EEGATLAS: ADAPTIVE TOPOLOGY LEARNING FOR GENERALIZABLE EEG REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Electroencephalography (EEG) supports applications from brain–computer interfaces to sleep monitoring. Generalizable representations are needed to transfer across subjects and paradigms, but informative channel relations vary with neural state. Existing encoders infer these relations from content alone or impose input-invariant spatial priors. EEGAtlas introduces a Dynamic Channel Relation Atlas that derives geometric bases from encoder-slot coordinates and selects a content-conditioned top-2 mixture for each sample and temporal patch, together with a shared relation. The resulting biases augment content attention. A Spatial–Temporal–Feature Alternating Block integrates this mechanism with temporal aggregation and token-wise feature transformation. Across five subject-independent frozen-probe tasks spanning event-related potentials, motor imagery, and sleep staging, EEGAtlas outperforms pretrained REVE, LaBraM, EEGPT, and CBraMod on Balanced Accuracy, Cohen's Kappa, and task-specific F1 or AUROC. Its Balanced Accuracy gains average 2.53 percentage points over the strongest task-wise baselines and reach 3.43 points on BCIC-2A. Dynamic relations also outperform static geometric relations in all three evaluated ablations, supporting content-conditioned spatial modeling. Code is available at `https://anonymous.4open.science/r/EEGAtlas-ICLR2027`.

1 INTRODUCTION

Electroencephalography (EEG) supports applications ranging from brain–computer interfaces to sleep monitoring, yet most systems still depend on task-specific labels and subject-specific calibration (Lotte et al., 2018; Roy et al., 2019). Generalizable EEG representation models seek to reduce this dependence by learning representations that transfer across subjects and EEG paradigms. Such transfer is difficult because subject, recording, and neural-state variability affect both local waveforms and the distributed interactions that make those waveforms informative. Electrode geometry provides a stable spatial scaffold, but the useful communication relations on that scaffold need not remain fixed across samples or time. A generalizable EEG encoder should therefore preserve geometric structure while adapting how its input slots communicate under the observed neural state.

Existing EEG representation learning spans task-specific encoders, large-scale pretrained models, and explicit modeling of electrode relations. DeepConvNet established convolutional EEG decoding (Schirrmeister et al., 2017), and EEG Conformer combines convolutional feature extraction with self-attention for supervised EEG decoding (Song et al., 2023). BENDR introduced large-scale self-supervised EEG pretraining (Kostas et al., 2021). EEGPT combines spatiotemporal representation alignment with masked waveform reconstruction to learn robust EEG features (Wang et al., 2024). LaBraM trains a vector-quantized neural tokenizer from spectral targets and predicts masked neural codes across channel patches (Jiang et al., 2024). CBraMod uses a criss-cross Transformer to model spatial and temporal dependencies in parallel and learns through masked EEG reconstruction (Wang et al., 2025). REVE combines large-scale masked pretraining with four-dimensional positional encoding to accommodate variable signal lengths and electrode arrangements (Ouahidi et al., 2025). This representation locates EEG tokens in space and time, but does not explicitly construct a geometry-derived family of pairwise relations or select among such relations for each sample and temporal patch. RGNN represents electrodes as anatomically grounded nodes and encodes local and interhemispheric relations with a biologically initialized, learnable adjacency matrix (Zhong et al., 2022). DGCNN learns a channel adjacency from training data (Song et al., 2020).

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.

These graph models establish learnable electrode relations, but do not select sparse mixtures of coordinate-derived pairwise bases at each temporal patch or use the selected mixture as a bias to content attention. Content-only attention provides input-dependent weights without such explicit geometric bases. Sparse conditional computation adapts feature transformations (Shazeer et al., 2017; Fedus et al., 2022) but does not determine which electrode relations should be active. The unresolved problem is how to select geometry-derived relations for each EEG sample and temporal patch while retaining ordinary content attention.

We introduce EEGAtlas around this distinction between a stable electrode scaffold and input-dependent communication relations. Its primary method, Dynamic Channel Relation Atlas, derives relation bases from standard electrode coordinates assigned to encoder slots and selects a top-2 mixture from each sample and temporal patch, together with an always-active shared relation. We further design a Spatial–Temporal–Feature Alternating Block that applies relation-aware spatial communication before same-channel temporal aggregation and token-wise feature transformation. This ordered block integrates the adaptive relation mechanism into a complete EEG encoder and serves as the architectural realization of the primary method. The encoder is pretrained through I-JEPA-style latent prediction and MAE waveform reconstruction (Assran et al., 2023; He et al., 2022) under a montage-consistent mask. We evaluate frozen representations on PhysioP300, KaggleERN, BCIC-2A, BCIC-2B, and SleepEDF (Citi et al., 2010; Perrin et al., 2012; Tangermann et al., 2012; Kemp et al., 2000) against pretrained REVE, LaBraM, EEGPT, and CBraMod (Ouahidi et al., 2025; Jiang et al., 2024; Wang et al., 2024; 2025). Dynamic, Static, and Content controls test adaptive relation selection, while component ablations evaluate the alternating block. Our main contributions are summarized as follows:

1. **Primary methodological contribution.** We propose Dynamic Channel Relation Atlas, a content-conditioned geometric relation model that combines an always-active shared relation with sample- and patch-specific routing over geometry-derived relation bases.
2. **Architectural realization.** We design a Spatial–Temporal–Feature Alternating Block that realizes adaptive relation modeling by sequentially integrating relation-aware spatial communication, same-channel temporal aggregation, and token-wise feature transformation.
3. We evaluate the resulting pretrained representations on five cross-subject tasks. Dynamic, Static, and Content controls isolate adaptive relation selection, while component ablations evaluate the alternating block.

2 METHOD

2.1 MODEL ARCHITECTURE

Given an EEG window $\boldsymbol{x} \in \mathbb{R}^{B\times C\times T}$, each encoder input slot is divided into $N = T/P$ patches of P samples. Slots correspond to recorded electrodes during pretraining and in most downstream tasks. A shared convolution embeds each slot patch into D dimensions and adds a learned slot-identity embedding. Each patch also receives S summary tokens, which aggregate information across visible slots and supply patch-level representations to the pretraining heads and frozen-probe evaluation.

The encoder contains eight layers. Layers 0 and 7 use standard channel attention (Vaswani et al., 2017) followed by a dense feedforward network, while Layers 1–6 use the alternating block. The central layers operate on the available tokens without requiring every sample to expose the same slot subset. For a context patch with C_v visible slots, Relation Attention reads C_v channel tokens and S summary tokens together. GroupAttention then reorders only the channel tokens by slot identity, while the summary tokens retain their patch-wise role.

Each intermediate encoder layer applies three pre-normalized residual operators in sequence

$$\begin{aligned} \boldsymbol{X}_{\mathrm{S}} &= \boldsymbol{X} + \mathcal{R}(\mathrm{LN}_{\mathrm{S}}(\boldsymbol{X})), \\ \boldsymbol{X}_{\mathrm{T}} &= \boldsymbol{X}_{\mathrm{S}} + \mathcal{G}(\mathrm{LN}_{\mathrm{T}}(\boldsymbol{X}_{\mathrm{S}})), \\ \boldsymbol{X}_{\mathrm{out}} &= \boldsymbol{X}_{\mathrm{T}} + \mathcal{F}(\mathrm{LN}_{\mathrm{F}}(\boldsymbol{X}_{\mathrm{T}})). \end{aligned} \tag{1}$$

Here $\mathcal{R}$, $\mathcal{G}$, and $\mathcal{F}$ denote Relation Attention, GroupAttention, and Progressive MoE. They model within-patch slot interactions, same-slot temporal evolution, and token-wise feature transformation,

respectively. Summary tokens bypass GroupAttention. The order matters for the information available to each operator. Temporal attention follows the spatially updated representation of the same input slot, and PMoE then transforms the resulting spatiotemporal token. The summary tokens continue to aggregate each patch through Relation Attention and receive a feature update from PMoE, but do not form channel-wise temporal sequences. Figure 1 summarizes the model architecture.

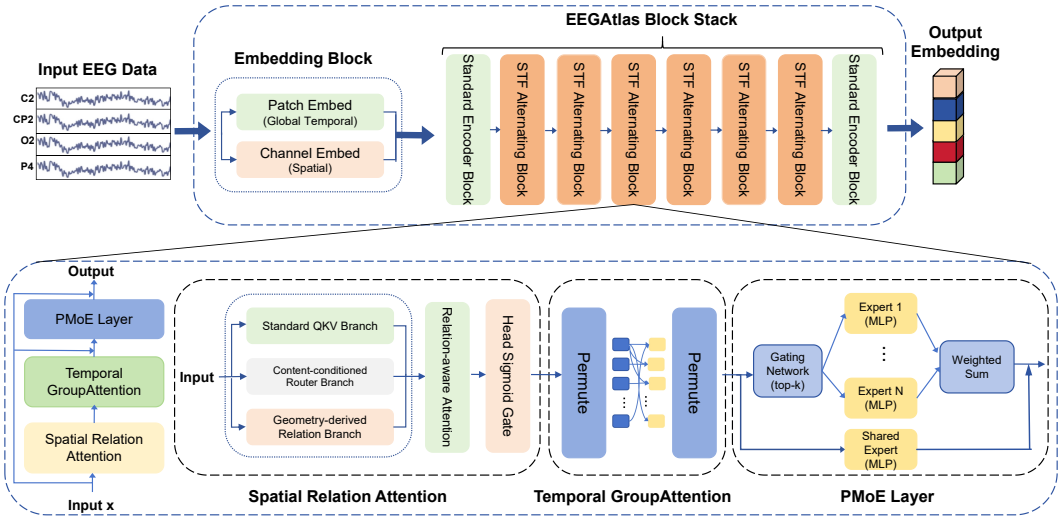

Figure 1: EEGAtlas Model Architecture

2.1.1 Dynamic Channel Relation Atlas

The Dynamic Channel Relation Atlas routes geometric relation bases using EEG content. Let $\boldsymbol{p}_i \in \mathbb{R}^3$ denote the normalized standard 10–20 coordinate assigned to encoder slot i (Oostenveld & Praamstra, 2001). For each directed slot pair, we construct

$$\boldsymbol{\phi}_{ij} = \left[\boldsymbol{p}_i - \boldsymbol{p}_j,\ \|\boldsymbol{p}_i - \boldsymbol{p}_j\|_2,\ \boldsymbol{p}_i^\top \boldsymbol{p}_j\right] \in \mathbb{R}^5. \tag{2}$$

The displacement records relative orientation, the distance records separation, and the inner product records coordinate alignment. These features are indexed by visible encoder slots. Four routed kernels and one shared kernel map $\boldsymbol{\phi}_{ij}$ to head-specific biases B^k_{ijh} and B^{sh}_{ijh}. The shared relation remains active for every sample and patch. Each kernel is a two-layer MLP with an H-dimensional output, so one geometric pair receives a separate bias in every attention head. Kernel parameters are shared across samples and patches; the input-dependent part arises from their mixture weights rather than from changing slot coordinates.

