et al., 2015)	Acoustic Stimulus	64	10	3399	992	A public domain dataset of EEG recordings taken during music perception and imagination comprised 10 subjects listening to and imagining 12 short music fragments.

B Implementation Details

B.1 Overview

Let the raw EEG signal be denoted as $\mathbf{x}_0 \in \mathbb{R}^{C \times P_0}$, where C is the number of channels and P_0 is the total number of time points. This signal is segmented into non-overlapping patches of length $P = f_s$. After batching, the batched input tensor is defined as $\mathbf{x}_p \in \mathbb{R}^{B \times T \times C \times P}$, where B is the batch size and T is the number of temporal patches. Throughout this paper, D_m denotes the model embedding dimension, $\bigoplus$ represents the feature concatenation operator, and RMS and ZN refer to Root Mean Square and Z-score normalization, respectively.

Our framework consists of a shared encoder and two distinct decoder configurations. The encoder includes: (1) a **Prior-Guided Continuous Tokenizer** and (2) a **Depth-Stratified Expert Trans-**

Table 5: Detailed information about evaluation datasets.

Dataset	Category	#Channel	Duration	#Train	#Valid	#Test	Task
TUAB	Abnormal Classification	23	30	247728	12315	12277	Binary Classification
TUEV	Anomalous Event Detection	21	5	87834	12473	13046	6-class Classification
TUSL	Slowing Event Classification	21,22	10	210	43	37	3-class Classification
SEED	Emotion Recognition	60	10	22455	7875	7560	3-class Classification
SEED-V	Emotion Recognition	60	10	3552	4638	4128	5-class Classification
HMC	Sleep Staging	4	30	91681	22804	22440	5-class Classification
Workload	Mental Stress	19	10	1537	300	297	Binary Classification
Siena	Seizure Detection	29	10	41631	5592	3607	Binary Classification
Mimul-11	Motor Imagery	60	5	31398	5000	4949	3-class Classification
PhysioMI	Motor Imagery	64	4	6210	1734	1803	4-class Classification
Things-EEG-2	Visual Target Detection	63	5	24915	8324	8331	Binary Classification
ADFTD	Alzheimer's Disease Identification	19	10	4743	1115	1155	3-class Classification

former. The stage-specific components include (3) an **EEG Reconstructor** for pre-training and (4) **Dual-Stream Spatial-Temporal Classifiers** for fine-tuning.

B.2 Prior-Guided Continuous Tokenizer

B.2.1 Feature Extraction and Positional Encoding

To capture temporal dynamics and global spectral characteristics, we extract features from both time and frequency domains. In the temporal domain, we apply stacked 1D convolutional blocks with multiple receptive fields to each patch. The normalized patch tensor $\mathbf{x}_p \in \mathbb{R}^{B\times T\times C\times P}$ is reshaped into $\mathbf{x}'_p \in \mathbb{R}^{(BTC)\times 1\times P}$ and processed via S convolutional branches. The aggregated temporal feature is defined as:

$$\mathbf{x}_t = \mathrm{RMS}\left(\mathrm{GELU}\left(\mathbf{W}_{\mathrm{emb}}^{\top}\bigoplus_{i=1}^{S}\mathbf{Y}_i\right)\right), \tag{11}$$

where $\mathbf{Y}_i$ denotes the output of the i-th branch. In the frequency domain, we extract spectral features using the Power Spectral Density (PSD) (Alsolamy & Fattouh, 2016). Specifically, we apply Fast Fourier Transform (FFT) to $\mathbf{x}_p$ to obtain its frequency representation. The one-sided PSD is computed using a Hanning window and converted to the decibel scale, yielding the spectral feature $\mathbf{x}_s$. The composite feature $\mathbf{x} \in \mathbb{R}^{B\times T\times C\times D_m}$ is formed by concatenating the normalized temporal and spectral features:

$$\mathbf{x} = \bigoplus(\mathbf{ZN}(\mathbf{x}_t), \mathbf{ZN}(\mathbf{x}_s)). \tag{12}$$

For the pre-training reconstruction task, the target signal $\mathbf{x}_{\mathrm{ref}}$ consists of the concatenated raw patch $\mathbf{x}_p$ and spectral features $\mathbf{x}_s$.

EEG channel configurations vary significantly across datasets. To address this, we define a standardized electrode superset $\mathbb{C}$ (based on the 10-10 system with auxiliary electrodes such as T1 and T2) and introduce a learnable channel positional embedding $\mathbf{E}_c \in \mathbb{R}^{|\mathbb{C}|\times D_m}$, which is added to the input features $\mathbf{x}$ and serves as a query in the EEG reconstructor. For the temporal dimension, we apply Rotary Positional Encoding (RoPE) within the attention to encode relative temporal positions.

B.2.2 Static Group Queries with Weak Sensor-Space Priors

A core strategy of our design involves representing EEG signals not only as channel patches but also as a set of *group patches* corresponding to neuro-functional organization. We define a fixed grouping scheme $\mathcal{P}$ (details in Appx. D.7) that organizes channels according to anatomical regions and coarse functional networks. To incorporate prior knowledge, we employ a prior matrix $\mathbf{M}_{\mathrm{s}} \in \mathbb{R}^{|\mathcal{P}|\times C}$ that defines channel-group associations, wherein in-group entries are set to 0 and out-of-group entries are assigned a negative bias (ablations in Appx. E.3). To prevent excessive rigidity, we introduce a learnable relaxation term $\mathbf{\Delta}_{\mathrm{pri}} \in \mathbb{R}^{|\mathcal{P}|\times C}$, which adjusts this bias while retaining the underlying structural guidance. Subsequently, we instantiate learnable parameters denoted as *static queries* $Q_s \in \mathbb{R}^{|\mathcal{P}|\times D_m}$, which serve as inputs to the group cross-attention modules. The attention mechanism

incorporating these priors is formulated as:

$$\mathrm{Attn}(Q_s, K, V) = \mathrm{Softmax}\left(\frac{Q_s K^\top}{\sqrt{d_k}} + \mathbf{M}_\mathrm{s} + \mathbf{\Delta}_\mathrm{pri}\right) V \tag{13}$$

where K and V are from the input $\mathbf{x}$, reshaped to $\mathbb{R}^{(B\times T)\times C\times D_m}$.

B.2.3 Dynamic Sub-second Queries through Short-Time Correlation

Static groups provide coarse region-network anchors. Dynamic groups complement them with learnable short-time channel affinities that vary with the sample and temporal position within each one-second patch. We reshape and embed each patch into **micro-frames**: $\mathbf{x} \to \tilde{\mathbf{x}} \in \mathbb{R}^{B\times T\times F\times C\times S}$, $P = F \cdot S$,

$$E = \phi(\tilde{\mathbf{x}}) \in \mathbb{R}^{B\times T\times F\times C\times H}, \tag{14}$$

$$Q' = EW_Q, K' = EW_K \in \mathbb{R}^{B\times T\times F\times C\times d}, \tag{15}$$

where F, S, and H are the number of frames, samples per frame, and frame embedding dimension, respectively; $\phi(\cdot)$ is a compact convolution projector, $W_Q, W_K \in \mathbb{R}^{H\times d}$ are projection matrices, and d is the correlation feature dimension. For a given patch, the **channel affinity** between micro-frames i and j is defined as:

$$A_{i,j} \;=\; \frac{1}{\sqrt{d}}\, Q'_i K'^\top_j \quad \in \mathbb{R}^{C\times C}, \tag{16}$$

where $Q'_i, K'_j \in \mathbb{R}^{C\times d}$ denote the slices of Q' and K' at frames i and j, respectively; $A_{i,j}(a, b)$ quantifies the affinity between channel a (at frame i) and channel b (at frame j). We define a lag-limited set of frame pairs, $\mathcal{G} = \{(i, j) \mid 0 \le i-j \le \mathbf{\Delta}_f\}$, where $\mathbf{\Delta}_f$ denotes the maximum frame lag, and each pair $(i, j) \in \mathcal{G}$ constitutes a single dynamic group g. The group biases are derived from each affinity matrix via a node encoder applied row-wise, $b_g \;=\; enc(A_g) \;\in \mathbb{R}^C$, where $A_g \in \mathbb{R}^{C\times C}$ denotes the affinity matrix for group g, and $enc(\cdot)$ is an MLP-based node encoder utilizing row statistics and top-k neighbor values. Aggregating these groups yields $\tilde{M}_d \in \mathbb{R}^{B\times T\times|\mathcal{G}|\times C}$, which serves as an additive dynamic attention bias analogous to Eq. 13, but utilizes a distinct set of parameters denoted as *dynamic queries* $\tilde{Q}_d \in \mathbb{R}^{|\mathcal{G}|\times D_m}$. To suppress uninformative lag pairs and reduce computation, we apply a **Gumbel Selector** to pick the most important frame-pairs in $\mathcal{G}$ for each sample-patch. We obtain differentiable importance scores with Gumbel-Softmax and select top-k groups:

$$\mathbf{\Lambda} = \mathrm{Mean_C}(\tilde{M}_d) + \mathrm{MLP}(\tilde{Q}_d) \in \mathbb{R}^{(B\cdot T)\times|\mathcal{G}|} \tag{17}$$

$$\Pi = \mathrm{TopK}(\mathrm{GumbelSoftmax}(\mathbf{\Lambda}; \tau), k) \in \mathbb{R}^{(B\cdot T)\times k} \tag{18}$$

where τ is temperature. The Π score is used to gather corresponding dynamic biases $\mathbf{M}_d \in \mathbb{R}^{B\times T\times k\times C}$ and queries $Q_d \in \mathbb{R}^{k\times D_m}$.

B.3 CKA-Calibrated Depth-Stratified Expert Transformer

During pretraining, PRiSE-EEG contains an ordered MoE Transformer stack consisting of tokenizer cross-attention blocks, temporal-wise blocks, group-wise blocks, and decoder blocks. For PRiSE-EEG-B, this stack has $L = 12$ blocks: 2 tokenizer cross-attention blocks, 4 group MoE blocks, 4 temporal MoE blocks, and 2 decoder MoE blocks. All blocks use the same MoE-FFN structure, so CKA-calibrated shared expert allocation can be applied uniformly across the stack.

Interleaved temporal and group layers. Given the tokenizer output $\mathbf{h}_0 \in \mathbb{R}^{B\times T\times G\times D_m}$, the encoder contains L_E Transformer layers. Temporal-wise and group-wise Transformer layers appear alternately, so each type has $L_E/2$ layers (Bertasius et al., 2021; Arnab et al., 2021; Li et al., 2025). For layer l, let $a_l \in \{\mathrm{temp}, \mathrm{group}\}$ denote its attention type. If $a_l = \mathrm{temp}$, self-attention is applied along the temporal dimension T for each group. If $a_l = \mathrm{group}$, self-attention is applied along the group dimension G for each time patch. Each temporal-wise or group-wise layer has its own depth-stratified MoE feed-forward block:

$$\tilde{\mathbf{h}}_l = \mathrm{Attn}_l^{a_l}(\mathbf{h}_l), \qquad \mathbf{h}_{l+1} = \tilde{\mathbf{h}}_l + \mathrm{DSE}_l(\mathrm{LN}(\tilde{\mathbf{h}}_l)). \tag{19}$$

Thus, depth stratification is applied to the FFN of every Transformer layer, regardless of whether the layer performs temporal-wise or group-wise attention.

