BEYOND LAYOUT COMPATIBILITY: UNDERSTANDING AND REUSING SPATIAL PRIORS IN EEG FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG foundation models aim to transfer knowledge from large-scale pre-training across heterogeneous acquisition settings, where positional representations link signal content to electrode location. Yet accommodating an electrode layout does not ensure effective use of pre-trained positional knowledge. We theoretically analyze how positional parameterization shapes electrode-specific learning, motivating the reuse of independently parameterized electrode vectors. Building on this analysis, we propose RADAR, a framework centered on Spatial Positional Remapping (SPR). SPR uses pre-trained electrode vectors as spatial anchors to initialize target encodings through geometric remapping, with backbone-specific interfaces enabling cross-model reuse. RADAR further incorporates Spatio-Temporal Regularization (STR) and Distributionally Robust Optimization (DRO) to examine whether positional reuse complements data- and optimization-level improvements. Across multiple EEG decoding tasks and foundation models, RADAR improves downstream performance, while SPR outperforms alternative positional encodings in most evaluated settings. Further analyses show that the learned spatial information remains useful outside its original electrode configurations and model architecture, with SPR providing additional gains under matched auxiliary training conditions. Taken together, these results show that layout-compatible models can still benefit from better use of positional knowledge, highlighting spatial prior reuse as a way to translate large-scale pre-training into greater downstream gains.

1 INTRODUCTION

Electroencephalography (EEG) noninvasively records brain electrical activity for studying neural function, supporting clinical assessment, and enabling brain-computer interfaces (McFarland & Wolpaw, 2011; Roy et al., 2019; Aggarwal & Chugh, 2022). However, differences in electrode count, spatial arrangement, and reference scheme make models developed for one acquisition configuration difficult to reuse in another (Mellot et al., 2023). Learning representations that remain useful across these heterogeneous configurations has therefore been a longstanding challenge for reusable EEG decoding systems (Biasiucci et al., 2019; Wu et al., 2022).

Inspired by the success of foundation models in other domains (Vaswani et al., 2017; Bommasani et al., 2022; OpenAI et al., 2024), EEG foundation models use large-scale pretraining to learn reusable representations across tasks and acquisition settings (Wu et al., 2025; Xiong et al., 2025; Lu et al., 2026). These architectures can accommodate heterogeneous electrode configurations (Yang et al., 2023; Wang et al., 2024). However, architectural compatibility with diverse electrode layouts does not guarantee that pre-trained spatial priors are functionally leveraged for downstream tasks. This distinction matters because the task relevance of EEG patterns depends on both waveform content and spatial context (Schirrmeister et al., 2017). Positional representations form a core interface for learning these signal-location associations and reusing them during downstream adaptation. The question is therefore whether a model that already supports the target layout effectively uses the spatial priors learned during large-scale pretraining. Existing ablations establish that positional design affects decoding within individual models, but systematic comparisons of such reuse across layouts and receiving backbones remain limited.

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.

To investigate how pretrained spatial knowledge can be reused, we first seek a positional prior whose learned representations remain explicitly associated with their electrode locations. To guide this choice, we theoretically compare how the parameterization of electrode-indexed tables and shared coordinate-based representations shapes electrode-specific learning. Our analysis shows that independently parameterized electrode vectors allow location-specific adjustment, whereas shared coordinate generation can couple learning across locations. This distinction motivates separating electrode-specific learning during pretraining from geometry-guided sharing during transfer. We therefore use LaBraM's positional table (Jiang et al., 2024), whose electrode vectors are independently parameterized and jointly pretrained with signal representations. Each vector provides an explicitly identified spatial anchor that can be reused when constructing target positional representations. To realize this reuse beyond LaBraM, we examine whether its anchors benefit other backbones at electrode locations already represented in the source table. We further ask how their learned spatial knowledge can be extended to target locations without corresponding anchors.

Building on this analysis, we introduce RADAR, a framework centered on Spatial Positional Remapping (SPR). SPR initializes target positional representations by retaining matching LaBraM anchors and geometrically combining neighboring anchors at new locations. Our theory establishes initialization stability and bounds local extension error under spatial regularity. Backbone-specific interfaces introduce these priors into different EEG foundation models for joint fine-tuning without additional large-scale pretraining. The remapping acts only on positional vectors and leaves the measured target signals unchanged.

To assess its value in practical adaptation, we examine whether SPR remains beneficial alongside complementary data and optimization strategies. We incorporate Spatio-Temporal Regularization (STR), which uses channel perturbations and signal noise augmentation (Rommel et al., 2022), and Distributionally Robust Optimization (DRO), which emphasizes higher-loss groups (Mohajerin Esfahani & Kuhn, 2018; Sagawa et al., 2020). Comparisons under matched auxiliary training conditions test whether SPR provides additional decoding gains when regularization and robust optimization are already in place. Together, these modules form the full RADAR framework.

Within-backbone comparisons on high-density and native low-density benchmarks show that SPR outperforms alternative positional schemes in most evaluated settings. The results support both geometric extension to new electrode locations and cross-backbone reuse without interpolation, demonstrating decoding gains beyond input compatibility. SPR also provides additional gains on multiple backbones under matched auxiliary training conditions.