For sample b and patch n, let $\bar{\boldsymbol{x}}_{bn}$ be the mean visible-channel token. The router selects two relation bases

$$\boldsymbol{g}_{bn} = \mathrm{Top2Norm}(\mathrm{softmax}(r(\bar{\boldsymbol{x}}_{bn}))), \tag{3}$$

where r is a two-layer router and $\mathrm{Top2Norm}$ retains and renormalizes the two largest probabilities. The Dynamic Channel Relation Map is

$$M^{bn}_{ijh} = \rho \left(B^{\mathrm{sh}}_{ijh} + \sum_{k=1}^{4} g_{bnk} B^k_{ijh} \right), \tag{4}$$

where ρ is a learned scale. The router reads the current patch representation rather than a fixed slot identity. Its mean over visible slots yields one mixture per sample and patch. The two selected weights sum to one, while unselected g_{bnk} are zero. The shared relation is added before scaling and does not depend on router selection.

For query token t and head h, the sigmoid output gate is $s_{bnth} = \sigma\big(\boldsymbol{w}_h^\top \mathrm{LN}_g(\boldsymbol{x}_{bnt})\big)$, and $\boldsymbol{s}_{bn}^h$ collects these values across query tokens. Relation Attention combines the geometric map with content similarity and gates the aggregated head outputs as

$$\begin{aligned} \boldsymbol{A}_{bn}^h &= \mathrm{softmax}\left(\frac{\boldsymbol{Q}_{bn}^h (\boldsymbol{K}_{bn}^h)^\top}{\sqrt{d_h}} + \boldsymbol{M}_{bn}^h\right), \\ \mathrm{RelationAttn}(\boldsymbol{X}_{bn}) &= \mathrm{Concat}_{h=1}^{H}\left[\boldsymbol{s}_{bn}^h \odot (\boldsymbol{A}_{bn}^h \boldsymbol{V}_{bn}^h)\right] \boldsymbol{W}_O. \end{aligned} \tag{5}$$

The gate acts before head concatenation and the shared output projection, without changing the normalized attention weights or the Dynamic Relation routing probabilities. The QKV and output projections remain shared. For a partial input, the relation matrix is indexed by visible encoder-slot IDs before it is added to the corresponding query–key logits. Content similarity and the geometric mixture jointly determine within-patch attention. For each head, the selected matrix has $C_v \times C_v$ channel entries. It is padded with zero relation bias for the S summary tokens, preserving their content attention. The routed mixture varies between patches without altering the coordinate-derived bases or the shared QKV projection.

2.1.2 Channel-Conditioned GroupAttention

GroupAttention regroups spatially updated tokens by encoder slot and attends over their visible patch positions. The regrouping changes the channel-token layout from (BN_v, C_v, D) to (B, C_v, N_v, D), yielding one temporal sequence per visible slot. For a token $\widetilde{\boldsymbol{x}}_{bcn}$, FiLM (Perez et al., 2018) conditions the shared projection as $\boldsymbol{z}_{bcn} = (\widetilde{\boldsymbol{x}}_{bcn} W_{qkv}) \odot (1 + \boldsymbol{\gamma}_c) + \boldsymbol{\beta}_c$, followed by a split into queries, keys and values. The scale $\boldsymbol{\gamma}_c$ and shift $\boldsymbol{\beta}_c$ use the canonical slot ID and begin at zero.

For each encoder slot c, temporal attention is

$$\widetilde{\boldsymbol{y}}_{bc} = \mathrm{softmax}\left(\frac{\widehat{\boldsymbol{Q}}_{bc} \widehat{\boldsymbol{K}}_{bc}^\top}{\sqrt{d_h}}\right) \boldsymbol{V}_{bc}. \tag{6}$$

The hats denote RoPE-transformed queries and keys (Su et al., 2024) at their original patch positions. Thus, hidden patches do not become artificial adjacent positions in the temporal sequence. The output is projected back to width D and restored to patch-major order; summary tokens receive no GroupAttention update.

2.1.3 Progressive MoE

Progressive MoE applies token-wise feature transformations through sparse expert selection (Shazeer et al., 2017; Riquelme et al., 2021). A linear router with temperature τ produces normalized top-2 weights g_{tk}^ℓ. Its router computes softmax probabilities from $(W_r^\ell \boldsymbol{x}_t + b_r^\ell)/\tau$ and retains the two largest after renormalization. For token $\boldsymbol{x}_t$ at layer ℓ,

$$\mathrm{PMoE}_\ell(\boldsymbol{x}_t) = E_{\mathrm{sh}}^\ell(\boldsymbol{x}_t) + \sum_k g_{tk}^\ell E_k^\ell(\boldsymbol{x}_t). \tag{7}$$

The shared expert processes all channel and summary tokens, while only two routed experts are active per token. The number of routed experts increases from three in Layers 1–2 to four in Layers 3–4 and six in Layers 5–6. Routed-expert load balancing is included during pretraining. Increasing the expert count with encoder depth changes the available feature transformations without changing the number activated for each token.

2.2 Pretraining Method

As shown in Figure 2, inspired by EEGPT (Wang et al., 2024), EEGAtlas jointly trains I-JEPA-style latent prediction and MAE waveform reconstruction under a Sample-wise Montage-Consistent 2D Mask (SMC-2D Mask). The masked online encoder supplies context summaries, while a full-view momentum encoder provides stop-gradient targets (Tarvainen & Valpola, 2017). The Predictor completes the summary sequence, which the Reconstructor combines with observed summaries for waveform decoding.

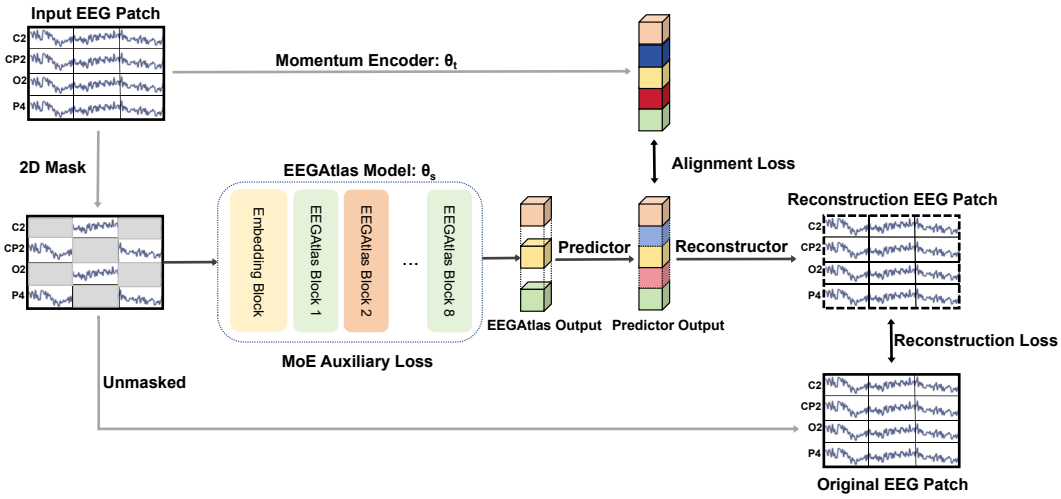

Figure 2: Pretraining of EEGAtlas

The SMC-2D Mask uses shared context patch IDs $\mathcal{P}_{\text{ctx}}$ within a batch and an independently sampled input-slot set $\mathcal{C}_b$ for each sample b. This set remains fixed across that sample's context patches. Each patch is sampled for $\mathcal{P}_{\text{ctx}}$ independently with probability p_{ctx}, resampling until context and target-only patches both occur. Every $\mathcal{C}_b$ contains C_v slots drawn uniformly without replacement. The binary context mask is

$$M_{\text{ctx}}[b, n, c] = \begin{cases} 1, & n \in \mathcal{P}_{\text{ctx}} \text{ and } c \in \mathcal{C}_b, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

The reconstruction mask M_{rec} includes all slots at target-only patches. Hidden slots at context patches form a second pool; for each sample, a fraction r_{rec} is selected uniformly without replacement, with the count rounded to the nearest integer. Reconstruction targets never overlap visible context tokens. The fixed slot set gives GroupAttention a coherent trajectory for every visible input slot, while the montage varies independently across samples. The momentum encoder still receives the complete window, so latent targets describe positions unavailable to the online encoder. The reconstruction target set also includes hidden slots inside otherwise visible patches, separating slot missingness from temporal target-only patches.

2.2.1 Momentum Encoder & Predictor

The online encoder f_θ produces S summary tokens for each context patch. The Predictor projects these summaries, inserts S learnable mask tokens at target-only positions, and restores the original patch order. Transformer layers with temporal RoPE infer the complete sequence, after which LayerNorm and a linear projection produce $\boldsymbol{z} \in \mathbb{R}^{B\times N\times S\times D}$.

The momentum encoder $f_{\bar{\theta}}$ shares the online architecture but processes the full window. Its normalized summary outputs form stop-gradient targets $\boldsymbol{h} = \text{sg}(\text{LN}(f_{\bar{\theta}}(\boldsymbol{x})))$. Latent prediction minimizes

$$\mathcal{L}_{\text{JEPA}} = \text{MSE}(\boldsymbol{z}, \boldsymbol{h}). \tag{9}$$

The loss covers context and target-only positions. It aligns refined context representations and inferred target summaries with the same full-view target.

The momentum encoder is initialized from the online encoder and updated by

$$\bar{\theta} \leftarrow m\bar{\theta} + (1-m)\theta, \tag{10}$$

where m is the EMA momentum coefficient. Gradients update the online encoder and Predictor, while the momentum encoder changes only through EMA.

2.2.2 RECONSTRUCTOR AND JOINT OBJECTIVE

The Reconstructor combines online summaries at context positions with Predictor summaries at target-only positions. For each selected slot–patch location, it adds a learnable reconstruction token conditioned by the slot embedding and temporal RoPE. Joint token processing followed by normalization and linear projection produces a P-sample waveform $\boldsymbol{r}_{bnc}$.

Let $\mathcal{R} = \{(b, n, c) : M_{\text{rec}}[b, n, c] = 1\}$. The target $\boldsymbol{y}_{bnc}$ is the input waveform patch normalized over its P samples. Reconstruction minimizes

$$\mathcal{L}_{\text{MAE}} = \text{MSE}_{\mathcal{R}}(\boldsymbol{r}, \boldsymbol{y}), \tag{11}$$

where $\text{MSE}_{\mathcal{R}}$ averages over selected locations and their P samples. Each target patch is normalized along its sample axis. Thus latent prediction supervises the complete summary sequence, while reconstruction acts only on selected slot–patch locations.