Dense pilot checkpoints. Before training PRiSE-EEG, we train a dense pilot version with the same tokenizer and interleaved temporal/group attention layout, but with standard dense FFNs. We save pilot checkpoints at selected pretraining steps $\mathcal{T} = \{t_1, \ldots, t_m\}$. At each step $t \in \mathcal{T}$, we feed a mixed-paradigm batch $\mathcal{B}_t = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_b}$, where y_i denotes the dataset or paradigm identity of sample i.

Sample-wise CKA. For block l, let N_l denote the number of tokens in that block representation. For sample i at pretraining step t, we reshape the block output as

$$\mathbf{R}_{l,t}^{(i)} = \mathrm{reshape}(\mathbf{h}_{l,t}^{(i)}) \in \mathbb{R}^{N_l \times D_m} \tag{20}$$

be its flattened token representation. We use linear CKA between two samples i and j:

$$\mathrm{CKA}_{l,t}(i,j) = \frac{\left\| (\bar{\mathbf{R}}_{l,t}^{(i)})^\top \bar{\mathbf{R}}_{l,t}^{(j)} \right\|_F^2}{\left\| (\bar{\mathbf{R}}_{l,t}^{(i)})^\top \bar{\mathbf{R}}_{l,t}^{(i)} \right\|_F \left\| (\bar{\mathbf{R}}_{l,t}^{(j)})^\top \bar{\mathbf{R}}_{l,t}^{(j)} \right\|_F + \epsilon}, \tag{21}$$

where $\bar{\mathbf{R}} = \mathbf{J}\mathbf{R}$, $\mathbf{J}_l = \mathbf{I}_{N_l} - \frac{1}{N_l}\mathbf{1}\mathbf{1}^\top$ centers the token dimension, $\| \cdot \|_F$ is the Frobenius norm, and ϵ is a small constant.

Within- and between-paradigm CKA. Let $\mathcal{Y}$ be the set of paradigms in the mixed batch. For paradigm a, define within-paradigm pairs

$$\mathcal{I}_{a,a} = \{(i,j) \mid i < j,\ y_i = a,\ y_j = a\}. \tag{22}$$

For two different paradigms a and b, define between-paradigm pairs

$$\mathcal{I}_{a,b} = \{(i,j) \mid y_i = a,\ y_j = b\}. \tag{23}$$

We first average over pairs inside each paradigm or paradigm pair, and then average over paradigms:

$$c_{l,t}^{\mathrm{intra}} = \frac{1}{|\mathcal{Y}|} \sum_{a \in \mathcal{Y}} \frac{1}{|\mathcal{I}_{a,a}|} \sum_{(i,j) \in \mathcal{I}_{a,a}} \mathrm{CKA}_{l,t}(i,j), \tag{24}$$

$$c_{l,t}^{\mathrm{inter}} = \frac{2}{|\mathcal{Y}|(|\mathcal{Y}|-1)} \sum_{\substack{a,b \in \mathcal{Y} \\ a<b}} \frac{1}{|\mathcal{I}_{a,b}|} \sum_{(i,j) \in \mathcal{I}_{a,b}} \mathrm{CKA}_{l,t}(i,j). \tag{25}$$

Here $c_{l,t}^{\mathrm{intra}}$ measures within-paradigm representation consistency at layer l and pretraining step t, while $c_{l,t}^{\mathrm{inter}}$ measures cross-paradigm alignment.

CKA-derived shared expert ratio. We define normalized sharedness as

$$s_{l,t} = \mathrm{clip}\left(\frac{c_{l,t}^{\mathrm{inter}}}{c_{l,t}^{\mathrm{intra}} + \epsilon}, 0, 1 \right), \qquad \bar{s}_l = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} s_{l,t}. \tag{26}$$

The shared expert proportion of layer l is

$$\rho_l^{\mathrm{sh}} = \rho_{\min} + (\rho_{\max} - \rho_{\min}) \sigma\left(\frac{\bar{s}_l - \tau}{T_{\mathrm{cka}}} \right), \tag{27}$$

where $\rho_{\min}$ and $\rho_{\max}$ bound the shared-expert proportion, τ is the sharedness threshold, and T_{cka} is the sigmoid temperature. Here E_l is the total number of experts, including one fixed shared expert, and ρ_l^{sh} is the total shared proportion. We set

$$E_l^{\mathrm{sh}} = \mathrm{round}(E_l \rho_l^{\mathrm{sh}}), \qquad E_l^{\mathrm{sp}} = E_l - E_l^{\mathrm{sh}}. \tag{28}$$

The allocation is computed once from the dense pilot pretraining steps and then fixed during PRiSE-EEG training.

Expert configuration and routing. Each layer of PRiSE-EEG-B contains one default shared expert and an eight-expert pool partitioned by CKA. The total shared count E_l^{sh} includes the default expert, leaving $E_l^{\mathrm{sh}} - 1$ shared and $9 - E_l^{\mathrm{sh}}$ specialized experts in the pool. With $\rho_{\min} = 1/9$, $\rho_{\max} = 5/9$, $\tau = 0.275$, and $T_{\mathrm{cka}} = 0.040$, the 12-layer allocation is:

$$[E_l^{\mathrm{sh}}]_{l=0}^{11} = [5, 5, 5, 4, 2, 1, 1, 2, 2, 2, 2, 2]. \tag{29}$$

The allocation is determined before PRiSE-EEG training and remains fixed thereafter. For token $\mathbf{h}_{l,n}$, the router assigns logits and probabilities only to the specialized expert pool:

$$a_{l,e}(\mathbf{h}_{l,n}) = \mathbf{w}_{l,e}^{\top} \mathbf{h}_{l,n}, \tag{30}$$

$$p_{l,e}(\mathbf{h}_{l,n}) = \frac{\exp(a_{l,e}(\mathbf{h}_{l,n}))}{\sum_{j \in \mathcal{E}_l^{\mathrm{sp}}} \exp(a_{l,j}(\mathbf{h}_{l,n}))}. \tag{31}$$

The top-K specialized experts, with $K = 2$, are selected according to

$$\mathcal{K}_l(\mathbf{h}_{l,n}) = \mathrm{TopK}_{e \in \mathcal{E}_l^{\mathrm{sp}}} \left(a_{l,e}(\mathbf{h}_{l,n}) + b_{l,e}, K \right). \tag{32}$$

The output follows Eq. 8: all shared experts process every token, whereas only the selected specialized experts are activated. Specialized-expert outputs are weighted by $p_{l,e}$ without renormalization over $\mathcal{K}_l$, and no token is dropped.

Load balancing. Following the DeepSeek-style bias-based balancing strategy (Liu et al., 2024), $b_{l,e}$ affects expert selection but not the routing probability. Let N_l denote the number of tokens routed by layer l in an update. The normalized load of specialized expert e is

$$u_{l,e} = \frac{1}{K N_l} \sum_{n=1}^{N_l} \mathbb{I}[e \in \mathcal{K}_l(\mathbf{h}_{l,n})]. \tag{33}$$

The selection bias is updated outside backpropagation as

$$b_{l,e} \leftarrow b_{l,e} - \eta_b \left(u_{l,e} - \frac{1}{E_l^{\mathrm{sp}}} \right), \tag{34}$$

where $\eta_b = 10^{-4}$ is the expert-bias update rate.

We additionally use sequence-wise auxiliary balancing. For routing sequence q with $N_{l,q}$ tokens:

$$f_{l,q,e} = \frac{1}{K N_{l,q}} \sum_{n=1}^{N_{l,q}} \mathbb{I}[e \in \mathcal{K}_l(\mathbf{h}_{l,q,n})], \tag{35}$$

$$\bar{p}_{l,q,e} = \frac{1}{N_{l,q}} \sum_{n=1}^{N_{l,q}} p_{l,e}(\mathbf{h}_{l,q,n}). \tag{36}$$

The layer-wise auxiliary loss is

$$\mathcal{L}_{\mathrm{aux},l} = \frac{E_l^{\mathrm{sp}}}{Q_l} \sum_{q=1}^{Q_l} \sum_{e \in \mathcal{E}_l^{\mathrm{sp}}} f_{l,q,e} \bar{p}_{l,q,e}, \tag{37}$$

where Q_l is the number of routing sequences in layer l.

B.4 Decoder

B.4.1 EEG Reconstructor at Pre-training Stage

During pre-training, the reconstructor maps the encoder context $\mathbf{h}$ to the channel space via a symmetric decoding mechanism employing learnable queries and Transformers. The module reconstructs both waveform samples and spectral targets in $\mathbf{x}_{\mathrm{ref}}$, with masked patches reconstruction and future patches forecasting objectives.

B.4.2 DUAL-STREAM AVERAGE POOLING MLP HEAD AT FINE-TUNING STAGE

For unified fine-tuning, we append a compact classification head to the encoder context $\mathbf{h} \in \mathbb{R}^{B\times T\times G\times D_m}$. We implement a dual-stream architecture that processes representations from two complementary perspectives: (1) a temporal branch that averages over query groups to yield a tensor of shape $\mathbb{R}^{B\times T\times D_m}$, which is subsequently flattened and mapped to $\mathbb{R}^{B\times D_m}$ via an MLP; and (2) a query branch that performs temporal averaging to produce a representation in $\mathbb{R}^{B\times G\times D_m}$, followed by a corresponding MLP. The resulting features are concatenated, normalized, fused via a linear layer, and projected to task logits. This configuration is designed to surpass the performance of standard average pooling while ensuring greater memory efficiency than fully flattened architectures.

B.5 PRE-TRAINING AND FINE-TUNING STRATEGY

B.5.1 PRE-TRAINING STAGE

Activation within Tokenizer. During pre-training, to ensure comprehensive optimization of all group priors, we fully activate both static and dynamic group queries.

Loss Function. We pre-train PRiSE-EEG using a self-supervised framework comprising multiple objectives:

1. **Masked Auto-Encoding (MAE):** We randomly mask 50% of the input along either temporal patches (MAE-TP) or channels (MAE-CH) and reconstruct the masked targets.
2. **Generative Pre-training (GPT):** We train an auto-regressive objective to predict the next patch conditioned on previous patches via causal attention.
3. **Auxiliary Sequence Routing Loss (Aux):** To prevent expert collapse, we impose a load balancing loss that encourages uniform patch distribution across experts.