Our main contributions are as follows.

- **A new perspective on EEG foundation model adaptation.** We distinguish input compatibility from effective reuse of pretrained spatial knowledge, treating positional representations as transferable priors for decoding at existing and new electrode locations.
- **Spatial prior reuse across layouts and architectures.** We introduce SPR within RADAR to retain matching pretrained vectors and geometrically combine neighboring anchors for new electrode locations. Backbone-specific interfaces enable reuse across EEG foundation models without signal remapping or additional large-scale pretraining.
- **A theoretical basis for positional knowledge transfer.** We compare independent electrode vectors with shared coordinate generators to characterize electrode-specific learning and guide prior selection. We establish initialization stability for fixed geometric weights and local extension error bounds under spatial regularity.
- **Empirical evidence for transferable spatial knowledge.** Matched comparisons show gains beyond input compatibility through extension to new locations and cross-backbone reuse without interpolation. Ablations and spatial probing support the value of pretrained content and geometric correspondence beyond independent parameterization.

2 Related Work

EEG Foundation Models. Large-scale self-supervised pretraining enables EEG foundation models to learn reusable signal representations across subjects and downstream tasks (Kuruppu et al., 2026). Differences in electrode sets and configurations motivate spatial representations that associate signal features with their recording locations while accommodating heterogeneous inputs.

Spatial Representations in EEG Foundation Models. Spatial representations differ in how they encode electrode identity, spatial context, and physical geometry. LaBraM (Jiang et al., 2024) learns electrode-indexed position vectors jointly with signal representations, but requires extensions for locations absent from its table. CBraMod (Wang et al., 2025) uses asymmetric depthwise convolution over channel-by-time features for conditional positional encoding, while CSBrain (Zhou et al., 2025) combines multiscale regional convolutions with cross-region interactions. These mechanisms accommodate varying channel sets, but channel-index neighborhoods and regional grouping do not specify precise interelectrode geometry. Their physical coverage can therefore change across layouts. MMM (Yi et al., 2023) and LUNA (Döner et al., 2025) project sinusoidal coordinate features through MLPs, whereas REVE (Ouahidi et al., 2025) combines spatial and temporal coordinates using Fourier features and a learnable branch. Such encodings support new locations, but their utility depends on how learned signal-location associations generalize. LUNA also aggregates channels into a fixed-size latent space through learned-query cross-attention, potentially limiting electrode-specific detail. Across these designs, accommodating a target layout and effectively exploiting positional knowledge remain distinct considerations.

Evaluating Spatial Representations. LaBraM ablates spatial embeddings during fine-tuning, while CBraMod and REVE compare encodings and CSBrain evaluates regional and multiscale structures. These results demonstrate the utility of spatial designs, but do not directly address whether pretrained positional representations can be selected and reused to further improve decoding. In contrast, our work compares positional schemes within the same pretrained receiving backbone under matched target observations and adaptation conditions, and transfers geometrically aligned pretrained position vectors across layouts and architectures. Our focus is thus the selection and reuse of positional representations after pretraining, treating learned spatial priors as transferable components rather than evaluating them solely within their original design.

3 METHOD

3.1 SELECTING A PRETRAINED POSITIONAL PRIOR

Because electrode location informs the interpretation of EEG patterns, positional reuse should retain the signal-location associations learned during pretraining. Selecting a positional prior therefore requires considering how source representations are learned, not only whether the encoding supports new coordinates. We examine this distinction by comparing independently parameterized electrode vectors with a shared coordinate generator.

How Positional Parameterization Shapes Learning. LaBraM and REVE provide two representative ways to learn positional representations. LaBraM assigns an independently learnable vector to each source electrode, whereas REVE generates representations from spatial coordinates and temporal indices through a shared mapping that combines Fourier features with a learnable branch. An electrode-indexed table permits separate adjustment of source vectors but requires an extension rule for locations without entries. A coordinate generator provides representations at new locations directly, while coupling their learning through shared parameters.

To isolate this parameterization difference, we compare one local update in a controlled setting. Both schemes use the same signal backbone and task head, whose parameters are held fixed. The training batch, source coordinates, and temporal conditions are also identical and fixed. Let $\mathcal{R} = \{r_i\}_{i=1}^{C_s}$ be the coordinates of C_s source electrodes and let d_s be the dimension of each positional vector. Let $\mathbf{p} \in \mathbb{R}^{C_s d_s}$ collect all source position vectors, and let $\mathbf{f}_\vartheta(\mathcal{R})$ denote the complete output of the coordinate generator at these locations. We evaluate both schemes at a common current positional representation by setting the direct table equal to $\mathbf{f}_\vartheta(\mathcal{R})$. Under these conditions, both schemes share the same differentiable loss $\mathcal{L}(\mathbf{p})$ and the same positional gradient $\mathbf{g} = \nabla_{\mathbf{p}}\mathcal{L}$ at the comparison point. The mapping $\mathbf{f}_\vartheta$ includes both fixed features and learnable components, and J_ϑ denotes its Jacobian with respect to the learnable parameters ϑ. Using Euclidean gradient descent with the same step size η, direct table optimization and shared coordinate generation give