The joint objective includes routed-expert load balancing from the online encoder

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{JEPA}} + \lambda_{\text{rec}}\mathcal{L}_{\text{MAE}} + \alpha \sum_{\ell=1}^{6} (\mathcal{L}^{\ell}_{\text{PMoE}} + \beta \mathcal{L}^{\ell}_{\text{relation}}). \tag{12}$$

Here λ_{rec} weights reconstruction, α scales routing regularization, and β sets the relative Relation-routing weight. Each auxiliary loss uses $K \sum_k I_k F_k$, where I_k is the mean pre-selection probability and F_k is the selected-expert fraction, excluding shared experts. PMoE aggregates these statistics over tokens and Relation Attention over sample–patch pairs in each of the six encoder layers.

3 EXPERIMENTS

3.1 DATASETS AND PREPROCESSING

Table 1: Pretraining and downstream EEG datasets. PhysioP300 is listed at its resampled 256 Hz rate; the other rates describe the source recordings.

Dataset	Paradigm	Subjects	Sampling rate (Hz)	Labels
SEED	Emotion	15	200	3
M3CV	Multiple tasks	106	250	–
THU-Benchmark	SSVEP	35	250	40
PhysioNet-MI	MI	109	160	4
Cue-Reactivity	Cue reactivity	120	250	2
PhysioP300	P300	9	256	2
KaggleERN	ERN	26	200	2
BCIC-2A	MI	9	250	4
BCIC-2B	MI	9	250	2
SleepEDF	Sleep staging	8	100	5

As shown in Table 1, the pretraining corpus comprises SEED, M3CV, THU-Benchmark, and PhysioNet-MI (Zheng & Lu, 2015; Huang et al., 2022; Wang et al., 2017; Schalk, 2009), alongside laboratory-collected Cue-Reactivity data. The corpus spans affective EEG, multiple tasks, SSVEP, motor activity, and cue reactivity. Downstream evaluation covers P300 detection on PhysioP300, ERN detection on KaggleERN, four- and two-class motor imagery on BCIC-2A and BCIC-2B, and five-class sleep staging on SleepEDF (Citi et al., 2010; Perrin et al., 2012; Tangermann et al., 2012; Kemp et al., 2000). Pretraining inputs are standardized to 58 canonical encoder slots and 1,024 points through source-specific re-referencing, scaling, interpolation, cropping, or segment concatenation. Downstream preprocessing retains task-specific windows: two seconds for the ERP tasks, four seconds at 0–38 Hz for motor imagery, and nonoverlapping 30-second epochs low-pass filtered at 30 Hz for SleepEDF. Full dataset construction is given in Appendix A.

3.2 MODEL IMPLEMENTATION DETAILS

Every pretraining and downstream input is divided into patches of 64 samples. The four-second pretraining window contains 16 such patches; downstream patch counts depend on the task-specific

input length. The encoder, Predictor, and Reconstructor each use eight layers, width 512, eight heads, and four summary tokens per patch. The SMC-2D Mask uses $p_{\text{ctx}} = 0.5$, retains $C_v = 29$ of 58 channels, and reconstructs all target-patch channels plus $r_{\text{rec}} = 0.2$ of hidden context locations. The objective weights are $\lambda_{\text{rec}} = 2$, $\alpha = 0.0025$, and $\beta = 0.25$. Pretraining runs for 50 epochs with batch size 24, AdamW (Loshchilov & Hutter, 2019), OneCycleLR (Smith & Topin, 2019), a peak learning rate of 6×10^{-5}, and EMA momentum increasing from 0.96 toward 1.0. Training trajectories and complete settings are given in Appendices B.1 and B.

3.3 Evaluation Setup

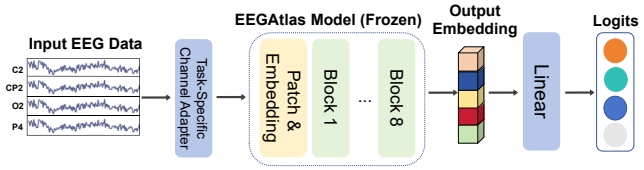

Figure 3: Frozen-probe evaluation protocol. Input adaptation is used only for the tasks specified in Appendix A.

Subject-independent transfer is evaluated with frozen probes (Varoquaux et al., 2017) as shown in Figure 3: the pretrained encoder remains fixed and only task-specific parameters are optimized. Following EEGPT (Wang et al., 2024), BCIC-2B and SleepEDF use an adaptive spatial filter to map recorded channels to encoder input slots. PhysioP300, BCIC-2A, and BCIC-2B use nine-fold leave-one-subject-out validation; KaggleERN follows the EEGPT split with ten fixed test participants and four folds over the remaining sixteen; and SleepEDF uses four subject-wise folds over eight participants. All evaluations use three fixed random seeds, with full settings in Appendix B.

We compare EEGAtlas with pretrained REVE (Ouahidi et al., 2025), LaBraM (Jiang et al., 2024), EEGPT (Wang et al., 2024), and CBraMod (Wang et al., 2025) using their documented checkpoints under the same subject partitions, probe capacity, and metric definitions. Balanced Accuracy (Brodersen et al., 2010) and Cohen's Kappa (Cohen, 1960) are reported for all tasks, with AUROC (Fawcett, 2006) for binary classification and weighted F1 (Sokolova & Lapalme, 2009) for multiclass classification.

3.4 Comparisons with Other Methods

Table 2 presents the frozen-probe evaluation results. EEGAtlas achieves the highest Balanced Accuracy, Cohen's Kappa, and task-specific third metric on all five datasets. Relative to the strongest baseline on each task, its Balanced Accuracy increases by 2.92, 2.76, 3.43, 1.85, and 1.69 percentage points on PhysioP300, KaggleERN, BCIC-2A, BCIC-2B, and SleepEDF, respectively. The mean improvement is 2.53 percentage points.

The largest gain appears on BCIC-2A, where EEGAtlas improves Balanced Accuracy by 3.43 points and weighted F1 by 3.74 points over the strongest baseline. REVE, LaBraM, and EEGPT exhibit complementary strengths across tasks, whereas EEGAtlas remains consistently strongest. Improvements on PhysioP300 and KaggleERN show consistent transfer across two ERP paradigms. EEGAtlas also leads on both motor-imagery datasets and SleepEDF, indicating that the representation transfers across ERP, motor imagery, and sleep staging. The margins in Kappa and AUROC on BCIC-2B are small, but all three metrics remain highest. On the other four tasks, the improvements in Kappa and the task-specific metric accompany the gains in Balanced Accuracy, showing that the comparison is consistent across complementary evaluation criteria.

3.5 Ablation Studies

Table 3 compares geometric structure and content-conditioned routing. Content-only removes geometric bias, Static Relation uses an input-invariant mixture of the same relation bases, and Dynamic Relation selects a top-2 mixture for each sample and temporal patch. All other pretraining and evaluation settings are matched.

Table 2: Results on Downstream Tasks

Datasets	Model	Balanced Accuracy	Cohen's Kappa	Weighted F1 / AUROC
PhysioP300	REVE	0.6495±0.0118	0.3168±0.0240	0.7147±0.0363
	LaBraM	0.6417±0.0031	0.3120±0.0072	0.7232±0.0042
	EEGPT	0.6549±0.0028	0.3092±0.0044	0.7013±0.0016
	CBraMod	0.5948±0.0047	0.2201±0.0079	0.6557±0.0046
	EEGAtlas	**0.6841±0.0032**	**0.3653±0.0069**	**0.7316±0.0028**
KaggleERN	REVE	0.5758±0.0012	0.1496±0.0021	0.6075±0.0147
	LaBraM	0.5643±0.0010	0.1303±0.0015	0.5915±0.0005
	EEGPT	0.5839±0.0013	0.1636±0.0005	0.6193±0.0013
	CBraMod	0.5212±0.0015	0.0679±0.0019	0.5421±0.0031
	EEGAtlas	**0.6115±0.0013**	**0.1874±0.0065**	**0.6445±0.0025**
BCIC-2A	REVE	0.4865±0.0058	0.3147±0.0077	0.4692±0.0017
	LaBraM	0.4579±0.0044	0.2773±0.0059	0.4494±0.0052
	EEGPT	0.4292±0.0297	0.2389±0.0396	0.4200±0.0316
	CBraMod	0.4433±0.0057	0.2577±0.0122	0.4311±0.0086
	EEGAtlas	**0.5208±0.0031**	**0.3611±0.0034**	**0.5066±0.0063**
BCIC-2B	REVE	0.7038±0.0060	0.4475±0.0120	0.8036±0.0073
	LaBraM	0.6903±0.0032	0.4606±0.0065	0.7980±0.0020
	EEGPT	0.7126±0.0030	0.4591±0.0059	0.8122±0.0057
	CBraMod	0.6809±0.0013	0.3898±0.0093	0.7635±0.0076
	EEGAtlas	**0.7311±0.0019**	**0.4623±0.0015**	**0.8137±0.0047**
SleepEDF	REVE	0.6537±0.0012	0.6196±0.0012	0.7148±0.0007
	LaBraM	0.6353±0.0087	0.6079±0.0085	0.6981±0.0024
	EEGPT	0.6489±0.0046	0.6183±0.0048	0.7142±0.0032
	CBraMod	0.6224±0.0065	0.6279±0.0069	0.7193±0.0057
	EEGAtlas	**0.6706±0.0042**	**0.6401±0.0006**	**0.7318±0.0052**

Table 3: Ablation of relation modeling. Results are Balanced Accuracy over three fixed random seeds.

Variant	Adaptive relation routing	Geometry	PhysioP300	BCIC-2A	SleepEDF
Content-only	No	No	0.6668±0.0048	0.4936±0.0067	0.6561±0.0039
Static Relation	No	Yes	0.6749±0.0041	0.5057±0.0052	0.6624±0.0035
Dynamic Relation	Yes	Yes	**0.6841±0.0032**	**0.5208±0.0031**	**0.6706±0.0042**

Static Relation exceeds Content-only by 0.81, 1.21, and 0.63 percentage points on PhysioP300, BCIC-2A, and SleepEDF. Dynamic Relation adds another 0.92, 1.51, and 0.82 points, respectively. This consistent ordering supports both the geometric prior and its content-conditioned selection, with the largest Dynamic-over-Static gain on four-class motor imagery.