We employ the reconstruction losses formulated as:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{|\Omega|}\sum_{i\in\Omega}\left\|\hat{\mathbf{x}}_i - \mathbf{x}_{\text{ref},i}\right\|_2, \tag{38}$$

where Ω denotes the set of masked patch indices, and

$$\mathcal{L}_{\text{GPT}} = \frac{1}{C(T-1)}\sum_{n=1}^{C}\sum_{t=1}^{T-1}\left\|\hat{\mathbf{x}}_{n,t+1} - \mathbf{x}_{\text{ref},n,t+1}\right\|_2. \tag{39}$$

The bias-based load correction is defined in Eq. 34. The sequence-wise terms in Eq. 37 are averaged over the L_E encoder layers:

$$\mathcal{L}_{\text{Aux}} = \frac{1}{L_E}\sum_{l=1}^{L_E}\mathcal{L}_{\text{aux},l}. \tag{40}$$

The overall pre-training loss is:

$$\mathcal{L}_{pt} = \frac{1}{3}\left(\mathcal{L}_{\text{MAE-TP}} + \mathcal{L}_{\text{MAE-CH}} + \mathcal{L}_{\text{GPT}}\right) + \beta\mathcal{L}_{\text{Aux}}. \tag{41}$$

where β governs the weight of the expert load balancing constraint.

B.5.2 UNIFIED FINE-TUNING STAGE

Activation within Tokenizer. Recognizing that the relevance of specific group priors varies across tasks, we apply differentiable top-k selection over the candidate dynamic groups using Gumbel-Softmax (Jang et al., 2017), activating a subset of k groups with their dynamic priors $\mathbf{M}_d$. The Gumbel temperature is annealed via an exponential decay to gradually sharpen the selection distribution.

Loss Function. We fine-tune the shared pre-trained encoder jointly across multiple downstream datasets, equipping each task with a dedicated **dual-stream classification head**. For a set of K tasks, the joint fine-tuning objective is defined as:

$$\mathcal{L}_{ft} = \sum_{k=1}^{K}\mathcal{L}_{\text{cls}}^{k} + \beta\mathcal{L}_{\text{Aux}}. \tag{42}$$

Each task employs a weighted cross-entropy loss:

$$\mathcal{L}_{\text{cls}}^{k} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C_k} w_c^k y_{ic}^k \log(p_\theta(y_{ic}^k|\mathbf{x}_i)), \tag{43}$$

where N is the batch size, C_k is the number of classes for task k, w_c^k is the class weight, y_{ic}^k is the ground truth label, and p_θ is the predicted probability. This unified strategy facilitates cross-dataset transfer and enhances data efficiency within a shared encoder.

C Gradient and Routing Analysis

We analyze encoder gradients to quantify whether different downstream datasets impose aligned or conflicting updates on the same shared encoder. Let $\mathcal{D} = \{D_1, \ldots, D_M\}$ be the downstream datasets. Starting from a fixed pretrained checkpoint, we attach dataset-specific classification heads and run Q consecutive optimization steps for each dataset without updating the checkpoint used for other datasets. Unless otherwise stated, $Q = 512$.

C.1 Projected Encoder Gradients

We partition encoder parameters into semantic groups $g \in \mathcal{G}$, such as tokenizer, temporal layers, group layers, or backbone FFNs. At step q for dataset D_i, let $\mathbf{g}_{g,i}^{(q)}$ be the flattened encoder gradient of group g. To reduce memory cost, we project gradients into $\mathbb{R}^d$ with a fixed CountSketch (Charikar et al., 2004) projector $\mathbf{S}_g$:

$$\mathbf{v}_{g,i}^{(q)} = \mathbf{S}_g \mathbf{g}_{g,i}^{(q)} \in \mathbb{R}^d. \tag{44}$$

The same projector $\mathbf{S}_g$ is used for all datasets and steps within group g.

C.2 Mean Gradient Cosine Similarity

For each group g and dataset D_i, we compute the normalized mean update direction

$$\bar{\mathbf{v}}_{g,i} = \frac{\sum_{q=1}^{Q} \mathbf{v}_{g,i}^{(q)}}{\left\|\sum_{q=1}^{Q} \mathbf{v}_{g,i}^{(q)}\right\|_2 + \epsilon}. \tag{45}$$

The pairwise mean gradient cosine similarity is

$$C_g(i,j) = \langle \bar{\mathbf{v}}_{g,i}, \bar{\mathbf{v}}_{g,j} \rangle . \tag{46}$$

Positive values indicate aligned average update directions, while negative values indicate conflicting mean directions. When reporting an overall heatmap, we average over groups:

$$C(i,j) = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} C_g(i,j). \tag{47}$$

C.3 Gradient Subspace Affinity

Mean directions can miss cases where two datasets share dominant update modes but have different average gradients. Therefore, we also compare low-rank gradient subspaces. For group g and dataset D_i, we stack projected gradients into

$$\mathbf{X}_{g,i} = \begin{bmatrix} (\mathbf{v}_{g,i}^{(1)})^\top \\ \cdots \\ (\mathbf{v}_{g,i}^{(Q)})^\top \end{bmatrix} \in \mathbb{R}^{Q\times d}. \tag{48}$$

Let $\mathbf{U}_{g,i} \in \mathbb{R}^{d\times k}$ be the top-k PCA basis of $\mathbf{X}_{g,i}$ with orthonormal columns. We use $k = 7$ by default. For datasets D_i and D_j, the subspace affinity is

$$\text{Affinity}_g(i,j) = \frac{1}{k}\sum_{\ell=1}^{k} \sigma_\ell \left(\mathbf{U}_{g,i}^\top \mathbf{U}_{g,j}\right), \tag{49}$$

where $\sigma_\ell(\cdot)$ is the ℓ-th singular value. Values close to 1 indicate highly overlapping dominant gradient subspaces, while values close to 0 indicate distinct update subspaces. The overall affinity is

$$\text{Affinity}(i,j) = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \text{Affinity}_g(i,j). \tag{50}$$

C.4 Expert Routing Similarity

To compare gradient geometry with expert specialization, we also compute the cumulative routing distribution of each dataset. Let $\mathbf{p}_i \in \mathbb{R}^E$ be the normalized count or probability mass assigned to all experts by samples from dataset D_i. The routing similarity is

$$R(i,j) = \frac{\langle \mathbf{p}_i, \mathbf{p}_j \rangle}{\|\mathbf{p}_i\|_2 \|\mathbf{p}_j\|_2 + \epsilon}. \tag{51}$$

Expert routing similarity are then averaged across encoder Transformer layers. Comparing $\text{Affinity}(i,j)$ and $R(i,j)$ shows whether datasets with similar dominant gradient subspaces also induce similar expert usage.

D Experimental Setup

D.1 Data Preprocessing

The preprocessing pipeline utilizes a systematic framework for managing heterogeneous EEG data using MNE-Python (Gramfort et al., 2013). The pipeline begins with a resampling stage to transform all source signals to $f_s = 256$ Hz, which standardizes patch division and ensures parity with the sampling rates used in baseline models. A band-pass Finite Impulse Response (FIR) filter (1-100 Hz), implemented via an overlap-add method, is applied to attenuate low-frequency noise and high-frequency artifacts. This method ensures filtering consistency across signals of varying durations. A notch filter is then applied at either 50 Hz or 60 Hz to remove power-line interference, with the frequency selected according to a manual review of each dataset's spectral profile or geographical origin. Then time series are segmented into consecutive 1-second patches. Electrode configurations are then mapped to a standardized 10-10 universal channel set. Additionally, data units are converted from microvolts (μV) to Volts to ensure compatibility with MNE-Python conventions. To optimize data handling, processed EEG signals are serialized in the `Parquet` format using Zstandard compression for storage efficiency, or `Apache Arrow` for high-throughput memory mapping. To facilitate large-scale distributed training, the pipeline supports direct dataset access from remote object storage via the S3 protocol. The implementation employs parallel execution to process the multi-terabyte pre-training corpus efficiently.

D.2 Comparing Baselines

We evaluate our framework against seven state-of-the-art EEG foundation models: BENDR (Kostas et al., 2021), BIOT (Yang et al., 2023), LaBraM (Jiang et al., 2024), EEGPT (Wang et al., 2024), CBraMod (Wang et al., 2025), CSBrain (Zhou et al., 2025), and REVE (Ouahidi et al., 2025). BENDR (Kostas et al., 2021) (3.97 M parameters) applies a BERT-style objective to a large-scale clinical EEG corpus. BIOT (Yang et al., 2023) (3.19 M) manages data heterogeneity by tokenizing various biosignals into a unified, sequence-based structure. LaBraM ($5.82M$) is pre-trained on 2,500 hours of data with integrated VQ-VAE (Van Den Oord et al., 2017) modules for dual-domain (frequency/phase) mask learning, EEGPT ($25.29M$) combines dual self-supervised universal representation learning and stabilization mechanisms, and CBraMod (Wang et al., 2025) ($4.92M$) utilizes criss-cross attention to capture the spatial-temporal separately at the same transformer layer, leading to refined representation ability. CSBrain (Zhou et al., 2025) ($8.86M$) employs cross-scale spatiotemporal tokenization (CST) and structured sparse attention to obtain robust EEG representations. REVE (Ouahidi et al., 2025) ($69.19M$) is characterized by the use of the largest pre-training dataset available and 4D Fourier positional encoding, which is paired with an improved MAE reconstruction objective. In addition, we select two strong supervised methods for comparison, and two supervised methods (EEGNet (Lawhern et al., 2018), EEGConformer (Song et al., 2022)). Among these methods, BENDR, BIOT, LaBraM, CBraMod, CSBrain and REVE exploit fully flatten MLP and fine-tune

on different downstream tasks separately during the evaluation. However, EEGPT exploits linear probe to frozen pre-trained backbone, making their framework support the multi-task evaluation with different linear probes. During experiment, we follow their own strategy and open-sourced hyperparameter settings to reproduce them on 12 downstream tasks based on their pre-trained EEG foundation models.

D.3 Evaluation Strategy

To partition each dataset, we perform a **subject-level split** for most of the datasets. Subject-dependent strategy is adopted for SEED and SEED-V to have a comparable metric to other baselines. We also employ a **greedy, multi-label stratified** splitting algorithm to ensure label distributions remain balanced across the training, validation, and test sets to predefined ratios:

- **Siena**: Subject 0-7, 9-13, 16-17 are assigned to the training, valid, test set correspondingly;
- **SEED**: Following prior research, the 15 trials are divided into three sets in a 9:3:3 ratio, and all sessions are merged together;
- **SEED-V**: Following prior research, the 15 trials are divided into three sets in a 1:1:1 ratio;
- **PhysioMI**: Subject 1-70,71-89, 90-109 are assigned to the training set, valid set, test set correspondingly;
- **Mimul-11**: Stratified splitting is employed to achieve ratios of 0.76, 0.12, and 0.12 for training, validating, and testing;
- **Workload**: Stratified splitting is employed to achieve ratios of 0.72, 0.14, and 0.14 for training, validating, and testing;
- **HMC**: Subjects are randomly split into training, validation, and test sets at a ratio of 103:24:24;
- **TUEV** and **TUSL**: Owing to the highly imbalanced label distribution in these datasets, the stratified splitting function is employed to create three splits from all the data, approximately aligning with predefined ratios of 0.8, 0.1, and 0.1;
- **TUAB**: The validation and test sets are obtained by equally splitting the original evaluation set by subject, while the training set remains unchanged;
- **Things-EEG-2**: Split by subject with predefined ratios of 0.6, 0.2, and 0.2;
- **ADFTD**: Stratified splitting is employed to achieve ratios of 0.70, 0.15, and 0.15 for training, validating, and testing.