$$\Delta\mathbf{p}_{\text{table}} = -\eta\mathbf{g}, \qquad \Delta\mathbf{p}_{\text{coord}} = -\eta J_\vartheta J_\vartheta^\top \mathbf{g} + O(\eta^2), \tag{1}$$

where the second expression assumes a locally twice continuously differentiable generator and a sufficiently small step size. If J_i is the Jacobian block for electrode i, the block $J_i J_a^\top$ determines how

the positional gradient at electrode a contributes to the first-order update at electrode i. Shared generation thus introduces cross-location coupling into positional updates, whereas an independent table allows each vector to follow its own positional gradient. This distinction characterizes electrode-wise update freedom and parameter sharing, with interelectrode dependencies still learned through the signal backbone and loss. The derivation is provided in Appendix B.1.

Motivated by this comparison, we build on independently parameterized source position vectors and introduce local sharing during transfer. We retain the electrode-specific vectors learned during pretraining and use geometric proximity to construct representations for target locations without corresponding source entries. LaBraM provides a pretrained positional table that makes this design possible without additional large-scale pretraining. This choice leads to Spatial Positional Remapping (SPR), the core module of RADAR, which we describe next.

3.2 Spatial Positional Remapping

Geometry-Guided Construction. SPR uses the pretrained LaBraM electrode table $P_S \in \mathbb{R}^{C_s \times d_s}$, with row i corresponding to source location s_i. Let $\mathcal{T} = \{t_j\}_{j=1}^{C_t}$ denote the coordinates of the C_t target electrodes. We first express source and target coordinates in a common head-centered frame and normalize each coordinate to unit length. Using $s_i, t_j \in \mathbb{S}^2$ for the normalized coordinates, we measure proximity by angular distance

$$d_{ji} = d(s_i, t_j) = \arccos\big(\mathrm{clip}(t_j^\top s_i, -1, 1)\big)\,. \tag{2}$$

For each target electrode t_j, let $\mathcal{N}_K(j)$ contain its K nearest source electrodes and let i_j index the nearest source electrode. We use $K = 4$ and treat a target location as matching its nearest source anchor when $d_{ji_j} < \varepsilon$, with $\varepsilon = 10^{-6}$. SPR copies the corresponding source vector at matching locations and otherwise combines neighboring vectors using distance-dependent weights. The remapping matrix $W \in \mathbb{R}^{C_t \times C_s}$ is

$$W_{ji} = \begin{cases} \mathbf{1}\{i = i_j\}, & d_{ji_j} < \varepsilon, \\ \dfrac{\exp(-d_{ji}/\tau)}{\sum_{a \in \mathcal{N}_K(j)} \exp(-d_{ja}/\tau)}, & d_{ji_j} \geq \varepsilon,\ i \in \mathcal{N}_K(j), \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where $\tau = 0.15$ radians controls the concentration of neighborhood weights. We initialize the target positional table as $P_T^{(0)} = WP_S$, where W specifies geometric correspondence and P_S supplies pretrained vector values. Remapping acts only on positional parameters and leaves the measured target channels and raw signal values unchanged. Special-token position vectors do not participate in electrode interpolation. For a fixed target layout, the remapping is computed once before downstream fine-tuning.

We next analyze the remapped initialization to determine how SPR retains pretrained anchors and under what conditions it can extend them to new locations. We first establish preservation at matching locations and stability to source-vector perturbations, then examine the approximation error at locations without matching anchors.

Anchor Preservation and Stability. The weights satisfy $W_{ji} \geq 0$ and $\sum_i W_{ji} = 1$, so each initialized target vector lies in the convex hull of its contributing source vectors. Matching locations retain their selected source vectors exactly, and $W = I$ when the ordered source and target layouts coincide and every match is unique. Moreover, for a fixed W and any source-table perturbation ΔP_S, define the maximum row norm as $\|P\|_{\mathrm{row},\infty} = \max_i \|P[i]\|_2$. Then

$$\|W\Delta P_S\|_{\mathrm{row},\infty} \leq \|\Delta P_S\|_{\mathrm{row},\infty}. \tag{4}$$

Thus, SPR preserves available anchors and does not amplify perturbations in their vector values under this norm.

Local Extension Error. Beyond preserving existing anchors, we assess when their local combinations can approximate target positional structure at new locations. To quantify this approximation, let g_S and g_T be source and target reference positional fields taking values in $\mathbb{R}^{d_s}$ within the same