Table 4 evaluates the Spatial–Temporal–Feature Alternating Block. The three variants replace Relation Attention with standard content attention, remove same-channel GroupAttention, or replace PMoE with a dense feedforward network. Each model is pretrained independently under the matched configuration.

Table 4: Component ablation of the Spatial–Temporal–Feature Alternating Block.

Variant	PhysioP300	BCIC-2A	SleepEDF
Full EEGAtlas	**0.6841±0.0032**	**0.5208±0.0031**	**0.6706±0.0042**
w/o Relation	0.6668±0.0048	0.4936±0.0067	0.6561±0.0039
w/o Group	0.6713±0.0045	0.5074±0.0056	0.6508±0.0051
w/o PMoE	0.6766±0.0039	0.5132±0.0047	0.6648±0.0038

Each substitution reduces performance. Replacing Relation Attention with standard attention has the largest mean effect at 1.97 percentage points, while removing GroupAttention causes the largest task-specific decline of 1.98 points on SleepEDF. Replacing PMoE yields smaller but consistent decreases of 0.58–0.76 points. The results support the complete alternating block as the architectural realization of Dynamic Channel Relation Atlas.

3.6 Dynamic Relation Analysis

We analyze BCIC-2A because its four motor-imagery classes provide a fine-grained test of whether relation biases vary with EEG content. Class-balanced Layer 4 biases from the frozen encoder are averaged over heads and temporal patches, centered by the class-pooled bias, and aggregated equally across held-out subjects. A matched test contrasts between-class and within-class Frobenius distances among sample-level maps and reports $100(d_{\text{between}} - d_{\text{within}})/\bar{d}$. Label permutation provides the corresponding null interval.

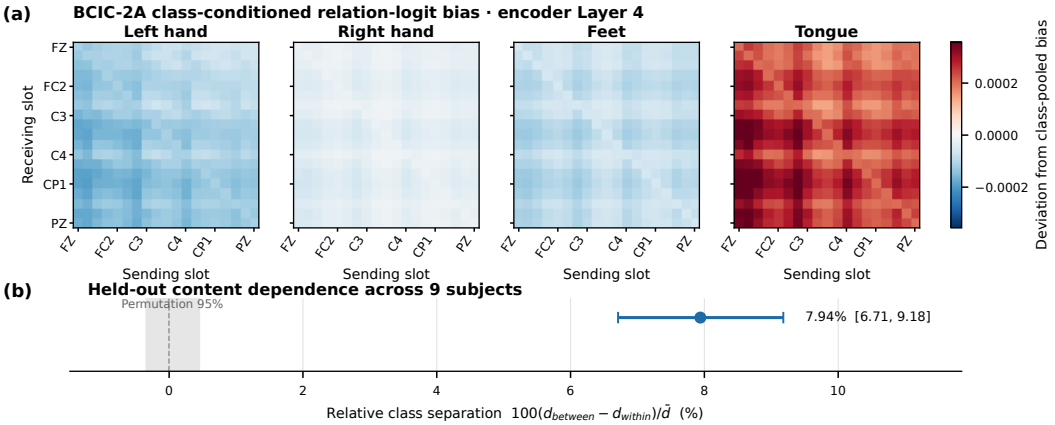

Figure 4: BCIC-2A dynamic relation analysis. (a) Class-conditioned Layer 4 relation-logit bias deviations. (b) Matched relation-bias separation across nine held-out subjects. The point and error bar denote the observed mean and its 95% bootstrap interval, while the shaded band denotes the 95% label-permutation interval.

Figure 4 shows distinct class-conditioned patterns, with tongue imagery producing the strongest positive deviation and left-hand and feet imagery showing predominantly negative deviations. The matched separation is 7.94%, with a 95% subject-level bootstrap interval of $[6.71, 9.18]\%$, entirely above the label-permutation interval of $[-0.34, 0.45]\%$. These results support content-dependent relation selection across motor-imagery classes. Task-resolved relation maps, temporal-routing results, expert-activation analyses, and the complete five-task comparison are reported in Appendix C.

4 Conclusion

EEGAtlas centers on Dynamic Channel Relation Atlas, which conditions geometry-derived encoder-slot relations on each sample and temporal patch. Its Spatial–Temporal–Feature Alternating Block integrates this mechanism with temporal aggregation and conditional feature transformation. Across five subject-independent frozen-probe tasks, EEGAtlas achieves the highest results and improves mean Balanced Accuracy by 2.53 percentage points over the strongest task-wise baselines. The Dynamic $>$ Static $>$ Content ordering and component ablations support content-dependent relation selection on a stable geometric scaffold and the complete alternating block.

Future work will investigate how the learned relation atlas evolves across layers, temporal scales, and neural states, and whether these dynamics support robust cross-task transfer in longer continuous recordings.

AI USE STATEMENT

Generative AI tools were primarily used to improve the clarity, concision, grammar, and academic style of the manuscript, and to check the consistency of terminology, notation, tables, figures, and cross-references. All AI-assisted revisions were reviewed and verified by the authors. The authors take full responsibility for the final content of this work, including text and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All datasets used in this study, except Cue-Reactivity, are publicly available research datasets and were accessed under their respective use conditions. Cue-Reactivity comprises laboratory-collected EEG recordings from participants who provided informed consent for research use. The recordings were used only for self-supervised pretraining and are not released with this paper. We report aggregate results rather than participant-level Cue-Reactivity data.

REPRODUCIBILITY STATEMENT

The model and training objectives are specified in Section 2, and the dataset, preprocessing, evaluation, and comparison protocols are described in Section 3. Appendix A provides detailed descriptions of the pretraining and downstream datasets, while Appendix B reports the implementation settings and pretraining results. Anonymous model and downstream-training code examples are available at `https://anonymous.4open.science/r/EEGAtlas-ICLR2027`. The Cue-Reactivity recordings are not distributed, as noted in the ethics statement.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Assran_Self-Supervised_Learning_From_Images_With_a_Joint-Embedding_Predictive_Architecture_CVPR_2023_paper.html.

Kay H. Brodersen, Cheng Soon Ong, Klaas E. Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. In *Proceedings of the 20th International Conference on Pattern Recognition*, 2010. doi: 10.1109/ICPR.2010.764.

Luca Citi, Riccardo Poli, and Caterina Cinel. Documenting, modelling and exploiting p300 amplitude changes due to variable target delays in donchin's speller. *Journal of Neural Engineering*, 7 (5):056006, 2010. doi: 10.1088/1741-2560/7/5/056006.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. doi: 10.1016/j.patrec.2005.10.010.

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. URL https://www.jmlr.org/papers/v23/21-0998.html.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.html.

Gan Huang, Zhenxing Hu, Weize Chen, Shaorong Zhang, Zhen Liang, Linling Li, Li Zhang, and Zhiguo Zhang. M3cv: A multi-subject, multi-session, and multi-task database for eeg-based biometrics challenge. *NeuroImage*, 264:119666, 2022. doi: 10.1016/j.neuroimage.2022.119666.

Wei-Bang Jiang, Liming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/47393e8594c82ce8fd83adc672cf9872-Abstract-Conference.html.

Bob Kemp, A. H. Zwinderman, Bert Tuk, H. A. C. Kamphuisen, and Janine J. L. Oberyé. Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000. doi: 10.1109/10.867928.

Demetres Kostas, Stéphane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15:653659, 2021. doi: 10.3389/fnhum.2021.653659.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update. *Journal of Neural Engineering*, 15(3):031005, 2018. doi: 10.1088/1741-2552/aab2f2.

Robert Oostenveld and Peter Praamstra. The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical Neurophysiology*, 112(4):713–719, 2001. doi: 10.1016/S1388-2457(00)00527-7.

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. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0760.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11671.

Margaux Perrin, Emmanuel Maby, Sébastien Daligault, Olivier Bertrand, and Jérémie Mattout. Objective and subjective evaluation of online error correction during p300-based spelling. *Advances in Human–Computer Interaction*, 2012:578295, 2012. doi: 10.1155/2012/578295.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/48237d9f2dea8c74c2a72126cf63d933-Abstract.html.

Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H. Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: A systematic review. *Journal of Neural Engineering*, 16(5):051001, 2019. doi: 10.1088/1741-2552/ab260c.

Gerwin Schalk. Eeg motor movement/imagery dataset. *PhysioNet*, 2009. doi: 10.13026/C28G6P. Version 1.0.0.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017. doi: 10.1002/hbm.23730.

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. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/pdf?id=BkFEbNwIg.

Leslie N. Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, 2019. doi: 10.1117/12.2520589.

Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4):427–437, 2009. doi: 10.1016/j.ipm.2009.03.002.

Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3):532–541, 2020. doi: 10.1109/TAFFC.2018.2817622.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023. doi: 10.1109/TNSRE.2022.3230250.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J. Miller, Gernot R. Müller-Putz, Guido Nolte, Gert Pfurtscheller, Hubert Preissl, Gerwin Schalk, Alois Schlögl, Carmen Vidaurre, Stephan Waldert, and Benjamin Blankertz. Review of the bci competition iv. *Frontiers in Neuroscience*, 6:55, 2012. doi: 10.3389/fnins.2012.00055.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/68053af2923e00204c3ca7c6a3150cf7-Abstract.html.

Gaël Varoquaux, Pradeep Reddy Raamana, Denis A. Engemann, Andrés Hoyos-Idrobo, Yannick Schwartz, and Bertrand Thirion. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, 145:166–179, 2017. doi: 10.1016/j.neuroimage.2016.10.038.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

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 *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1239.

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 *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bbbd6d915cb90be21c1254a82d45cedd-Abstract-Conference.html.

Yijun Wang, Xiaogang Chen, Xiaorong Gao, and Shangkai Gao. A benchmark dataset for SSVEP-based brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1746–1752, 2017. doi: 10.1109/TNSRE.2016.2627556.

Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015. doi: 10.1109/TAMD.2015.2431497.

Peixiang Zhong, Di Wang, and Chunyan Miao. EEG-based emotion recognition using regularized graph neural networks. *IEEE Transactions on Affective Computing*, 13(3):1290–1301, 2022. doi: 10.1109/TAFFC.2020.2994159.

A Pretraining and Downstream Dataset Details

A.1 Pretraining Datasets

The pretraining corpus combines five sources spanning affective responses, multiple cognitive and sensory paradigms, SSVEP, motor behavior, and cue reactivity. Dataset labels are retained only as source metadata and are not used by the self-supervised objectives. The descriptions below distinguish the original recording protocols from the transformations used to construct the EEGAtlas pretraining tensors.