Table 6: Configurations for model variants.

	PRiSE-EEG-S	PRiSE-EEG-M	PRiSE-EEG-B	PRiSE-EEG-L	PRiSE-EEG-XL
#Parameters	9.41 M	27.06 M	41.06 M	78.86 M	125.58 M
#Activated Parameters	6.46 M	11.73 M	18.05 M	36.39 M	64.85 M
Model Dim	192	256	384	512	640
MoE Expert Dim	256	256	256	256	256
Tokenizer Cross Attention Block	1	2	2	2	2
Group MoE Block	3	3	4	6	8
Temporal MoE Block	3	3	4	6	8
Decoder Block	1	2	2	3	4
Batch Size	256	128	128	128	64
Attn. Head	4	4	8	8	8

D.4 Evaluation Metrics

To account for class imbalance in the downstream datasets, we utilize the following evaluation metrics:

- **Balanced Accuracy**: the arithmetic mean of recall (sensitivity) across all classes, mitigating the impact of imbalanced class distributions. It is effective for evaluating classification models on datasets with significant disparities in class proportions, which is formulated as:

$$\text{B-Acc} = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FN_i}, \tag{52}$$

Table 7: Region electrode groupings used in the Prior-Guided Continuous Tokenization.

Group	Electrodes
Prefrontal	FP1, FPZ, FP2, AF9, AF7, AF5, AF3, AF1, AFZ, AF2, AF4, AF6, AF8, AF10
Frontal	F9, F7, F5, F3, F1, FZ, F2, F4, F6, F8, F10, FT9, FT7, FC5, FC3, FC1, FCZ, FC2, FC4, FC6, FT8, FT10
Central	FC1, FCZ, FC2, C5, C3, C1, CZ, C2, C4, C6, CP1, CPZ, CP2
Parietal	TP9, TP7, CP5, CP3, CP1, CPZ, CP2, CP4, CP6, TP8, TP10, P9, P7, P5, P3, P1, PZ, P2, P4, P6, P8, P10
Occipital	PO9, PO7, PO5, PO3, PO1, POZ, PO2, PO4, PO6, PO8, PO10, O1, OZ, O2, I1, IZ, I2
Left Temporal	T1, T7, T9, FT7, FT9, TP7, TP9
Right Temporal	T2, T8, T10, FT8, FT10, TP8, TP10
Midline	FPZ, AFZ, FZ, FCZ, CZ, CPZ, PZ, POZ, OZ, IZ

where C is the number of classes, TP_i and FN_i denote true positives and false negatives for class i.

- **Weighted F1**: a harmonic mean of precision and recall, weighted by the number of true instances in each class. This metric accounts for class imbalance by assigning higher importance to classes with larger sample sizes, ensuring a more representative evaluation of model effectiveness, which is formulated as

$$\text{Pre}_i = \frac{TP_i}{TP_i + FP_i}, \quad \text{Rec}_i = \frac{TP_i}{TP_i + FN_i}, \tag{53}$$

$$\text{W-F1} = \sum_{i=1}^{C} w_i \cdot \frac{2 \cdot \text{Pre}_i \cdot \text{Rec}_i}{\text{Pre}_i + \text{Rec}_i}, \tag{54}$$

  where FP_i denotes false positives for class i, and w_i is the weight of class i based on its support.

- **AUROC**: area under the ROC curve. It reflects the model's ability to discriminate between classes across all possible decision boundaries, which is formulated as

$$\text{TPR} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{FP}{FP + TN}, \tag{55}$$

$$\text{AUROC} = \int_0^1 \text{TPR}(f)\, df, \quad f = \text{FPR}. \tag{56}$$

- **AUC-PR**: area under the precision-recall curve. It provides a holistic evaluation of model performance under class imbalance, which is formulated as

$$\text{AUC-PR} = \int_0^1 \text{Pre}(r)\, dr, \quad r = \text{Rec}. \tag{57}$$

- **Cohen's Kappa**: the agreement level between predicted and true labels by comparing observed and expected frequencies along the diagonal of a confusion matrix. It is particularly suited for multi-class classification scenarios, which is formulated as

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{58}$$

  where p_o and p_e are observed and expected agreement.

Among these metrics, AUROC and AUC-PR are used to evaluate binary classification tasks, while Cohen's Kappa and Weighted F1 are applied to multi-category classification. Together, these metrics provide a robust evaluation framework under class imbalance.

Table 8: Network electrode groupings used in the Prior-Guided Continuous Tokenization.

Group	Electrodes
DMN	FP1, FPZ, FP2, AFZ, FZ, FCZ, CZ, CPZ, PZ, POZ, OZ, P5, P7, P6, P8, TP7, TP8, AF1, AF2
ECN	AF1, AF2, AF3, AF4, AF5, AF6, AF7, AF8, AF9, AF10, F1, F3, F5, F2, F4, F6, FC1, FC3, FC2, FC4, P1, P3, P2, P4, CP1, CP3, CP2, CP4
SN	FZ, FCZ, CZ, F1, F2, FC1, FC2, C1, C2, FC3, FC4, C3, C4
DAN	F1, F3, F2, F4, P1, P3, P5, P2, P4, P6, PO3, PO4
VAN	F10, F8, FT10, FT8, T8, TP8, C6, CP6, P10, P8, P6, T10, TP10
Visual	P9, P10, PO9, PO7, PO5, PO3, PO1, POZ, PO2, PO4, PO6, PO8, PO10, O1, OZ, O2, I1, IZ, I2, T10
Somatomotor	FC5, FC3, FC1, FC2, FC4, FC6, C5, C3, C1, CZ, C2, C4, C6, CP5, CP3, CP1, CP2, CP4, CP6
Language	F9, F7, F5, FT9, FT7, T7, TP7, C5, CP5, P9, P7, P5, T9, TP9

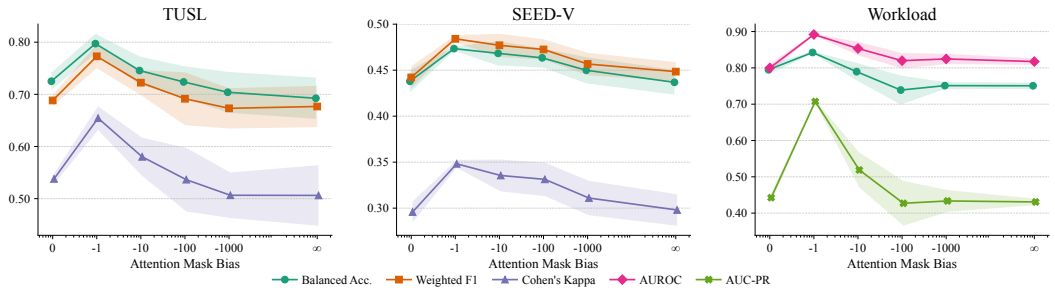


Figure 6: Ablation Study of attention bias strength for static path on TUSL, Workload and SEED-V datasets.

D.5 Training Settings

Data. To facilitate high-throughput loading, all preprocessed samples are converted into the `Arrow` format, which accelerates distributed computation and enables direct data access for the GPU.

Environment. Experiments are conducted using Python 3.11.13, PyTorch 2.7.1, and CUDA 12.8 on four NVIDIA A100 GPUs for pre-training and one A100 GPU for fine-tuning. We utilize Automatic Mixed Precision (AMP) with the `bfloat16` data type to optimize memory usage and employ `GradScaler` to maintain stability.

Optimizer. We employ AdamW with a two-phase learning rate scheduler combining linear warm-up and cosine annealing. The warm-up factor is 0.1, and the gradient clipping threshold is 1.0. For MoE load balancing, the expert-bias update rate is $\eta_b = 1 \times 10^{-4}$, and the auxiliary sequence balancing coefficient is $\beta = 1 \times 10^{-5}$. For LoRA-based adaptation, the rank (r) and scaling factor (α) are set to 8 and 16.0, respectively.

Pre-training. The pre-training learning rate is 1×10^{-4}, with attention and feedforward dropout rates set to 0.1. We apply a maximum `DropPath` rate of 0.1, utilizing a linear stochastic depth schedule that increases with layer depth. Training is conducted for 5 epochs, including a 1-epoch warm-up.

Fine-tuning. The fine-tuning learning rate is 5×10^{-5} with a weight decay of 0.05. The process spans 20 epochs, consisting of one initial epoch with a frozen backbone to stabilize classifier initialization, followed by 2 warm-up epochs. **PRiSE-EEG** is optimized in a unified multi-task setting, where the shared backbone is updated using data from all downstream sets simultaneously. The classification

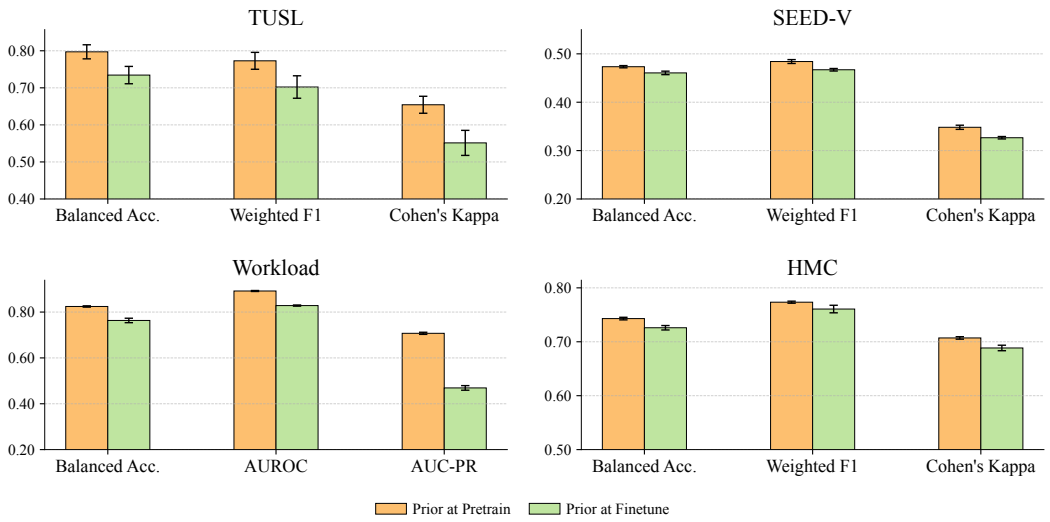


Figure 7: Ablation Study of the stage that the neural prior has been introduced on TUSL, SEED-V, HMC and Workload.

head's learning rate is five times that of the backbone. The Gumbel selector is activated during this stage with $k = 12$ and a temperature τ that decays from 1.0 to 0.1. Model selection is based on peak validation performance, with final results reported on the held-out test set for each task.