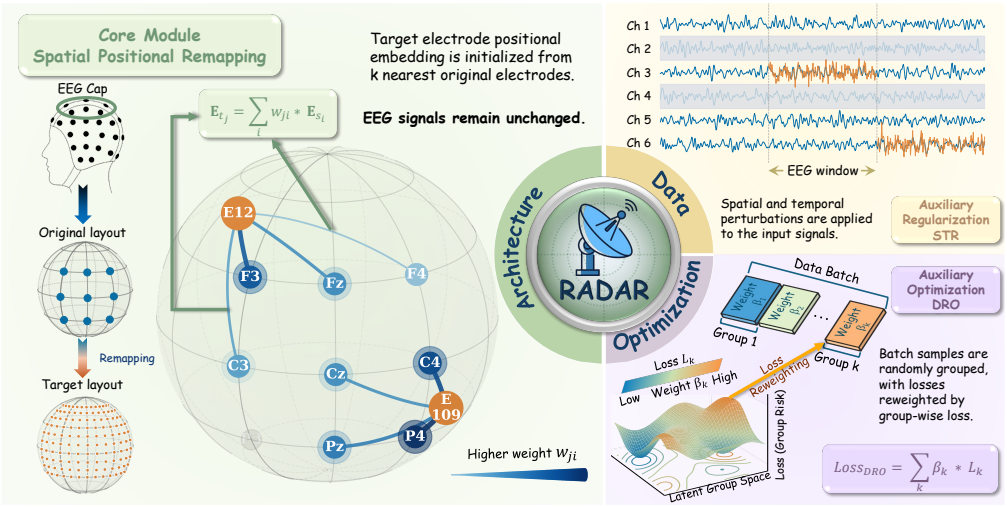

Figure 1: Overview of RADAR, a framework centered on Spatial Positional Remapping (SPR) for downstream adaptation of EEG foundation models. Auxiliary Spatio-Temporal Regularization (STR) and Distributionally Robust Optimization (DRO) complement SPR at the data and optimization levels, respectively. Electrode positions are illustrative and do not reflect actual spatial locations.

source embedding space. Write the source vectors as $P_S[i] = g_S(s_i) + e_i$, where e_i is the anchor error. For a target location t_j, assume the source reference field satisfies the local regularity condition

$$\|g_S(s_i) - g_S(t_j)\|_2 \le L d(s_i, t_j) \quad \text{for all } i \text{ with } W_{ji} > 0, \tag{5}$$

where $L \ge 0$ bounds local variation over the selected neighborhood. Define the weighted neighborhood radius $r_W(t_j) = \sum_i W_{ji} d(s_i, t_j)$ and the source-target discrepancy $\delta(t_j) = \|g_T(t_j) - g_S(t_j)\|_2$. The initialized target vector $\hat{p}_j = P_T^{(0)}[j]$ then satisfies

$$\|\hat{p}_j - g_T(t_j)\|_2 \le L r_W(t_j) + \left\| \sum_i W_{ji} e_i \right\|_2 + \delta(t_j). \tag{6}$$

The three terms capture geometric coverage, the error in the reused anchors, and the mismatch between source and target positional structure. This bound motivates local combinations by relating approximation error to the distance from supporting anchors, while making the dependence on anchor quality and source-target transferability explicit. These properties concern the remapped initialization. During downstream adaptation, the target table is fine-tuned independently of the fixed source table, together with the receiving backbone and task head. Proofs are provided in Appendices B.2 and B.3, and backbone-specific integration is described in Appendix E.8.

3.3 Auxiliary Downstream Adaptation

STR and DRO complement positional reuse at the data and optimization levels, respectively. Neither module changes the geometric construction of the target positional table.

Roles Within RADAR. We include STR and DRO for two complementary reasons. Under matched STR and DRO settings, comparisons with and without SPR test whether positional reuse provides additional decoding gains when regularization and robust optimization are already in place. As illustrated in Figure 1, these three modules form RADAR, an end-to-end adaptation framework that handles downstream fine-tuning across the three orthogonal axes of model architecture, data, and optimization, thereby better unleashing the decoding potential of foundation models obtained from large-scale pre-training.

Spatio-Temporal Regularization. Given a target EEG input X, STR forms a perturbed training observation $\tilde{X} = \mathcal{A}_\xi(X)$ through random channel masking and additive Gaussian noise, where ξ controls the stochastic augmentations. Channel masking encourages the model to use information distributed across electrodes, while noise augmentation regularizes sensitivity to waveform fluctuations. The perturbations preserve channel indexing and its correspondence with positional vectors. Both augmentations are applied during training and disabled at standard inference.

Distributionally Robust Optimization. DRO emphasizes higher-loss groups through batchwise loss reweighting. Each shuffled mini-batch of size B is randomly partitioned into G equally sized groups $\{\mathcal{B}_g\}_{g=1}^{G}$, with G dividing B. Let $\hat{y}_n$ denote the current prediction for input $\tilde{X}_n$, let y_n be its target label, and let ℓ be the task loss. The group losses, weights, and training objective are

$$L_g = \frac{1}{|\mathcal{B}_g|} \sum_{n \in \mathcal{B}_g} \ell(\hat{y}_n, y_n), \quad \beta_g = \frac{\exp(\lambda L_g)}{\sum_{a=1}^{G} \exp(\lambda L_a)}, \quad \mathcal{J} = \sum_{g=1}^{G} \beta_g L_g, \tag{7}$$

where $\lambda \geq 0$ controls the emphasis on higher-loss groups. Gradients propagate through both the group losses and the softmax weights, with the weighting rationale detailed in Appendix C.3. At $\lambda = 0$, equal group sizes make $\mathcal{J}$ identical to the mean sample loss, while increasing λ concentrates weight on groups with larger losses. We apply a linear warmup to λ during early training epochs.