Emotion elicitation: SEED. SEED contains recordings from 15 participants who viewed 15 approximately four-minute film clips selected to elicit positive, neutral, and negative emotions (Zheng & Lu, 2015). Each participant completed three sessions on different days. The released preprocessed waveforms contain 62 EEG channels sampled at 200 Hz, with one of the three emotion labels assigned to each film trial. **Preprocessing.** We retained the 58 electrodes shared by the pretraining montage and applied average re-referencing over these channels. Each cached segment was converted to a 2,560-point sequence by nearest-neighbor temporal interpolation and multiplied by 10^{-3}. During pretraining, a contiguous 1,024-point window was sampled at a uniformly random valid starting position, producing the four-second model input.

Multi-paradigm EEG: M3CV. M3CV contains 64-channel EEG and one marker channel recorded at 250 Hz from 106 participants (Huang et al., 2022). Ninety-five participants completed a second session on another day. The database covers 14 signal types organized under six paradigms, including resting state, transient and steady-state sensory responses, cognitive oddball responses, motor execution, and selective attention. **Preprocessing.** The preparation pipeline uses the cached Enrollment subset. For every 1,000-point record, the 58 common electrodes were selected and average re-referenced. The sequence was multiplied by 10^{-3} and interpolated to 1,024 points, yielding one four-second pretraining window. The subject-identity metadata supplied with the database was not used for representation learning.

Steady-state visual evoked potentials: THU-Benchmark. THU-Benchmark comprises 64-channel EEG from 35 participants performing a 40-target SSVEP spelling task (Wang et al., 2017). Each participant completed six blocks, and every block presented the 40 targets once in randomized order. Each trial consisted of a 0.5-second cue, five seconds of simultaneous flickering stimulation, and a 0.5-second interval. The released signals were downsampled from 1,000 to 250 Hz. **Preprocessing.** The cached data were processed following the EEGPT protocol. All 40 SSVEP targets were retained. The signals were resampled to 256 Hz, scaled to mV, and globally average re-referenced after selecting the 58 common electrodes. The interval from 0 to 4 seconds after trial onset formed the 1,024-point pretraining input. Stimulus frequency and phase labels were not provided to the pretraining objective.

Motor imagery and motor execution: PhysioNet-MI. PhysioNet-MI consists of more than 1,500 one- and two-minute recordings from 109 participants, acquired with 64 EEG channels at 160 Hz using the BCI2000 system (Schalk, 2009). After two baseline runs, participants completed three repetitions of four task paradigms involving executed or imagined movements of the left or right fist and both fists or both feet. **Preprocessing.** All eight movement conditions were retained. Trials were extracted from 0 to 6 seconds after each annotated task onset. The 58 common electrodes were selected and average re-referenced, after which the signals were multiplied by 10^3 and interpolated to 1,536 points. During pretraining, a contiguous 1,024-point window was sampled from a uniformly random valid position within each six-second trial.

Cue responses: Cue-Reactivity. The laboratory-collected Cue-Reactivity corpus contains 60-channel EEG from 120 participants recorded at 250 Hz during a cue-reactivity experiment. Two dataset-level categories are available but are not used during self-supervised pretraining. **Preprocessing.** The preparation pipeline retained the 58 electrodes in the common montage, excluding PO5 and PO6 from the original layout. Each channel was centered by subtracting its temporal mean, and every input segment was interpolated to 256 points. Consecutive groups of four segments were concatenated along the temporal dimension to form a $58 \times 1{,}024$ tensor. Any trailing group with

fewer than four segments was discarded. Input files were shuffled with seed 42 and divided into 90% training and 10% validation partitions before tensor construction.

For SEED, M3CV, THU-Benchmark, and PhysioNet-MI, prepared samples were assigned independently to the training or validation partition with probabilities 0.9 and 0.1. The exact retained sample counts and recording durations are taken from the processed corpus manifest rather than inferred from the nominal dataset metadata.

A.2 Downstream Datasets

The downstream suite contains two event-related potential tasks, two motor-imagery tasks, and one sleep-staging task. All five datasets are public and remain separate from the labeled sources used during pretraining. The classifier and metric protocol is specified in Section 3.3.

To reduce differences caused by channel alignment and enable a more controlled comparison with EEGPT, BCIC-2B and SleepEDF adopt its adaptive spatial-filter design (Wang et al., 2024). A trainable 1×1 filter and the task head are optimized while the pretrained encoder remains frozen. This operation performs learned channel alignment rather than anatomical channel reconstruction and is not treated as an EEGAtlas contribution. The filter sizes are $3 \to 7$ for BCIC-2B and $2 \to 13$ for SleepEDF. PhysioP300, KaggleERN, and BCIC-2A use their retained EEG electrodes directly; PhysioP300 additionally learns only a channel-wise scale before the frozen encoder.

P300 detection: PhysioP300. The source database contains 20 annotated recordings from each of 10 participants, acquired with 64 EEG and four EOG channels at 2,048 Hz while the participants attended target characters in a matrix speller (Citi et al., 2010). The experiments retain nine participants with IDs 1–7, 9, and 11. Each visual event is labeled as target or non-target according to the attended character. **Preprocessing and evaluation.** The 256 Hz rate in Table 1 is the resampled rate used for evaluation. EEG events are extracted from -0.1 to 2.0 seconds relative to stimulus onset, filtered at 0–120 Hz, resampled to 256 Hz, and scaled by 10^3. Marker events are discarded. Fifty-eight EEG electrodes are retained, average re-referencing is applied, and the first 512 samples form the two-second input. Evaluation uses nine leave-one-subject-out folds.

Error-related response detection: KaggleERN. KaggleERN contains recordings from 26 participants who completed five sessions of a P300 spelling task. Feedback events indicate whether the inferred character was correct or erroneous. EEG was acquired from 56 electrodes at 600 Hz and released at 200 Hz for the challenge (Perrin et al., 2012). **Preprocessing and evaluation.** Nineteen electrodes distributed over frontal, central, parietal, and occipital regions are retained and supplied directly to the encoder with their recorded identities. For every feedback event, a two-second segment beginning 0.7 seconds before feedback is extracted without additional filtering. The 400-sample sequence is nearest-neighbor interpolated to 512 samples at model input. Participants 1, 3, 4, 5, 8, 9, 10, 15, 19, and 25 constitute the held-out set in every split. Each split trains on 12 of the remaining 16 participants, rotating the excluded groups $\{2, 6, 7, 11\}$, $\{12, 13, 14, 16\}$, $\{17, 18, 20, 21\}$, and $\{22, 23, 24, 26\}$.

Four-class motor imagery: BCIC-2A. BCIC-2A comprises EEG from nine participants performing imagined movements of the left hand, right hand, both feet, and tongue (Tangermann et al., 2012). Recordings contain 22 EEG and three EOG channels sampled at 250 Hz. Two sessions were recorded on different days, with 72 trials per class in each session. Only the 288 labeled trials in the official training session of each participant are used for nine-subject leave-one-subject-out evaluation; the official test session is not used. **Preprocessing and evaluation.** Trials are extracted from 0 to 4 seconds after the motor-imagery cue. The EOG channels are removed, the EEG is filtered at 0–38 Hz, converted to microvolts, and standardized with an exponential moving estimator. All 22 EEG channels are retained and supplied directly to the encoder with their recorded electrode identities and coordinates. The trial sequence is zero-padded to 1,024 samples to form 16 patches of 64 samples.

Two-class motor imagery: BCIC-2B. BCIC-2B contains five sessions from each of nine participants performing left- and right-hand motor imagery (Tangermann et al., 2012). The recordings provide three bipolar EEG channels and three EOG channels at 250 Hz. The experiments use only

the first three official training sessions, giving 400 trials per participant for nine-subject leave-one-subject-out evaluation. The two official test sessions are not used. **Preprocessing and evaluation.** Four-second trials are extracted from motor-imagery onset. EOG channels are removed, and the EEG is filtered at 0–38 Hz, converted to microvolts, and exponentially standardized. The retained EEG channels are C3, Cz, and C4. A $3 \rightarrow 7$ adaptive spatial filter aligns them to the encoder. Following the explicit EEGPT configuration, its output slots are C5, C3, C1, Cz, C2, C4, and C6 (Wang et al., 2024). They represent learned mixtures of the three recorded channels. The trial sequence is zero-padded to 1,024 samples to form 16 patches of 64 samples.

Sleep staging: SleepEDF. SleepEDF provides overnight polysomnography with technician-scored hypnograms. The recordings used through the Sleep Physionet interface contain two EEG derivations sampled at 100 Hz (Kemp et al., 2000). Given the available computational resources, eight participants were randomly selected for four-fold evaluation, with IDs 40, 42, 45, 46, 47, 48, 49, and 52. **Preprocessing and evaluation.** Signals are converted to microvolts and low-pass filtered at 30 Hz. Wake periods more than 30 minutes before the first sleep annotation or after the last are removed. The signals are divided into nonoverlapping 30-second epochs and standardized independently by channel. Each 100 Hz epoch is interpolated to 7,680 samples at model input, yielding 120 patches of 64 samples while retaining the 30-second epoch duration. Stages W, N1, N2, N3, and REM define the five labels, with the original stages 3 and 4 merged into N3. The retained EEG derivations are Fpz–Cz and Pz–Oz. EEGPT does not enumerate adaptive-filter output channels for this task. In our evaluation, a $2 \rightarrow 13$ filter produces the encoder slots F3, F4, C3, C4, P3, P4, Fpz, Fz, Cz, CPz, Pz, POz, and Oz. These slots are learned mixtures of the two recorded derivations. Four subject-wise folds hold out the pairs $\{40, 42\}$, $\{45, 46\}$, $\{47, 48\}$, and $\{49, 52\}$, respectively.

B Implementation Details and Pretraining Results

This appendix reports the optimization behavior of the production run and the complete numerical configuration used to reproduce the model. Architectural derivations remain in Section 2, while the tables below specify the settings that instantiate those components.

B.1 Pretraining Results

Figure B1 reports epoch-level averages over the 50-epoch pretraining run. The total validation objective decreases from 1.202 to 0.959 and closely follows the training objective, which ends at 0.953. The weighted MAE term reaches 0.848 at the final epoch, while the I-JEPA term shows a short early increase followed by a sustained decrease to 0.069. The weighted routing regularization remains within a narrow range and ends at 0.042. These trajectories show that latent prediction, waveform reconstruction, and sparse routing were optimized without a late-stage divergence.