D.6 Model Configurations

PRiSE-EEG is developed in five scaling variants, with parameter counts of 9.41M, 27.06M, 41.06M, 78.86M, and 125.58M. The model scale is primarily adjusted by varying the number of Transformer blocks and the latent embedding dimensionality. The patch size is set to $f_s = 256$, corresponding to a 1-second temporal duration. For the masked auto-encoding objectives, we employ a 50% masking ratio applied independently to either temporal patches or entire channel sequences. The hidden dimensionality of the dual-stream MLP classification head is 128. Detailed specifications for each model variant are provided in Tab. 6.

D.7 Electrode Correspondence

The Prior-Guided Continuous Tokenization incorporates neuroscientific guidance by integrating priors from anatomical regions and functional brain networks. Electrodes from the 10-10 standard montage are partitioned into groups as defined in Tabs. 7 and 8. Specifically, each electrode is assigned to its spatially proximal region of interest (ROI) within the reference atlas. We utilize electrode coordinates registered in MRI space via the MNE-standard montage (Gramfort et al., 2013), with group correspondences visually verified against the Human Brainnetome Atlas (Fan et al., 2016) using MRIcroGL (Rorden, 2025). Within this partition scheme, certain electrodes are assigned to overlapping groups, reflecting their involvement in multiple functional networks or their location at anatomical boundaries. While these groupings provide an initial neuro-informed prior, we acknowledge that these assignments are subject to future refinement as more granular electrophysiological mapping data becomes available. The functional network groupings encompass the Default Mode Network (DMN), Executive Control Network (ECN), Salience Network (SN), Dorsal Attention Network (DAN), Ventral Attention Network (VAN), Somatomotor Network, Visual Network, and Language Network.

E EXTENDED RESULTS

E.1 DETAILED EXPERIMENT RESULTS

Tab. 9 presents detailed model performance metrics across the default configuration and four alternative training strategies. Multi-task unified fine-tuning achieves consistent improvements over single-task training, confirming that pre-trained representations provide a significant performance advantage compared to training from random initialization. Besides, **PRiSE-EEG demonstrates strong representation transferability even with a frozen encoder**, validating the feature extraction capabilities of the Prior-Guided Continuous Tokenization and the CKA-Calibrated Depth-Stratified Expert Transformer. LoRA-based fine-tuning performs worse than full-parameter fine-tuning under the tested configuration ($r = 8$, $\alpha = 16$). This suggests that the selected low-rank adaptation setting may provide insufficient flexibility for these tasks, rather than establishing a structural incompatibility between LoRA and MoE.

Table 9: Performance comparison of 5 configurations of PRiSE-EEG on 12 BCI tasks under **full-parameter unified**, **full-parameter single-task**, **efficient LoRA**, **freezing backbone** and **from scratch** fine-tuning.

Dataset	Metrics	Unified	Single-Task	LoRA	Freezing	Scratch	Baseline Best
SEED	B-Acc	**73.92±0.31**	72.60±0.51	59.26±0.30	64.96±0.12	58.90±0.27	71.92±0.39
	F1/AUROC	**73.29±0.22**	72.00±0.62	58.90±0.28	64.66±1.04	58.68±0.27	71.84±0.20
	Kappa/AUCPR	**61.02±0.45**	59.10±0.52	39.01±0.11	47.60±0.18	41.00±0.16	58.02±0.59
PhysioMI	B-Acc	**65.55±0.38**	63.53±1.01	32.62±0.14	41.81±1.20	30.08±1.26	64.80±0.53
	F1/AUROC	**65.83±0.56**	63.60±0.98	30.99±0.27	41.69±1.95	29.27±1.11	64.84±0.94
	Kappa/AUCPR	**54.78±0.69**	52.10±1.02	10.17±0.17	22.40±0.91	6.79±0.82	53.04±0.85
Workload	B-Acc	**82.47±0.30**	77.33±1.53	59.70±0.40	74.96±1.85	74.74±1.17	74.23±1.36
	F1/AUROC	**89.21±0.19**	84.03±1.32	83.03±0.32	83.97±1.31	82.77±1.10	82.53±0.75
	Kappa/AUCPR	**70.75±0.47**	68.17±1.47	62.66±0.43	62.16±1.30	61.09±0.95	61.48±1.90
TUEV	B-Acc	**75.28±0.51**	73.60±0.98	70.13±0.30	71.05±2.58	67.05±2.83	71.20±1.34
	F1/AUROC	**88.25±0.51**	87.90±0.79	81.87±0.57	79.87±3.18	79.03±1.57	85.07±0.85
	Kappa/AUCPR	**80.90±0.62**	80.00±0.85	73.55±0.21	75.00±1.99	67.53±1.51	74.67±1.44
TUAB	B-Acc	**82.22±0.12**	81.83±0.52	81.11±0.31	81.39±0.29	78.38±1.10	81.75±0.65
	F1/AUROC	**89.93±0.31**	88.77±0.65	89.49±0.40	88.24±0.55	85.01±1.42	89.60±1.05
	Kappa/AUCPR	**90.11±0.15**	89.03±0.61	89.86±0.33	89.14±1.16	85.16±1.18	90.00±1.56
HMC	B-Acc	**74.33±0.21**	73.03±1.35	70.83±0.13	71.09±0.43	68.57±1.79	72.18±0.24
	F1/AUROC	**77.31±0.20**	76.26±1.62	74.24±0.21	71.86±0.79	69.14±1.16	74.16±0.27
	Kappa/AUCPR	**70.69±0.19**	68.76±1.53	66.31±0.48	65.18±0.96	64.76±2.04	66.43±0.16
Siena	B-Acc	**86.89±2.31**	84.31±0.77	79.54±0.52	80.81±2.46	74.22±1.29	83.96±0.55
	F1/AUROC	89.12±2.30	90.30±1.94	87.63±2.61	88.07±2.56	82.90±2.17	**93.15±1.49**
	Kappa/AUCPR	**99.85±0.03**	99.83±0.03	99.81±0.05	99.81±0.16	99.71±0.03	99.84±0.04
TUSL	B-Acc	**79.82±1.95**	78.33±1.59	33.33±0.00	61.30±2.93	62.44±1.75	77.54±1.95
	F1/AUROC	**77.34±2.30**	76.13±2.35	25.71±0.03	43.66±3.46	54.64±1.72	73.63±2.70
	Kappa/AUCPR	**65.45±2.31**	64.60±2.86	00.00±0.00	31.30±2.01	41.21±1.57	65.12±3.48
Mimul-11	B-Acc	**51.33±0.24**	49.60±0.42	40.13±0.11	44.00±2.11	41.01±1.10	50.51±0.63
	F1/AUROC	**58.82±0.23**	55.23±0.55	46.77±0.14	48.92±1.91	44.59±0.97	58.09±0.61
	Kappa/AUCPR	31.49±0.98	29.13±1.10	13.37±0.28	15.24±1.76	12.09±2.15	**33.63±1.80**
Things EEG 2	B-Acc	**65.15±0.48**	64.14±0.91	50.00±0.01	55.46±1.50	53.38±1.75	63.76±1.57
	F1/AUROC	**86.63±0.35**	75.02±0.84	59.97±0.10	61.48±0.46	57.56±1.66	77.54±0.70
	Kappa/AUCPR	42.82±1.19	**43.70±1.96**	15.78±0.02	17.21±0.34	15.08±1.14	40.89±1.20
SEED-V	B-Acc	**47.40±0.25**	42.27±1.52	22.06±0.12	29.58±0.92	26.30±1.51	40.98±1.58
	F1/AUROC	**48.39±0.42**	42.42±1.32	16.61±0.19	29.78±1.16	25.85±1.45	41.18±1.46
	Kappa/AUCPR	**34.84±0.45**	27.59±1.33	2.56±0.25	12.01±1.09	8.62±1.71	25.58±1.87
ADFTD	B-Acc	**58.75±0.66**	56.43±0.95	49.82±0.23	51.58±1.56	49.46±2.40	52.20±0.93
	F1/AUROC	**64.12±0.63**	59.93±1.15	47.26±0.38	53.57±2.54	54.08±2.06	55.29±1.50
	Kappa/AUCPR	**39.41±0.79**	37.07±1.36	28.96±0.29	28.71±2.51	28.41±1.48	31.38±1.82

E.2 MULTI-TASK COMPARISON WITH BASELINES

Table 10 reports an additional full-parameter multi-task comparison with BENDR, BIOT, LaBraM, EEGPT, CBraMod, CSBrain, and REVE. All seven baselines use average-pooling classification heads. This comparison assesses performance when the existing models also use joint training. PRiSE-EEG achieves Balanced Accuracy of 82.47%, 65.55%, 74.33%, and 47.40% on Workload, PhysioMI, HMC, and SEED-V, respectively; the highest baseline means are 74.67%, 58.82%, 72.18%, and 43.26%. But baselines still obtain higher means on TUSL, illustrating task-dependent differences.

Table 10: Performance comparison of baseline models and PRiSE-EEG under full-parameter multi-task results on the 12 BCI tasks.