4 Experiments

4.1 Experiment Setups

Datasets and Metrics. To assess SPR's ability to extend pretrained positional knowledge to unseen electrode layouts, we focus on high-density EEG datasets whose spatial configurations are not broadly represented across existing foundation-model pretraining corpora. We evaluate RADAR on four tasks from three representative datasets: MAMEM1 (256-channel, 5-class, Oikonomou et al., 2016), HGD (128-channel, 4-class, Schirrmeister et al., 2017), and CCD (128-channel, 2-class and regression, Shirazi et al., 2024). We report balanced accuracy for the three classification tasks (Brodersen et al., 2010) and normalized root mean squared error (NRMSE) for reaction time regression, with higher balanced accuracy and lower NRMSE indicating better performance.

Models and Baselines. We include three non-foundation baselines: EEGNeX (Chen et al., 2024), EEG Conformer (Song et al., 2023), and HTNet (Peterson et al., 2021). EEGNeX and EEG Conformer are task-specific models trained from scratch. HTNet serves as a baseline explicitly designed to address heterogeneous electrode configurations through projection into a common spatial representation. We evaluate five EEG foundation models: LaBraM, CBraMod, CSBrain, LUNA, and REVE. For each backbone, we compare its native positional encoding, a REVE-style coordinate-based encoding (REVE-PE), and SPR. These positional comparisons use the same pretrained receiving backbone and target observations under training settings, isolating the contribution of the positional scheme. We additionally compare SPR with signal-level remapping to 90 electrode locations represented in LaBraM's pretrained positional table, using nearest-neighbor assignment (Signal-NN) or four-neighbor interpolation (Signal-KNN4). As a further input-level comparison, we evaluate HTNet-inspired remapping that aggregates electrode signals into 64 geometry-defined virtual regions of interest (ROIs) before they enter each model.

Implementation Details. All experiments follow a cross-subject protocol, with strictly disjoint subject sets for training, validation, and testing. For each task, we use five predefined subject-level partitions and repeat each partition with three random seeds (42, 43, and 44), yielding 15 runs per method and task. We report the mean over all 15 runs and the standard deviation across the five fold-level means. All remaining experimental details are deferred to Appendix E.

4.2 Main Results

Our main experiment tests whether reusing pretrained positional knowledge improves decoding within a fixed receiving backbone and whether this benefit extends across architectures. Table 1 presents comparisons of native positional encodings, REVE-PE, and SPR within each backbone.

Table 1: Performance of different positional encoding methods across backbones. Average indicates the mean across classification tasks. Brown text reports the standard deviation across five dataset splits. Light gray rows highlight our SPR method.

Model	HGD ↑	MAMEM1 ↑	CCD-class ↑	CCD-reg ↓	Average ↑
Traditional Models					
EEGNeX	0.649 ±.057	0.294 ±.082	0.726 ±.023	**0.972** ±.013	0.556
EEGConformer	**0.689** ±.053	0.320 ±.094	**0.738** ±.008	0.976 ±.014	**0.582**
HTNet	0.607 ±.051	**0.338** ±.109	0.666 ±.071	0.980 ±.013	0.537
Foundation Models					
LaBraM	0.664 ±.062	0.447 ±.133	0.763 ±.036	0.956 ±.015	0.625
w/ REVE-PE	0.636 ±.058	0.495 ±.149	0.770 ±.028	0.960 ±.013	0.634
w/ SPR	**0.727** ±.041	**0.579** ±.209	**0.781** ±.031	**0.946** ±.013	**0.696**
CBraMod	0.667 ±.059	0.364 ±.117	0.726 ±.034	0.961 ±.014	0.586
w/ REVE-PE	0.642 ±.079	0.242 ±.085	0.630 ±.106	0.962 ±.009	0.505
w/ SPR	**0.686** ±.025	**0.377** ±.129	**0.746** ±.029	**0.956** ±.012	**0.603**
CSBrain	0.643 ±.048	0.367 ±.131	0.701 ±.028	0.963 ±.014	0.570
w/ REVE-PE	0.640 ±.054	0.328 ±.116	0.698 ±.027	0.973 ±.020	0.555
w/ SPR	**0.681** ±.072	**0.376** ±.106	**0.702** ±.034	**0.962** ±.010	**0.586**
LUNA	0.607 ±.090	**0.272** ±.056	0.625 ±.066	1.061 ±.039	0.501
w/ REVE-PE	0.625 ±.086	0.234 ±.022	0.564 ±.072	1.060 ±.040	0.474
w/ SPR	**0.626** ±.070	0.270 ±.077	**0.673** ±.040	**1.059** ±.054	**0.523**
REVE	0.657 ±.126	**0.573** ±.205	**0.748** ±.027	0.958 ±.026	**0.659**
w/ SPR	**0.665** ±.058	0.569 ±.198	0.734 ±.026	**0.950** ±.013	0.656

Benefits Across Foundation Models. SPR improves average classification performance on LaBraM, CBraMod, CSBrain, and LUNA. The largest gains occur on LaBraM, which supplies SPR's positional prior. CBraMod and CSBrain also improve on all four tasks, while LUNA shows a notable gain in HGD and CCD classification, demonstrating benefits beyond the source backbone. On REVE, SPR remains broadly comparable to the native encoding, improving HGD classification and CCD regression while yielding slightly lower scores on the other tasks. REVE's strong native baseline may reflect both its jointly pretrained positional encoding and its pretraining exposure to HGD and GSN-family layouts, as detailed in Appendix E.6. Together, these results support SPR as a reusable spatial prior across backbones with different positional designs and pretraining experience.