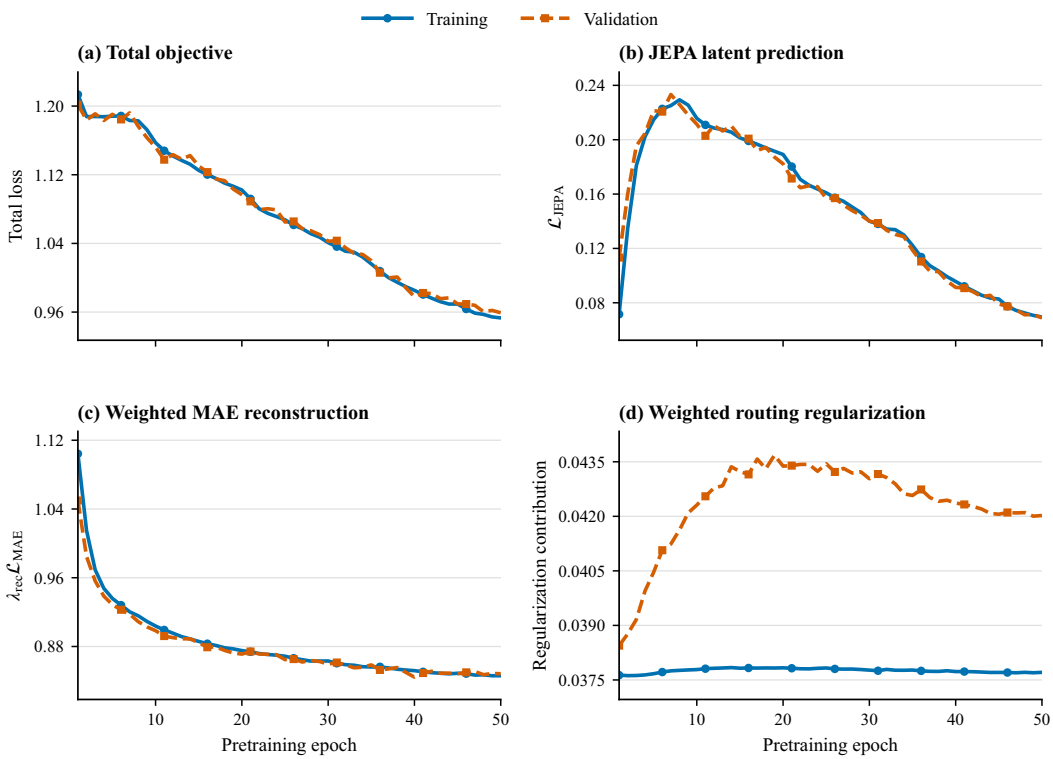

Figure B1: Pretraining loss dynamics over 50 epochs. Panels show (a) the total objective, (b) the I-JEPA latent-prediction loss, (c) the weighted MAE reconstruction loss, and (d) the weighted routing regularization. Solid and dashed curves denote training and validation values, respectively.

B.2 Complete Implementation Configuration

The five tables separate settings by their functional role. This organization keeps the encoder, adaptive relation mechanism, sparse feature transformation, pretraining procedure, and downstream evaluation protocol independently auditable.

Table B1: Input representation and encoder configuration.

Setting	Value
Pretraining montage size C	58 channels
Pretraining sampling rate / window duration	256 Hz / 4 s
Pretraining samples T / patch length P / patches N	1024 / 64 / 16
Patch embedding	1×64 Conv2d, stride 1×64
Encoder depth / model width D	8 / 512
Attention heads H / head width d_h	8 / 64
Summary tokens per temporal patch S	4
Feedforward hidden width D_{ff}	2048
Alternating / dense boundary layers	1–6 / 0 and 7
Boundary-layer feedforward activation	GELU
Query, key, and value bias	Enabled
Normalization / numerical epsilon	LayerNorm / 10^{-6}
Parameter initialization standard deviation	0.02
Projection / attention dropout	0 / 0

Table B1 defines the common patch length and model capacity used by the online encoder, momentum encoder, and downstream feature extractor. Downstream tasks use the same 64-sample patches with task-specific sequence lengths. Four summary tokens compress each temporal patch into a $4D = 2048$-dimensional representation for prediction and transfer.

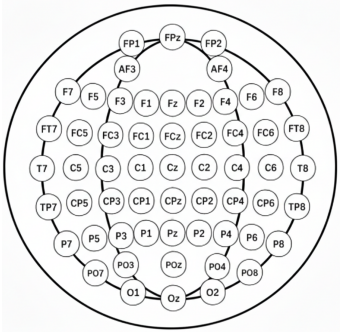

Figure B2: Standard 58-channel electrode montage used during pretraining.

Table B2: Dynamic Channel Relation Atlas and GroupAttention configuration.

Setting	Value
Enabled encoder layers	1–6
Electrode-coordinate system	Normalized standard 10–20 coordinates
Directed pair-feature dimension	5
Relation kernel dimensions	$5 \to 64 \to 8$
Routed / shared relation bases	4 / 1
Routed bases selected per sample and patch	Top-2
Relation router dimensions	$512 \to 128 \to 4$
Relation scale ρ	Learned, initialized to 0.1
Summary-token relation bias	Zero-padded rows and columns
Relation-attention output gate	Per-token, per-head sigmoid gate
GroupAttention model width / heads	512 / 8
Channel-conditioned FiLM table	$58 \times (6D)$
FiLM initialization	Zero scale and shift offsets
Temporal position encoding	RoPE with original patch indices
Summary-token GroupAttention update	Disabled

The Relation settings determine how coordinate-derived bases are selected from visible content and added to content-attention logits. GroupAttention then aggregates each channel along its original temporal positions. The scale ρ is learned jointly with the encoder rather than tuned as a fixed coefficient.

Table B3: PMoE, Predictor, and Reconstructor configuration.

Setting	Value
PMoE-enabled encoder layers	1–6
Routed-expert schedule for Layers 0–7	$[0, 3, 3, 4, 4, 6, 6, 0]$
Routed experts selected per token	Top-2
Shared expert	Always active in Layers 1–6
Expert dimensions / activation	$512 \to 2048 \to 512$ / ReLU
PMoE routing temperature τ	2.5
PMoE expert dropout	0
Predictor depth / width / heads	8 / 512 / 8
Predictor target tokens	4 learned tokens per target patch
Predictor positional encoding / output width	RoPE / 512
Reconstructor depth / width / heads	8 / 512 / 8
Reconstructor target token	One learned token per masked channel–patch location
Reconstructor positional conditioning	Channel embedding and temporal RoPE
Reconstructor output	64 waveform samples per location
Predictor and Reconstructor attention gates	Disabled

PMoE supplies the token-wise feature transformation in the alternating block while holding the activated routed-expert count fixed at two. The Predictor and Reconstructor implement the latent-prediction and waveform-reconstruction branches of the pretraining procedure.

Table B4: Masking, objectives, and pretraining optimization.

Setting	Value
Context-patch probability p_{ctx}	0.5
Visible channels per sample C_v	29 of 58
Hidden context locations reconstructed r_{rec}	0.2
Latent-prediction / reconstruction criterion	Mean squared error / mean squared error
Reconstruction weight λ_{rec}	2
Overall routing weight α	0.0025
Relative Relation-routing weight β	0.25
EMA momentum m	Linear from 0.96 toward 1.0
Training epochs / batch size	50 / 24
Optimizer / learning-rate schedule	AdamW / OneCycleLR per step
Peak learning rate / warmup fraction	6×10^{-5} / 0.2
OneCycleLR initial / final division factors	2 / 8
Weight decay	10^{-6}
Weight-decay exclusions	Biases and one-dimensional parameters
Numerical precision / matrix-multiplication precision	16-bit mixed / medium
Gradient clipping	Value clipping at 1.2
Pretraining random seed	42
Training / validation loader workers	8 / 2
Loader policy	Persistent workers; incomplete batches dropped

These settings instantiate the SMC-2D Mask and the joint objective defined in Section 2.2. The reconstruction and routing curves in Figure B1 already include their applied weights, so their plotted magnitudes correspond directly to their contributions to the total objective.

Table B5: Downstream frozen-probe configuration.

Setting	PhysioP300	KaggleERN	BCIC-2A	BCIC-2B	SleepEDF
Validation	9-fold LOSO	4 subject folds	9-fold LOSO	9-fold LOSO	4 subject folds
Probe feature projection	$2048 \to 16$	$2048 \to 16$	$2048 \to 16$	$2048 \to 16$	$2048 \to 64$
Task head	Flatten$\to 2$	Flatten$\to 2$	Flatten$\to 4$	Flatten$\to 2$	4-layer decoder$\to 5$
Head dropout	0.5	0.5	0.5	0.5	0.5
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW
Epochs / batch size	100 / 24	100 / 24	100 / 24	100 / 24	20 / 12
Peak learning rate	8×10^{-4}	8×10^{-4}	8×10^{-4}	8×10^{-4}	4×10^{-4}
Weight decay	0.03	0.02	0.01	0.03	0.02
OneCycleLR warmup fraction	0.1	0.2	0.2	0.2	0.2
Random seeds	7, 42, 718	7, 42, 718	7, 42, 718	7, 42, 718	7, 42, 718
Third metric	AUROC	AUROC	Weighted F1	AUROC	Weighted F1

For every task, the pretrained encoder remains fixed and only downstream-specific parameters are optimized with AdamW and OneCycleLR. Downstream evaluations were run on an NVIDIA GeForce RTX 4070 SUPER GPU. The fixed random seeds 7, 42, and 718 and the same subject partitions are used for EEGAtlas and all baselines.

C Expert Activation and Dynamic Relation Analysis

This appendix examines the two conditional routing mechanisms in EEGAtlas. Figure C1 first summarizes PMoE expert use across tasks and depth. Figures C2–C6 then report task-resolved Dynamic Relation results, followed by the matched five-task summary in Figure C7. All statistics are extracted from saved downstream models during inference, without changing model parameters or predictions. For the Relation analysis, classes are balanced within each held-out subject or fold before aggregation and every held-out unit receives equal weight. Layer 4 relation-logit biases

are averaged over attention heads and temporal patches, then centered by the task-level class mean. Heatmap axes identify encoder inputs. BCIC-2B and SleepEDF use learned spatial filters to produce input slots, whereas PhysioP300, KaggleERN, and BCIC-2A use retained electrodes directly. Temporal curves report class-conditional top-2 gate mass at each patch, with 95% confidence intervals computed across held-out units.

C.1 Downstream PMoE Expert Activation

Figure C1 reports normalized top-2 gate mass at Encoder Blocks 2 and 5, indexed from zero. Samples are balanced across classes within each held-out subject or fold before aggregation, and every held-out unit receives equal weight. The shared expert is omitted because it is active for every token. The displayed values therefore describe only the allocation among routed experts and sum to one within each task and layer.