Dataset	Metric	BENDR	BIOT	LaBraM	EEGPT	CBraMod	CSBrain	REVE	PRiSE-EEG
SEED	B-Acc	60.82 ± 0.19	65.62 ± 0.92	67.71 ± 0.55	68.75 ± 0.40	72.25 ± 1.90	71.92 ± 0.39	72.62 ± 0.92	$\mathbf{73.92 \pm 0.31}$
	W-F1	60.84 ± 0.20	65.57 ± 0.81	68.11 ± 0.66	67.42 ± 0.78	72.90 ± 1.63	71.84 ± 0.20	72.75 ± 0.76	$\mathbf{73.29 \pm 0.22}$
	Kappa	41.59 ± 0.16	50.11 ± 0.99	52.22 ± 0.83	52.37 ± 0.28	58.68 ± 2.91	58.02 ± 0.59	$\mathbf{62.57 \pm 0.94}$	61.02 ± 0.45
PhysioMI	B-Acc	44.77 ± 0.31	33.02 ± 0.29	43.19 ± 1.03	50.52 ± 0.80	41.80 ± 0.69	55.43 ± 0.37	58.82 ± 0.45	$\mathbf{65.55 \pm 0.38}$
	W-F1	44.63 ± 0.03	32.75 ± 0.36	42.38 ± 1.15	50.26 ± 0.76	40.35 ± 0.87	55.10 ± 0.63	58.77 ± 0.56	$\mathbf{65.83 \pm 0.56}$
	Kappa	26.33 ± 0.40	10.67 ± 0.40	24.25 ± 1.37	34.00 ± 1.06	22.42 ± 0.92	40.48 ± 0.48	45.10 ± 0.60	$\mathbf{54.78 \pm 0.69}$
Workload	B-Acc	63.32 ± 2.29	71.37 ± 2.19	67.77 ± 0.95	72.66 ± 0.90	74.12 ± 0.79	71.27 ± 0.75	74.67 ± 1.46	$\mathbf{82.47 \pm 0.30}$
	AUROC	74.09 ± 1.04	72.43 ± 1.94	77.32 ± 2.88	81.16 ± 3.42	83.89 ± 0.93	79.63 ± 1.20	82.32 ± 1.47	$\mathbf{89.21 \pm 0.19}$
	AUC-PR	54.78 ± 1.84	69.52 ± 0.23	61.13 ± 3.03	64.30 ± 6.48	$\mathbf{72.58 \pm 1.55}$	55.10 ± 2.21	68.22 ± 3.92	70.75 ± 0.47
TUEV	B-Acc	67.46 ± 2.59	61.35 ± 1.35	71.53 ± 0.19	70.85 ± 1.16	69.41 ± 1.86	71.82 ± 0.69	74.27 ± 0.87	$\mathbf{75.28 \pm 0.51}$
	W-F1	84.56 ± 1.82	80.55 ± 0.46	86.22 ± 0.73	86.35 ± 0.47	83.44 ± 0.31	87.00 ± 0.64	$\mathbf{89.37 \pm 0.33}$	88.25 ± 0.51
	Kappa	72.85 ± 2.87	68.37 ± 0.61	77.42 ± 1.20	77.69 ± 1.06	72.02 ± 0.65	77.84 ± 1.55	$\mathbf{82.12 \pm 0.76}$	80.90 ± 0.62
TUAB	B-Acc	81.72 ± 0.26	80.85 ± 0.36	80.47 ± 0.12	80.68 ± 0.37	81.55 ± 0.14	79.18 ± 0.48	81.18 ± 0.46	$\mathbf{82.22 \pm 0.12}$
	AUROC	89.88 ± 0.27	88.67 ± 0.62	88.33 ± 0.59	$\mathbf{90.10 \pm 0.52}$	89.57 ± 0.26	86.32 ± 0.28	88.32 ± 0.37	89.93 ± 0.31
	AUC-PR	90.22 ± 0.15	89.01 ± 0.47	88.65 ± 0.28	$\mathbf{90.35 \pm 0.54}$	89.67 ± 0.02	85.78 ± 0.25	89.33 ± 0.61	90.11 ± 0.15
HMC	B-Acc	70.21 ± 0.72	71.45 ± 1.05	71.63 ± 1.04	71.30 ± 0.91	72.03 ± 0.14	72.18 ± 0.24	71.93 ± 0.45	$\mathbf{74.33 \pm 0.21}$
	W-F1	72.44 ± 0.65	72.89 ± 0.34	72.28 ± 1.08	72.46 ± 0.96	75.20 ± 0.29	74.16 ± 0.27	72.07 ± 0.73	$\mathbf{77.31 \pm 0.20}$
	Kappa	64.57 ± 0.90	64.89 ± 0.71	65.22 ± 1.23	64.86 ± 0.77	67.43 ± 0.12	66.43 ± 0.16	64.40 ± 0.54	$\mathbf{70.69 \pm 0.19}$
Siena	B-Acc	78.96 ± 3.08	74.35 ± 0.71	73.62 ± 0.62	83.28 ± 1.56	82.75 ± 2.10	76.42 ± 1.43	81.87 ± 2.03	$\mathbf{86.89 \pm 2.31}$
	AUROC	91.31 ± 4.74	81.20 ± 0.92	89.43 ± 3.59	$\mathbf{93.44 \pm 0.12}$	92.03 ± 1.90	92.70 ± 0.67	90.70 ± 1.25	89.12 ± 2.30
	AUC-PR	99.81 ± 0.11	99.26 ± 0.05	99.82 ± 0.07	99.86 ± 0.00	99.83 ± 0.04	99.87 ± 0.05	$\mathbf{99.88 \pm 0.03}$	99.85 ± 0.03
TUSL	B-Acc	73.35 ± 0.98	69.90 ± 4.10	75.86 ± 1.15	76.12 ± 2.05	79.56 ± 3.24	$\mathbf{80.56 \pm 0.91}$	79.85 ± 1.27	79.82 ± 1.95
	W-F1	70.90 ± 1.73	71.03 ± 5.19	72.74 ± 1.33	$\mathbf{84.19 \pm 1.04}$	76.24 ± 4.45	76.72 ± 0.38	77.81 ± 1.89	77.34 ± 2.30
	Kappa	56.65 ± 2.55	56.00 ± 7.50	59.60 ± 1.81	$\mathbf{84.30 \pm 1.27}$	64.72 ± 5.96	65.57 ± 0.41	66.62 ± 2.35	65.45 ± 2.31
Mimul-11	B-Acc	50.98 ± 1.52	40.43 ± 0.63	45.05 ± 0.66	50.77 ± 0.61	44.53 ± 0.19	41.64 ± 0.54	45.22 ± 0.13	$\mathbf{51.33 \pm 0.24}$
	W-F1	57.85 ± 2.80	49.02 ± 0.73	53.42 ± 1.98	57.44 ± 1.11	52.80 ± 0.57	50.70 ± 0.61	53.73 ± 0.65	$\mathbf{58.82 \pm 0.23}$
	Kappa	$\mathbf{36.72 \pm 4.93}$	14.63 ± 1.88	22.60 ± 2.20	31.16 ± 2.43	19.23 ± 0.77	15.94 ± 1.33	21.53 ± 1.16	31.49 ± 0.98
Things-EEG-2	B-Acc	63.62 ± 1.38	54.15 ± 0.25	56.47 ± 0.32	51.54 ± 0.76	59.13 ± 0.72	57.49 ± 0.27	59.43 ± 1.20	$\mathbf{65.15 \pm 0.48}$
	AUROC	77.23 ± 0.88	63.01 ± 0.52	62.82 ± 0.84	66.64 ± 0.04	69.58 ± 1.41	62.47 ± 1.27	69.00 ± 0.53	$\mathbf{86.63 \pm 0.35}$
	AUC-PR	42.29 ± 1.75	20.12 ± 0.33	22.23 ± 0.64	20.30 ± 0.65	31.85 ± 1.21	24.66 ± 1.93	32.07 ± 1.85	$\mathbf{42.82 \pm 1.19}$
SEED-V	B-Acc	21.75 ± 0.66	30.54 ± 0.18	43.26 ± 0.29	43.07 ± 0.60	40.56 ± 0.77	38.02 ± 0.23	40.84 ± 0.98	$\mathbf{47.40 \pm 0.25}$
	W-F1	19.94 ± 2.58	30.97 ± 0.26	43.20 ± 0.81	42.99 ± 0.54	41.72 ± 0.72	38.07 ± 1.03	40.58 ± 2.00	$\mathbf{48.39 \pm 0.42}$
	Kappa	2.90 ± 0.68	14.14 ± 0.21	28.93 ± 0.50	28.62 ± 0.86	26.22 ± 0.97	22.48 ± 0.42	25.59 ± 1.48	$\mathbf{34.84 \pm 0.45}$
ADFTD	B-Acc	55.71 ± 3.76	52.40 ± 1.98	50.27 ± 1.74	52.91 ± 1.57	51.95 ± 2.84	48.96 ± 2.65	52.12 ± 3.78	$\mathbf{58.75 \pm 0.66}$
	W-F1	56.97 ± 2.75	53.48 ± 2.71	52.36 ± 2.23	54.82 ± 1.30	54.41 ± 3.13	50.17 ± 3.76	51.10 ± 3.90	$\mathbf{64.12 \pm 0.63}$
	Kappa	33.98 ± 4.63	36.27 ± 3.65	26.17 ± 3.15	31.14 ± 1.85	30.06 ± 4.35	24.96 ± 4.13	30.70 ± 6.41	$\mathbf{39.41 \pm 0.79}$

Table 11: Balanced Accuracy for the full model and ablation variants (mean $\pm$ standard deviation, %).

Configuration	TUSL	SEED-V	Workload	HMC
Full PRiSE-EEG	79.82 ± 1.95	47.40 ± 0.25	82.47 ± 0.30	74.33 ± 0.21
Static path only	67.31 ± 1.65	43.41 ± 0.20	78.15 ± 0.49	71.13 ± 0.59
Dynamic path only	64.89 ± 1.77	40.75 ± 0.86	74.93 ± 1.92	72.59 ± 0.30
Without Gumbel selection	69.11 ± 2.02	46.31 ± 0.44	78.89 ± 0.43	72.24 ± 0.39
Two additional shared experts per layer	77.12 ± 1.40	46.12 ± 0.60	80.11 ± 0.67	72.01 ± 0.69
Standard dense FFN	73.88 ± 0.94	45.19 ± 0.45	79.25 ± 0.60	71.40 ± 0.42
Reversed expert allocation	58.80 ± 2.72	29.58 ± 1.99	65.26 ± 1.36	66.40 ± 1.14
Flattened classification head	77.92 ± 1.01	39.41 ± 0.49	79.56 ± 0.73	72.70 ± 0.30

E.3 Additional Ablation Results

Table 11 lists the detailed numbers for Balanced Accuracy of the ablation variants shown in Fig. 3. The fixed allocation uses two additional shared experts from the eight-expert pool, giving three shared and six specialized experts in total per layer. Using the full-model values as the reference, combining static and dynamic paths yields higher Balanced Accuracy than either path alone on all four tasks. Relative to this fixed allocation, the depth-stratified configuration improves TUSL, SEED-V, Workload, and HMC by 2.70%, 1.28%, 2.36%, and 2.32%, respectively. Reversing the allocation lowers the corresponding results by 21.02%, 17.82%, 17.21%, and 7.93%.

We first evaluate the sensitivity of the attention bias strength (ranging from 0 to -10^6) within the Prior-Guided Continuous Tokenization, as illustrated in Fig. 6. Compared to the baseline without prior guidance (bias = 0), the application of a soft attention bias yields measurable performance

gains of 5-10% across the TUSL, Workload, and SEED-V datasets. However, as the bias strength approaches a hard mask configuration, performance across all three datasets progressively declines. Furthermore, enforcing a hard mask (bias = $-\infty$) induces training instability during pre-training, as the model fails to optimize the parameters in the layers preceding the tokenizer. This phenomenon is likely attributable to an information bottleneck generated by the sparse attention mask, which obstructs effective gradient propagation. These findings validate the tokenizer design, confirming that a soft, non-zero attention bias is optimal for learning versatile EEG representations. We also investigate the optimal stage for introducing neuroscientific priors. As shown in Fig. 7, incorporating prior knowledge during pre-training results in substantially better downstream performance compared to its introduction during the fine-tuning stage alone. This result underscores the importance of anchoring the model to neuroscientific principles from the onset of the learning process.

E.4 Prior Preservation and Grouping Sanity Check

PRiSE-EEG uses static region-network priors only as weak sensor-space biases, and the learnable relaxation term Δ_{pri} is expected to adjust rather than overwrite the initial grouping structure. We therefore measure whether the optimized prior matrix $\mathbf{M}_s + \Delta_{\mathrm{pri}}$ preserves the coarse in-group/out-group separation after training. In Tab. 12, we report the relative update magnitude $\|\Delta_{\mathrm{pri}}\|/\|\mathbf{M}_s\|$, the average in-group bias, the average out-group bias, and the in-group minus out-group gap. We also compare against a random learnable grouping baseline under the same training budget.