Comparison with Coordinate-Based Encoding. Despite REVE's strong performance, introducing REVE-PE into other pretrained backbones does not reproduce its advantage. Average classification performance drops for CBraMod, CSBrain, and LUNA relative to their native encodings. SPR instead outperforms REVE-PE on every evaluated task in these three backbones and LaBraM. This contrast suggests that pretrained electrode vectors retain transferable positional information, whereas the effectiveness of a shared coordinate mapping may depend more strongly on its integration with the receiving backbone. This interpretation aligns with the design rationale in Section 3.1, which distinguishes independent electrode-wise updates from the cross-location coupling induced by shared parameters. SPR preserves the learned anchors and introduces geometric sharing during transfer, separating positional knowledge acquisition from target-layout extension.

Beyond Input Compatibility. CBraMod and CSBrain already accommodate varying electrode configurations through their native spatial mechanisms, while LUNA explicitly generates encodings for new electrode coordinates. All three nevertheless achieve higher average classification performance with SPR in the main experiments, demonstrating gains beyond input compatibility. HTNet also explicitly addresses electrode heterogeneity, but its average classification performance is lower than that of most evaluated foundation models. We further examine positional adaptation on five native low-density benchmarks, where all evaluated backbones already support the input layouts (Appendix Table 9). Here, SPR initializes the target electrode table by directly reusing the corresponding LaBraM vectors without interpolation and coincides with native LaBraM. Across these benchmarks, SPR improves average performance over native encoding on CBraMod, CSBrain, and LUNA and outperforms REVE-PE on all four non-REVE backbones, whereas REVE favors its native encoding overall. SPR also exceeds Random PE in average balanced accuracy on all five back-

bones, supporting the value of pretrained positional content without interpolation. Together, these results show that positional adaptation can improve decoding beyond input compatibility, both when extending positional priors to new locations and when directly reusing them without interpolation.

Complementary Layout Evaluations. We further examine whether SPR's benefits persist when fewer target electrodes are available. On MAMEM1 and CCD classification reduced to 16, 32, or 64 channels, SPR improves native LaBraM in all six settings and outperforms REVE-PE overall (Appendix Table 10). We also test whether remapping EEG signals to LaBraM's existing electrode positions offers an effective alternative to adapting positional representations. Signal-NN and Signal-KNN4 achieve average balanced accuracies of 66.3% and 65.5%, respectively. SPR achieves 69.6%, outperforming both signal-level alternatives (Appendix Table 11). As a complementary input-level comparison, the HTNet-inspired ROI configurations yield lower average classification performance than the native references on all five backbones (Appendix F.6). Together, these comparisons support one of SPR's design choices to preserve the original electrode signals.

4.3 Analyzing Positional Knowledge Transfer

We use targeted comparisons to examine the roles of pretrained positional content, geometric remapping, and parameterization. Random PE randomly initializes independent target electrode vectors, whereas Unaligned Pretrained PE assigns LaBraM vectors without geometric alignment. Random-Init REVE-PE uses a REVE-style coordinate encoder with randomly initialized learnable parameters. Figure 2 summarizes the comparisons, with full results in Appendix Tables 4 and 5, respectively.

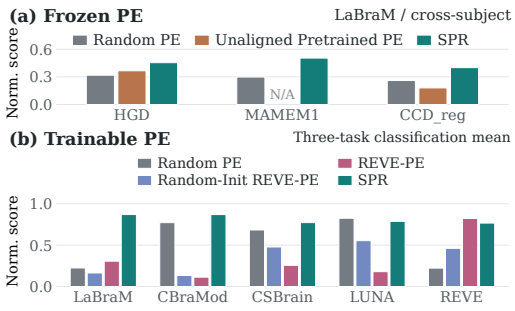

Figure 2: Cross-subject positional encoding analysis. (a) Frozen LaBraM embeddings. (b) Trainable embeddings averaged over three classification tasks. Scores are normalized per task in (a) and per backbone in (b). N/A: MAMEM1's 256 channels exceed LaBraM's 128 embeddings.

SPR Provides Useful Fixed Positional Priors. We first freeze positional embeddings while fine-tuning the LaBraM backbone and task head. On MAMEM1, SPR substantially outperforms Random PE, supporting the value of geometrically structured positional representations. On HGD and CCD, SPR consistently yields the strongest performance. Collectively, these findings show that SPR effectively transfers the spatial priors learned during large-scale pre-training to entirely new sensor layouts, thereby preserving high-fidelity spatial semantics.