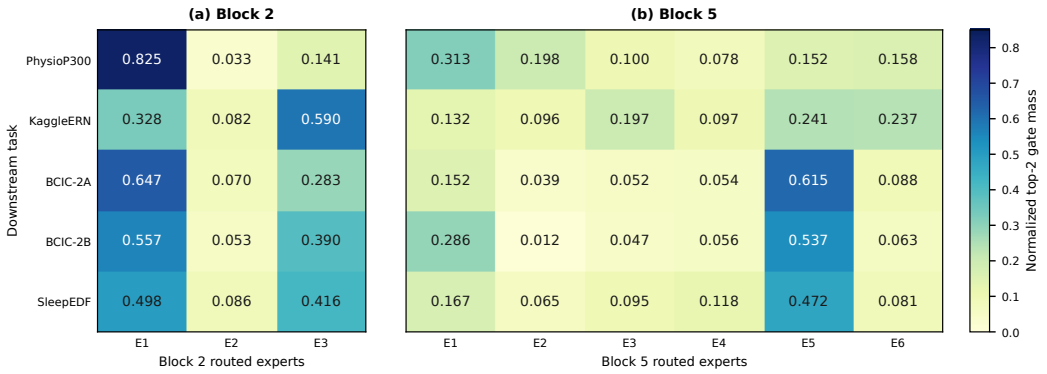

Figure C1: Downstream PMoE expert activation. Normalized top-2 gate mass is shown for the three routed experts in Encoder Block 2 and the six routed experts in Encoder Block 5. Block indices start at zero. Classes are balanced within each held-out subject or fold, after which units are averaged with equal weight. The always-active shared expert is not shown.

Block 2 exhibits a compact task-dependent allocation. Expert 1 receives most of the routed mass for PhysioP300, BCIC-2A, and BCIC-2B, whereas Expert 3 is dominant for KaggleERN. SleepEDF distributes most mass between Experts 1 and 3. Block 5 uses a broader expert set on PhysioP300 and KaggleERN, while Expert 5 receives the largest mass on BCIC-2A, BCIC-2B, and SleepEDF. Every routed expert receives nonzero aggregate mass, and the preferred expert changes across tasks and depth. These results show differentiated routing without complete collapse to one routed expert. Activation frequency alone does not, however, assign a physiological meaning to an individual expert.

C.2 Task-Resolved Dynamic Relation Analysis

PhysioP300. The P300 analysis uses nine held-out subjects and eight temporal patches per trial. Target and non-target trials produce opposing deviations around the class-pooled relation pattern. Both classes show a late decrease of Expert 1 and increasing allocation to Experts 2 and 3, indicating a shared temporal transition with smaller class-dependent differences. The matched relation separation is 3.72% with a 95% bootstrap interval of [1.44%, 5.96%], which partially overlaps the permutation interval of [-1.12%, 1.85%].

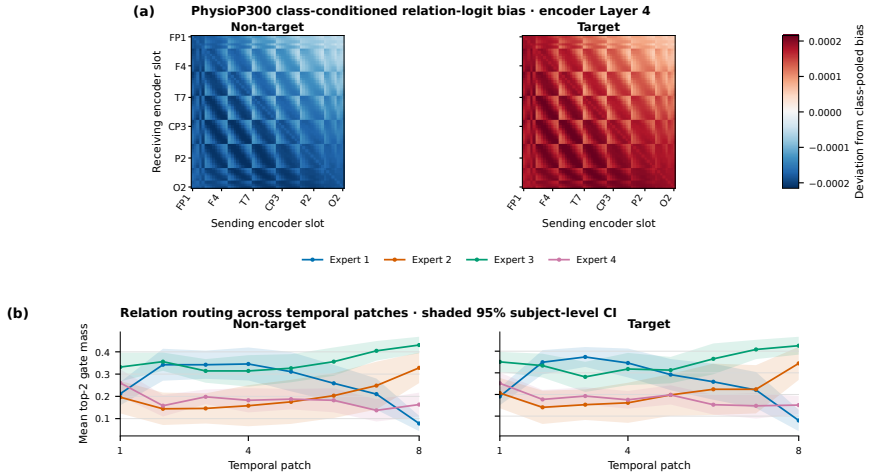

Figure C2: PhysioP300 relation bias and top-2 routing over eight temporal patches. Shaded regions denote 95% subject-level confidence intervals.

KaggleERN. KaggleERN is evaluated over four held-out folds and eight temporal patches. Correct and error feedback show opposite relation-bias deviations. Experts 3 and 4 receive most of the routing mass in both classes, while smaller temporal differences appear in the allocation to Experts 1 and 2. The matched separation is 2.63% [0.38%, 5.21%], with partial overlap with the permutation interval of [-1.63%, 3.39%].

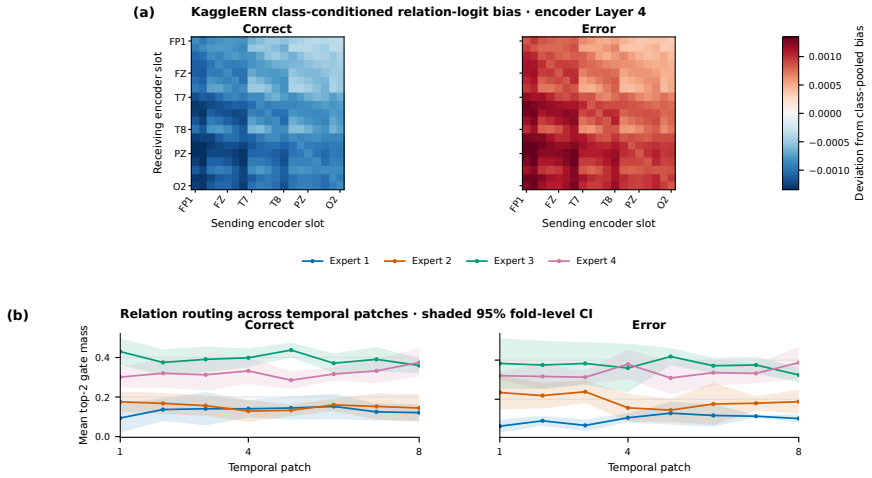

Figure C3: KaggleERN relation bias and top-2 routing over eight temporal patches. Shaded regions denote 95% fold-level confidence intervals.

BCIC-2A. The four-class motor-imagery analysis uses nine held-out subjects and 16 temporal patches. Tongue imagery produces the strongest positive deviation, while left-hand and feet imagery are predominantly negative. Expert 4 carries the largest routing mass across classes, as Expert 2 increases and Expert 3 decreases toward the end of the trial. The matched separation reaches 7.94% [6.71%, 9.18%], entirely above the permutation interval of [-0.34%, 0.45%].

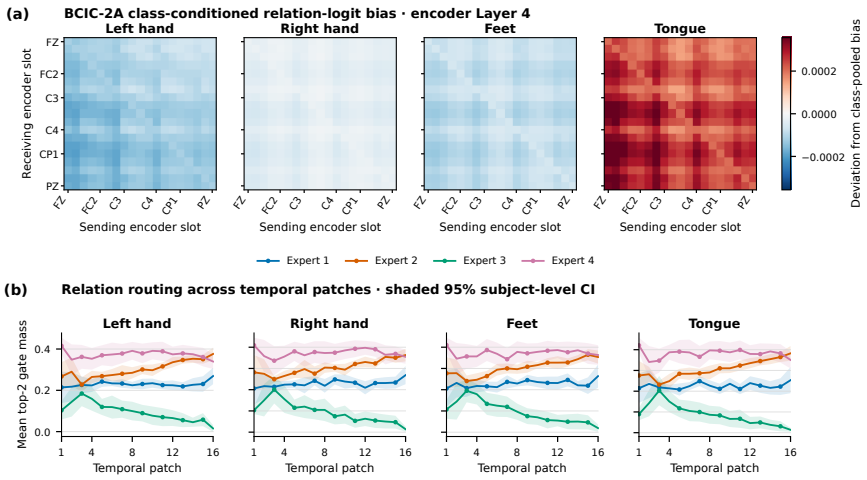

Figure C4: BCIC-2A relation bias and top-2 routing over 16 temporal patches. Shaded regions denote 95% subject-level confidence intervals.

BCIC-2B. BCIC-2B contains two motor-imagery classes evaluated over nine held-out subjects and 16 temporal patches. Left- and right-hand imagery produce opposing relation-bias deviations. Expert 4 remains dominant in both classes, while Expert 3 decreases and Experts 1 and 2 retain intermediate mass toward later patches. The matched separation is 4.56% [2.11%, 6.88%], entirely above the permutation interval of [-1.22%, 1.86%].

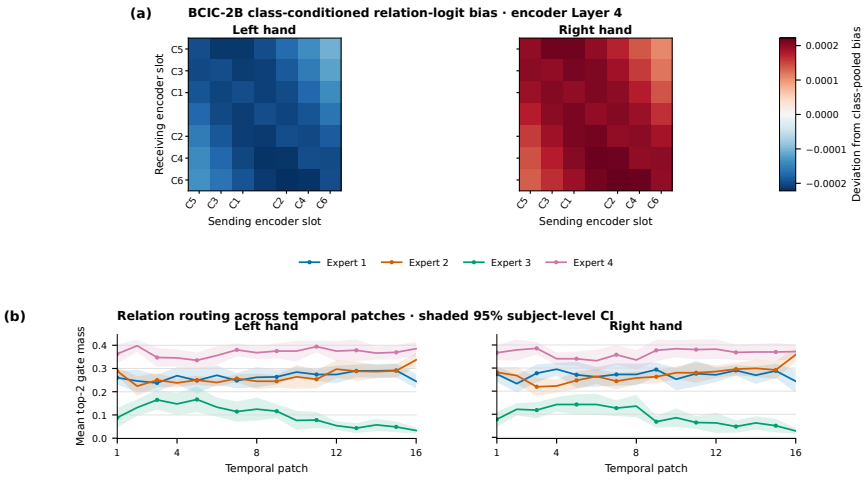

Figure C5: BCIC-2B relation bias and top-2 routing over 16 temporal patches. Shaded regions denote 95% subject-level confidence intervals.

SleepEDF. The sleep-staging analysis uses four held-out folds and 120 temporal patches per 30-second epoch. N1 and REM produce positive deviations, N3 produces a pronounced negative deviation, and Wake and N2 remain closer to the pooled pattern. Expert 2 dominates four stages, whereas N3 assigns the largest mass to Expert 4, yielding the clearest class dependence among the five tasks. The matched separation is 42.21% [30.91%, 52.73%], compared with a permutation interval of $[-1.33\%, 1.73\%]$.