Table 12: Prior preservation and grouping sanity check. The full PRiSE-EEG prior keeps a clear gap after training, while random learnable grouping shows a weaker structural separation.

Variant	$\|\Delta_{\mathrm{pri}}\|/\|\mathbf{M}_s\| \downarrow$	In-group bias $\uparrow$	Out-group bias $\downarrow$	In-group $-$ out-group $\uparrow$
Full PRiSE-EEG prior	0.2801 ± 0.0517	-0.1638 ± 0.0405	-1.0487 ± 0.0310	0.8849
Random learnable grouping	0.4345 ± 0.0892	-0.6911 ± 0.0783	-0.8255 ± 0.0712	0.1344

The learned update remains modest relative to the initial prior matrix, and the in-group/out-group gap is largely preserved. This indicates that the prior relaxation term provides data-driven correction without collapsing the predefined sensor-group structure.

E.5 Comparison with Recent EEG Tokenizers

MMM organizes spatial information using a unified topology and a multi-level channel hierarchy, with regional representations supporting different channel configurations focusing on emotion recognition (Yi et al., 2023). PRiSE-EEG shares the idea of structured spatial aggregation, but turns static region-network groupings into learnable soft attention biases and adds sample-dependent short-time dynamic groups. Soft biases allow out-of-group channels to participate, while dynamic groups supplement sample- and time-varying channel relations. Together they form continuous multi-channel tokens; the subsequent CKA-calibrated experts organize shared and specialized capacity across depth.

We compare PRiSE-EEG with recent tokenizer-centered EEG foundation models. CodeBrain uses a decoupled temporal-frequency discrete tokenizer with a multi-scale EEGSSM backbone, while TFM-Tokenizer learns single-channel time-frequency motif tokens. In contrast, PRiSE-EEG keeps EEG patches continuous and multi-channel, and combines weak prior-guided tokenization with CKA-calibrated depth-stratified experts. PRiSE-EEG outperforms both discrete-tokenizer baselines on the shared evaluation datasets under the matched pipeline. These results suggest that continuous multi-channel prior-guided tokenization is complementary to discrete EEG vocabularies, and that the proposed depth-stratified expert allocation provides additional cross-paradigm capacity.

Table 13: Comparison with recent tokenizer-centered EEG foundation models. "Reported" follows the original paper when available; "released code" follows the public implementation; "our pipeline" uses the same preprocessing, split cache, seed control, and fine-tuning budget as PRiSE-EEG. Balanced Accuracy is reported.

Model	TUEV	SEED-V	HMC	Workload
CodeBrain reported	64.28±0.62	41.37±0.23	–	75.14±2.03
CodeBrain by released code	63.19±1.06	41.24±0.39	71.93±0.54	75.25±0.33
CodeBrain by our pipeline	64.43±1.22	41.54±0.45	72.06±0.48	78.73±0.32
TFM-Tokenizer reported	49.43±5.16	–	–	–
TFM-Tokenizer by released code	51.12±0.60	35.50±0.18	66.74±0.54	58.47±0.59
TFM-Tokenizer by our pipeline	53.65±0.55	36.31±0.28	66.10±0.43	60.35±0.52
PRiSE-EEG	**75.25**±0.48	**47.34**±0.23	**74.30**±0.24	**82.44**±0.28

F VISUALIZATION

F.1 INTERPRETABLE DECISION LOCALIZATION

We apply Grad-CAM (Selvaraju et al., 2017) to the SEED, ADFTD, Workload, and Things-EEG-2 datasets to analyze how **PRiSE-EEG** extracts task-relevant spatial information from EEG signals. Specifically, Grad-CAM is a visualization technique that highlights critical regions in the input data that influence the predictions of model. By computing gradient-based weights from the target class score to the original EEG features, the resulting heatmaps provide intuitive explanations for model decisions. Although class-specific attribution patterns may vary across individuals, the averaged Grad-CAM maps summarize the scalp regions that contribute to the model's predictions. In our implementation, we generate heatmaps from the final layer of our dual-stream MLP classifier, allowing us to visualize channel-wise evidence for the model's decisions. As shown in Fig. 8, apparent lateralization is consistent with previous research on the functional networks of emotion (Liu et al., 2023). For Sad emotion, our framework pays more attention to F7, F8, and P8 electrodes, suggesting that the right frontal and temporal lobes may be dominant in negative emotion. For Neutral emotion, electrodes P2 and P6 receive relatively high attribution, indicating that the model uses parietal scalp features for this class. This spatial pattern may indicate activity in the Default Mode Network (DMN).

As shown in Fig. 9, the Grad-CAM visualizations for ADFTD show different scalp attribution patterns for Alzheimer's Disease (AD) and Frontotemporal Dementia (FTD). For the classification of AD, PRiSE-EEG assigns the highest importance to electrodes situated over the temporal and parietal lobes. When classifying FTD, attribution is more prominent at frontal electrodes. This spatial distinction is broadly consistent with reported regional differences between AD and FTD (Nardone et al., 2018). AD is characterized by pathology that predominantly affects posterior cortical areas, leading to classic EEG findings such as decreased alpha-band coherence in the temporal-parietal-occipital regions and general slowing of the background rhythm. In contrast, FTD is defined by progressive degeneration of the frontal and anterior temporal lobes, which can manifest as focal slowing.

As shown in Fig. 10, the Workload visualizations highlight electrodes Fp2, F7, F8, C4, T8, P8, O2, and F4, with different spatial distributions between background and arithmetic conditions. In the non-math condition, attribution is concentrated at frontal electrodes Fp2, F7, and F8. During arithmetic, it extends to temporal and parietal sites, suggesting that the classifier draws on more distributed spatial evidence. This shift is broadly compatible with the involvement of frontal and

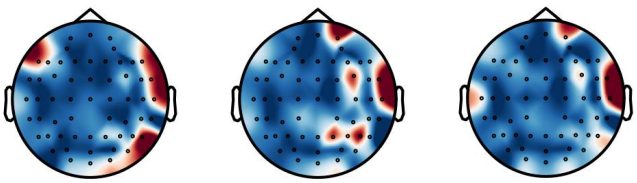


Figure 8: Grad-CAM visualization of PRiSE-EEG on SEED showing the region of model interest for Sad, Neutral and Happy emotion. Red colors indicating high relevance.

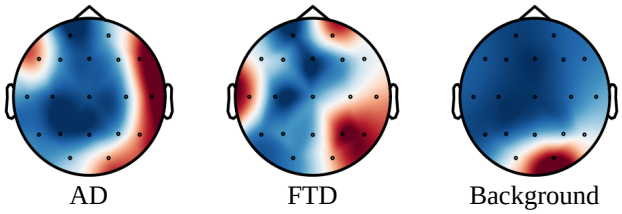
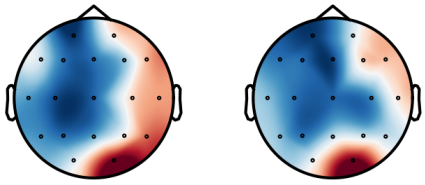


Figure 9: Grad-CAM visualization of our PRiSE-EEG on ADFTD dataset showing the region of model interest for Alzheimer's Disease, Frontotemporal Dementia and healthy control group. Red colors indicating high relevance.



(a) Arithmetic (b) Background

Figure 10: Grad-CAM visualization of PRiSE-EEG on Workload dataset showing the region of model interest for Arithmetic and Background status. Red colors indicating high relevance.

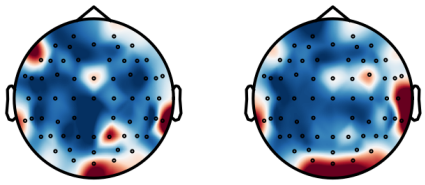
parietal regions in mathematical processing (Vansteensel et al., 2014; Artemenko, 2021). The contrast suggests that PRiSE-EEG uses condition-dependent spatial information to distinguish arithmetic workload from background recordings.

For Things-EEG-2, the Grad-CAM maps in Fig. 11 show concentrated attribution over parietal and left frontal electrodes for target stimuli, whereas this pattern is less prominent for non-target stimuli. The parietal emphasis is compatible with the scalp distribution of target-related P300 responses, particularly the P3b component (Boucher et al., 2010; Polich, 2007). P3b is a positive-going ERP component that typically peaks approximately 300-600 ms after a rare, task-relevant target and is most prominent over centro-parietal and parieto-occipital sites. It has been associated with attentional allocation, working-memory updating, and stimulus evaluation. These properties provide a possible physiological context for the target-related attribution pattern. The maps suggest that PRiSE-EEG uses spatial features relevant to target discrimination, although establishing that these features specifically reflect P3b would require time-resolved waveform analysis.

F.2 Latent Feature Clustering via t-SNE

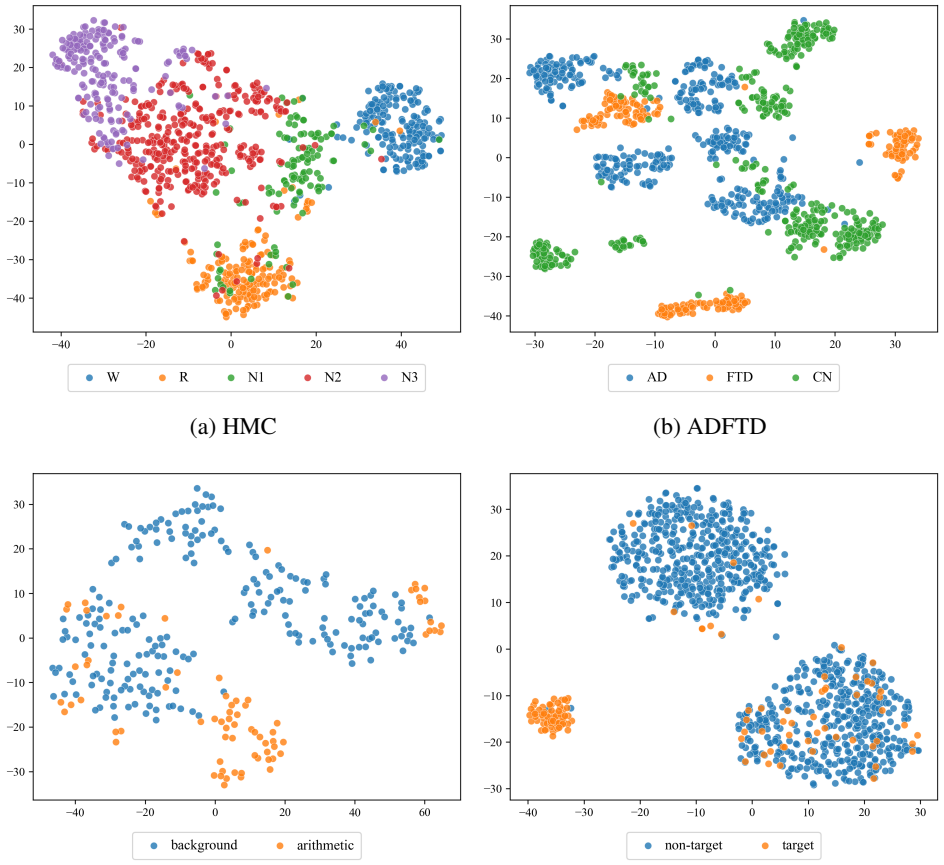
We use t-SNE (van der Maaten & Hinton, 2008), a nonlinear dimensionality reduction technique, to visualize the final classification features from four key datasets (Fig. 12). This allows us to inspect the separability of features and demonstrate the representation ability of our framework.