Coordinate Mapping Transfer Is Backbone-Dependent. For the remaining comparisons, we allow positional parameters to update during fine-tuning across all five backbones. On CBraMod, CSBrain, and LUNA, Random-Init REVE-PE achieves higher average classification performance than pretrained REVE-PE. Under the evaluated adaptations, transferring a pretrained coordinate mapping does not consistently outperform learning a coordinate encoder during downstream fine-tuning. This suggests that effective coordinate-based positional transfer depends on integration with the receiving backbone, beyond the mapping's effectiveness in its source model.

Independent Vectors Offer Benefits Without Positional Pretraining. With both positional modules randomly initialized and the pretrained receiving backbone unchanged, Random PE achieves higher average classification performance than Random-Init REVE-PE on LaBraM, CBraMod, CSBrain, and LUNA. Independent electrode vectors thus provide an effective adaptation mechanism even without pretrained positional weights. This observation is consistent with Section 3.1, where independent vectors permit electrode-specific updates while shared parameters couple updates across locations. REVE instead favors the mapping form, as expected for a backbone designed and pretrained with coordinate-based positional encoding.

SPR's Gains Extend Beyond Independent Parameterization. We finally examine whether the advantage of independent parameterization alone accounts for SPR's gains. Both Random PE and Unaligned Pretrained PE retain independent electrode vectors, but the former omits pretrained content and the latter omits geometric correspondence. SPR outperforms both controls in most com-

parable settings, indicating that neither independent positional parameters nor unaligned pretrained vectors suffice to consistently recover SPR's benefits. The exceptions on LUNA may reflect backbone limitations that lead to its weaker decoding performance, particularly on MAMEM1, and constrain its ability to benefit from transferred positional priors. These comparisons support the value of combining pretrained vector content with target geometry, as formulated in Section 3.2, to make existing positional knowledge useful in the target layout. Additional remapping comparisons favor distance-weighted SPR over nearest-neighbor assignment and uniform averaging, supporting its local aggregation rule in Appendix F.7.

4.4 SPR UNDER AUXILIARY ADAPTATION

We examine whether SPR provides additional benefits alongside other adaptation strategies. We compare the native model, SPR alone, STR and DRO with native positional encoding, and the full RADAR framework. Figure 3 shows a brief summary, while Appendix Table 6 reports detailed results. Under matched STR and DRO settings, adding SPR improves average classification performance, with full RADAR achieving the highest classification average among the four configurations on LaBraM, CBraMod, and CSBrain. For LUNA, adding STR and DRO to SPR lowers the classification average because performance deteriorates on CCD classification, despite improvements on HGD and MAMEM1. REVE, included in the appendix, benefits more from the auxiliary strategies when retaining its native encoding, possibly because the backbone and encoding were pretrained together. Beyond assessing SPR's incremental value, RADAR integrates positional knowledge reuse and complementary training strategies into a unified adaptation framework. Its component comparisons provide empirical guidance for selecting adaptation components to suit the receiving backbone and target task. Augmentation comparisons and auxiliary hyperparameter sensitivity are provided in Appendices F.8 and F.9.

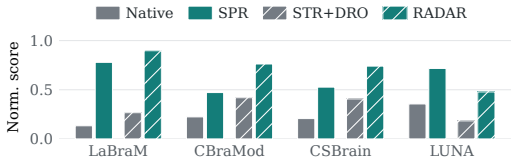

Figure 3: Cross-subject component comparisons averaged over three classification tasks. Scores are normalized per backbone, with higher values indicating better performance.

4.5 DECODING SPATIAL STRUCTURE FROM POSITIONAL EMBEDDINGS

We further examine the spatial information preserved by SPR through linear coordinate probing. With all position tables frozen, we fit and select a linear ridge probe to predict unit-sphere electrode coordinates from positional embeddings using only source electrodes, then evaluate it directly on the remapped GSN-128 and GSN-256 tables. We apply the same protocol to random source embeddings under matched remapping and to pretrained embeddings with shuffled electrode correspondences, with the detailed protocol in Appendix F.3 and full results in Table 8. SPR yields mean angular errors of 7.18° and 8.36° on the two layouts, respectively, compared with over 49° for both controls, with consistent advantages in NRMSE and R^2. These results support the preservation of linearly decodable spatial structure through SPR, reflecting the value of pretrained vectors and their electrode correspondences beyond geometric interpolation alone.

5 CONCLUSION

We presented RADAR to advance EEG foundation model adaptation beyond layout compatibility toward effective reuse of pretrained spatial knowledge. Its core module, SPR, geometrically transfers electrode-specific representations across layouts and backbones without additional large-scale pretraining. Our theoretical analysis guides prior selection and establishes stability and local approximation bounds for the remapped initialization. Matched decoding comparisons and spatial probing provide evidence for transferable spatial information in pretrained electrode vectors, with decoding gains also observed alongside complementary training strategies. Together, these findings support a broader view of positional representations as reusable knowledge rather than components confined to their original layouts and models. This perspective highlights the importance of designing and evaluating foundation models not only for input flexibility but also for their ability to preserve and exploit learned spatial structure, thereby more fully realizing the value of large-scale pretraining.