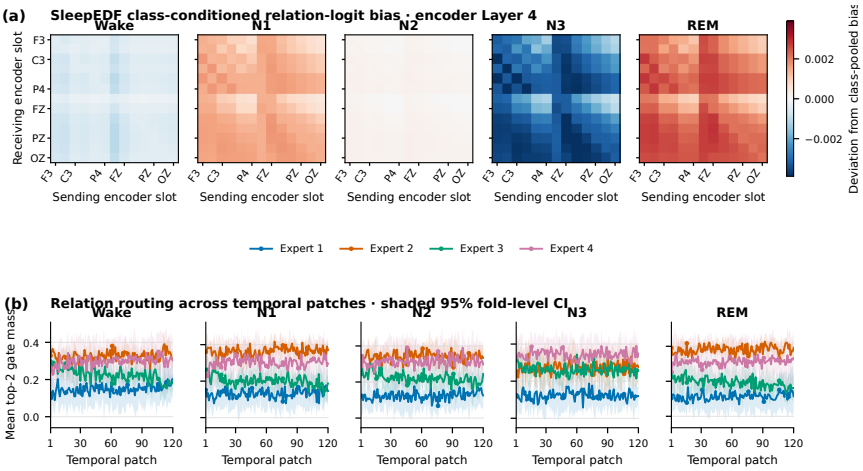

Figure C6: SleepEDF relation bias and top-2 routing over 120 temporal patches. Shaded regions denote 95% fold-level confidence intervals.

C.3 Cross-Task Summary

Figure C7 applies the matched Layer 4 Frobenius-distance statistic from Section 3.6 to all five tasks. Every task yields positive mean separation. The mean effects and 95% bootstrap intervals are 3.72% [1.44%, 5.96%] for PhysioP300, 2.63% [0.38%, 5.21%] for KaggleERN, 7.94% [6.71%, 9.18%] for BCIC-2A, 4.56% [2.11%, 6.88%] for BCIC-2B, and 42.21% [30.91%, 52.73%] for SleepEDF. The intervals for BCIC-2A, BCIC-2B, and SleepEDF lie entirely above their corresponding permutation ranges. PhysioP300 and KaggleERN show smaller positive effects with partial interval overlap. The matched analysis therefore identifies content-dependent relation structure across tasks, with an especially pronounced effect for sleep staging.

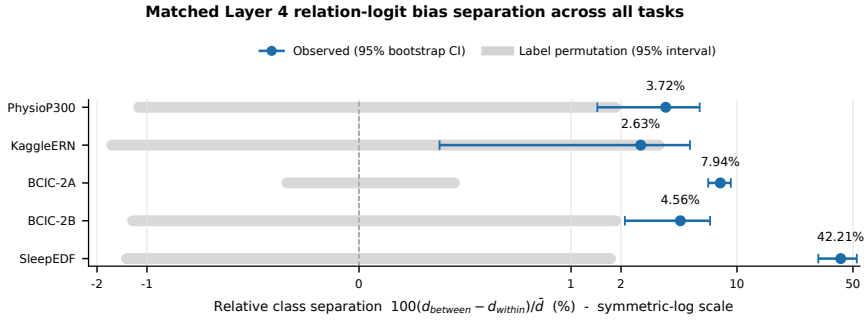

Figure C7: Matched Dynamic Relation analysis across five downstream tasks. Points and error bars show the mean relative class separation and 95% bootstrap interval of the Layer 4 relation-logit biases. Gray bars denote the corresponding 95% label-permutation intervals.

D Downstream Confusion Matrices

This section reports the final class-level error patterns for the five downstream tasks. Each matrix is normalized independently over samples from the corresponding true class, so every row sums to 100% and each diagonal entry is the recall of that class. The matrices follow the same subject-independent evaluation protocol as the main comparison.

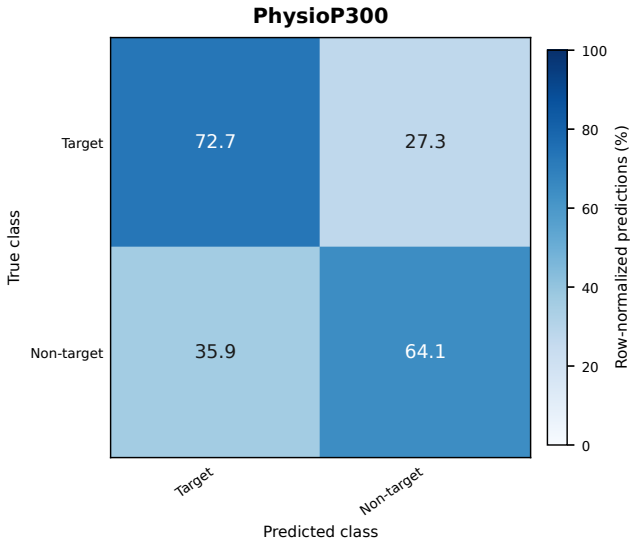

Figure D1: Row-normalized confusion matrix for PhysioP300.

PhysioP300 attains recalls of 72.7% for target trials and 64.1% for non-target trials. The higher target recall indicates that the transferred representation retains a clear stimulus-evoked response signal despite the class imbalance characteristic of P300 detection.

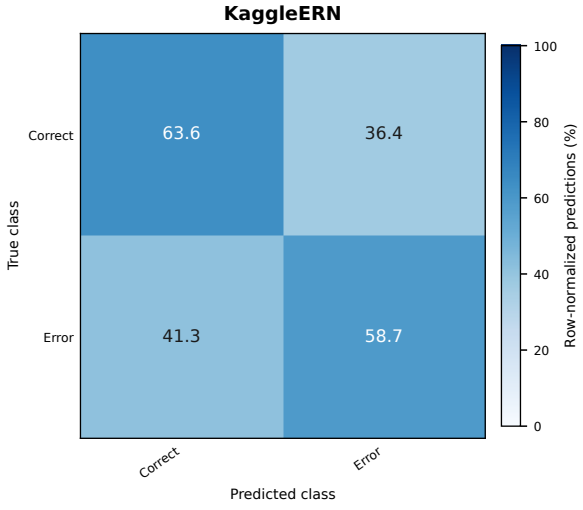

Figure D2: Row-normalized confusion matrix for KaggleERN.

KaggleERN yields recalls of 63.6% for correct-feedback trials and 58.7% for error-feedback trials. The two recalls are relatively close, showing that the reported Balanced Accuracy is not produced by recognizing only the more frequent class.

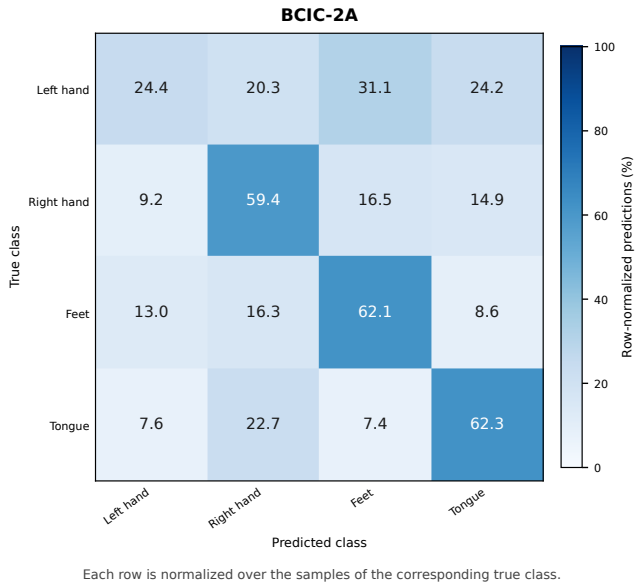

Figure D3: Row-normalized confusion matrix for four-class motor imagery on BCIC-2A.

BCIC-2A shows the largest class-wise variation. Right-hand, feet, and tongue imagery reach recalls of 59.4%, 62.1%, and 62.3%, respectively. Left-hand imagery is more difficult, with a recall of 24.4%, and is most frequently predicted as feet at 31.1%. This asymmetry explains why the weighted F1 is lower than the mean class recall and identifies left-hand imagery as the main source of residual error in this task.

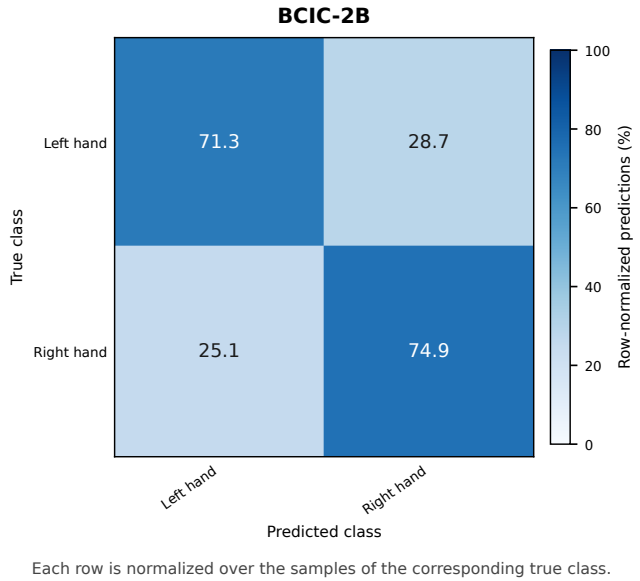

Figure D4: Row-normalized confusion matrix for two-class motor imagery on BCIC-2B.

The two BCIC-2B classes are recognized consistently, with recalls of 71.3% for left-hand imagery and 74.9% for right-hand imagery. The small recall gap indicates balanced transfer across the two motor-imagery classes.

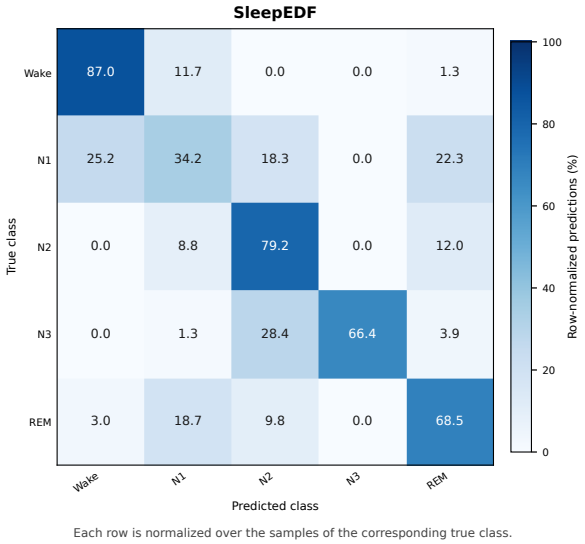

Figure D5: Row-normalized confusion matrix for five-class sleep staging on SleepEDF.

SleepEDF achieves the highest recalls for Wake and N2 at 87.0% and 79.2%. N1 remains the most difficult stage at 34.2% recall and is primarily confused with Wake, REM, and N2. N3 is mainly confused with the adjacent N2 stage, whereas REM is most frequently confused with N1. These errors are concentrated between physiologically adjacent or transitional sleep stages rather than being distributed uniformly across all labels.