- For the HMC dataset, as shown in Fig. 12a, the five sleep stages form tightly grouped clusters, reflecting high intra-class consistency with physiological patterns, which also confirms the representation ability of PRiSE-EEG to disentangle discriminative features while tolerating noise.
- For the ADFTD dataset, the embeddings for AD, FTD, and healthy controls form separable clusters, corroborating the Grad-CAM findings as in Fig. 12b. These clusters are less compact than those for sleep staging, which accurately reflects the higher inter-subject variability and more subtle neurophysiological differences inherent in complex neurodegenerative diseases. Nonetheless, the model successfully carves out a feature space that captures these nuanced distinctions.
- For the Workload dataset, the two classes, "arithmetic" and "background" form two separated clusters in Fig. 12c. This confirms that the model learns a robust representation that can



(a) Target (b) Non-Target

Figure 11: Grad-CAM visualization of PRiSE-EEG on Things-EEG-2 dataset showing the region of model interest for Target and Non-Target status. Red colors indicating high relevance.



(a) HMC (b) ADFTD

(c) Workload (d) Things-EEG-2

Figure 12: t-SNE visualizations of feature embeddings reduced to 40 dimensions by PCA on different datasets with PRiSE-EEG. t-SNE runs for 1000 iterations in which the perplexity is 30 and the number of samples is 100.

reliably distinguish between a task-engaged cognitive state and a resting state, mirroring the functional network separation identified in the Grad-CAM analysis.

- For Things-EEG-2, the “target and “non-target classes form distinct clusters in the t-SNE visualization, suggesting that the learned representations contain information useful for target discrimination.

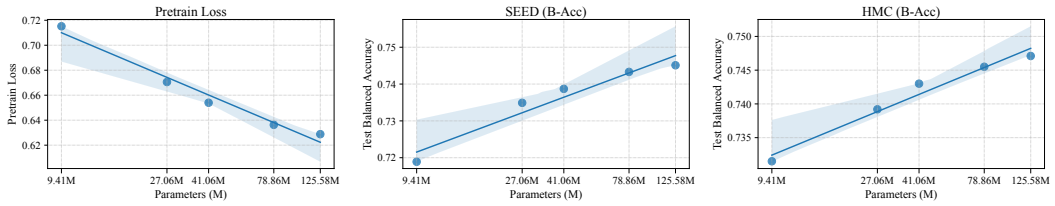


Figure 13: Scaling behavior with model size N: pretraining validation loss, SEED Balanced Accuracy, and HMC Balanced Accuracy.

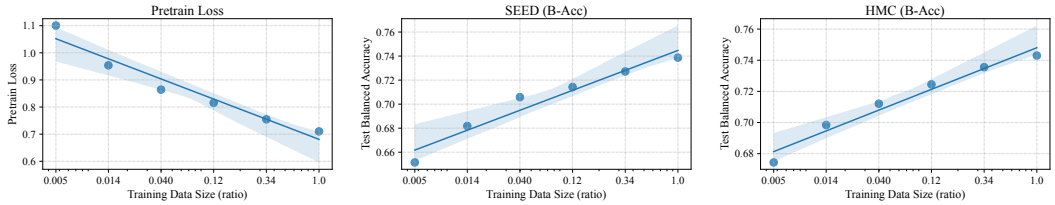


Figure 14: Scaling behavior with pretraining data size, ranging from 140 to 28,000 hours: pretraining validation loss, SEED Balanced Accuracy, and HMC Balanced Accuracy. The data ratio P denotes the fraction of the full 28,000-hour corpus.

G Scaling Behavior

In this section, we study the relationship between model performance and training scale, including the model size and pre-training data size. To study the effect of model size, we use five size variants of PRiSE-EEG (Fig. 13) and track the impact on both pre-training loss and downstream performance. To study the effect of data size, we vary the pre-training corpus from 140 to 28,000 hours and again evaluate the impact on pre-training loss and downstream performance. Pre-training loss is measured on a held-out validation set of the pre-training data, while downstream Balanced Accuracy is reported separately for HMC and SEED. The data ratio P is the number of pretraining hours divided by 28,000.

G.1 For Model Size

As shown in Fig. 13, our scaling law experiments reveal a clear and predictable relationship between model size and performance. The results on the pre-training loss function show that the scaling law of the test $\mathcal{L}_p$ with model size (N) is: $\mathcal{L}_p = -0.034 * \ln(\text{N}) + 0.786$, where R^2 is 0.974. The results on the SEED dataset show that the scaling law of the test balanced accuracy with model size (N) is: $BAcc = 0.010 * \ln(\text{N}) + 0.699$, where R^2 is 0.940. The results on the HMC dataset show that the scaling law of the test balanced accuracy with model size (N) is: $BAcc = 0.006 * \ln(\text{N}) + 0.719$, where R^2 is 0.969. The consistently high R^2 values across both pre-training and downstream tasks confirm a strong, predictable scaling behavior, which demonstrate that the PRiSE-EEG is robustly scalable and will reliably improve with increasing parameters.

G.2 For Data Size

In Fig. 14, we provide the results of the scale law experiments on the pre-training loss function ($\mathcal{L}_p$), SEED, and HMC dataset. The results on the pre-training loss function show that the scaling law of the test $\mathcal{L}_p$ with data ratio (P) is: $\mathcal{L}_p = -0.069 * \ln(\text{P}) + 0.681$, where R^2 is 0.945. The results on the SEED dataset show that the scaling law of the test balanced accuracy with data ratio (P) is: $BAcc = 0.015 * \ln(\text{P}) + 0.745$, where R^2 is 0.945. The results on the HMC dataset show that the scaling law of the test balanced accuracy with data ratio (P) is: $BAcc = 0.012 * \ln(\text{P}) + 0.748$, where R^2 is 0.964. The high R^2 values across all metrics confirm that PRiSE-EEG's performance scales predictably and effectively with the volume of pre-training data. This indicates that performance is not yet saturated and will likely continue to improve as more data becomes available.

H Computational Cost and Analysis Overhead

H.1 Capacity-matched Computational Cost

We use the PRiSE-EEG-B configuration for all cost measurements. Compared with the standard 41.06M-parameter model, the downstream instances include a measurement-specific four-class classification head, increasing the total parameter count to approximately 43.8M. We compare the depth-stratified MoE, a uniform two-shared-expert MoE, and a total-parameter-matched dense model.

Measurements use an RTX 4070 Ti SUPER, PyTorch 2.7.1, CUDA 12.8, BF16 autocast, FP32 parameters and AdamW states, eager execution and four CPU threads. Newly initialized models receive GPU-resident synthetic signals with 20 channels, 10 seconds, and 256 Hz sampling. For every instance, inference uses 10 warm-up iterations and 40 timed iterations; training uses three warm-up iterations and 15 timed iterations. CUDA synchronization is performed for each measurement, and wall-clock means are reported.

An operator counter records supported matrix multiplication, convolution, and scaled dot-product attention operations, counting a multiply-add as two operations. FFT, normalization, activations, softmax, indexing, and optimizer arithmetic are excluded. Memory is PyTorch peak allocated memory. A complete fine-tuning step includes gradient clearing, forward evaluation, loss computation, backward propagation, gradient scaling and clipping, and an AdamW update; it excludes data loading, validation, checkpoint saving, and multi-GPU communication.

Table 14: Capacity-matched inference and fine-tuning costs. FLOPs are counted at $B = 1$; memory and fine-tuning steps are measured at $B = 8$.

Instance	Total params (M)	Step GFLOPs/sample	$B = 1$ latency (ms)	$B = 8$ latency (ms)	$B = 8$ samples/s	Inference peak (MiB)	Fine-tuning step (ms)	Fine-tuning peak (MiB)
Stratified MoE	43.790	20.457	17.37	28.09	284.78	414.1	77.67	2500.4
Two-shared MoE	43.793	18.972	17.66	28.81	277.67	413.6	80.70	2419.8
Total-parameter-matched dense	43.761	32.514	5.67	18.02	443.94	460.3	52.44	2339.8

Stratified and uniform sharing have similar measured inference latencies: 17.37 versus 17.66 ms at $B = 1$, and 28.09 versus 28.81 ms at $B = 8$, corresponding to 284.78 and 277.67 samples/s. Complete fine-tuning steps take 77.67 and 80.70 ms. Although the MoE models require fewer GFLOPs than the parameter-matched dense model, their measured latency is higher. Sparse expert execution introduces token dispatch and combination, irregular memory access, and multiple small expert-specific matrix multiplications, which increase kernel-launch and data-movement overhead and reduce GPU utilization compared with the large regular operations of a dense FFN.

H.2 Representation Analysis Overhead

We measure the computational overhead of the representation analysis used to characterize layer-wise sharedness. The experiment trains a 12-layer dense Transformer for 512 steps with batch size eight, using the three self-supervised objectives adopted for pretraining.

Representations are collected every 16 training steps, resulting in 32 evaluation points. At each point, four fixed probe batches are used for each objective. The extracted representations are projected to 1024 dimensions before CKA and RSA.

Table 15: Wall-clock cost of the representation analysis. The total includes initialization, data loading, and visualization.

Stage	Time (s)
Pilot training (512 steps)	83.473
Representation extraction	21.954
CKA and RSA computation	77.394
Visualization	22.348
Total	214.341

The complete analysis takes 214.341 s with a peak memory usage of 2175.5 MiB. Pilot training and CKA/RSA computation take 83.473 s and 77.394 s, respectively, and constitute the main computational cost.

I Limitations and Future Work

First, while PRiSE-EEG scales to 125.58M parameters and 28,000 hours of data, this remains modest compared to the scale of contemporary language and vision foundation models. Furthermore, the inherent low signal-to-noise ratio of EEG data highlights a continual need for more advanced preprocessing techniques and self-supervised paradigms to learn truly robust representations. Second, while our unified fine-tuning strategy is effective, integrating new tasks still requires retraining the model. A significant open challenge for the field is developing a truly universal EEG model that can generalize to novel tasks with zero- or few-shot prompting, similar to the capabilities of large language models. Finally, our current work focuses on robust EEG representation; further research is needed to extend the model to multimodal scenarios such as Video-EEG, Image-EEG, Text-EEG, and fMRI-EEG for broader applicability.