AI USE STATEMENT

In this work, we used generative AI tools for editing software code, improving the readability of the manuscript, and identifying relevant literature. We have not used generative AI tools for other tasks with required disclosure. All AI-assisted contributions were reviewed by the authors, who made the final decisions regarding code changes, manuscript wording, and relevant literature selection. We take full responsibility for the final content of this work, including all text, claims, code, and references produced or revised with the aid of generative AI.

ETHICS STATEMENT

All experiments in this work use publicly available EEG benchmark datasets. This work does not introduce new data collection, new participant recruitment, or new deployment scenarios beyond those already established in the EEG research community.

REPRODUCIBILITY STATEMENT

We describe the proposed method in Section 3, with supporting derivations and auxiliary module details in Appendices B and C, respectively. Dataset descriptions and preprocessing procedures are provided in Appendix D. Data partitions, evaluation metrics, training hyperparameters, pretrained checkpoints, and implementation details are documented in Appendix E. We provide the core code for our method in the supplementary material. Upon acceptance, we will release the code on GitHub, together with the necessary instructions for reproducing the experiments.

REFERENCES

Swati Aggarwal and Nupur Chugh. Review of machine learning techniques for EEG based brain computer interface. *Archives of Computational Methods in Engineering*, 29(5):3001–3020, 2022.

Lindsay M Alexander, Jasmine Escalera, Lei Ai, Charissa Andreotti, Karina Febre, Alexander Mangone, Natan Vega-Potler, Nicolas Langer, Alexis Alexander, Meagan Kovacs, et al. An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific data*, 4(1):1–26, 2017.

Bruno Aristimunha, Dung Truong, Pierre Guetschel, Seyed Yahya Shirazi, Isabelle Guyon, Alexandre R. Franco, Michael P. Milham, Aviv Dotan, Scott Makeig, Alexandre Gramfort, Jean-Remi King, Marie-Constance Corsi, Pedro A. Valdés-Sosa, Amit Majumdar, Alan Evans, Terrence J Sejnowski, Oren Shriki, Sylvain Chevallier, and Arnaud Delorme. EEG foundation challenge: From cross-task to cross-subject EEG decoding, 2025.

Bruno Aristimunha, Igor Carrara, Pierre Guetschel, Sara Sedlar, Pedro Rodrigues, Jan Sosulski, Divyesh Narayanan, Erik Bjareholt, Quentin Barthelemy, Robin Tibor Schirrmeister, Reinmar Kobler, Emmanuel Kalunga, Ludovic Darmet, Cattan Gregoire, Ali Abdul Hussain, Ramiro Gatti, Vladislav Goncharenko, Anton Andreev, Alberto Tates, Simon Kojima, Jordy Thielen, Davoud Hajhassani, Paul-Adrien Graignic, Katelyn Begany, Benedetto Leto, Ethan Davis, Zachary Munro, Michele Romani, Baris Talar, Emily Schrag, Bhargav Kowshik, Ralph Angelo Badang, Grace Xu, Danae Polychroniadou, Henrique Lefundes da Silva, Azra Bano, Aditya Singh, Anna Sokolova, Meysam Rahimipour, Thomas Moreau, Yannick Roy, Vinay Jayaram, Alexandre Barachant, and Sylvain Chevallier. Mother of all bci benchmarks, 2026. URL `https://github.com/NeuroTechX/moabb`.

Andrea Biasiucci, Benedetta Franceschiello, and Micah M Murray. Electroencephalography. *Current Biology*, 29(3):R80–R85, 2019.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori

Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022. URL https://arxiv.org/abs/2108.07258.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pp. 3121–3124. IEEE, 2010.

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

Xia Chen, Xiangbin Teng, Han Chen, Yafeng Pan, and Philipp Geyer. Toward reliable signals decoding for electroencephalogram: A benchmark study to EEGNeX. *Biomedical Signal Processing and Control*, 87:105475, 2024.

Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. LUNA: Efficient and topology-agnostic foundation model for EEG signal analysis. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Zijian Dong, Ruilin Li, Yilei Wu, Thuan T Nguyen, Joanna S Chong, Fang Ji, Nathanael R Tong, Christopher L Chen, and Juan H Zhou. Brain-JEPA: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. *Advances in Neural Information Processing Systems*, 37:86048–86073, 2024.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. Physiobank, physiotoolkit, and physionet. *Circulation*, 101(23):e215–e220, 2000. doi: 10.1161/01.CIR.101.23.e215.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024.

Gayal Kuruppu, Neeraj Wagh, Vaclav Kremen, and Yogatheesan Varatharajah. EEG foundation models: a critical review of current progress and future directions. *Journal of neural engineering*, 23(2):021001, 2026.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. In *International Conference on Learning Representations*, 2021.

Ziling Lu, Zongsheng Li, Xinke Shen, Kexin Lou, Yingyue Xin, Xiaoqi Chen, Shinan Wang, Xiang Chen, Jiahao Fan, Chenyu Huang, et al. OmniEEG-Bench: A standardized evaluation benchmark for EEG foundation models. *arXiv preprint arXiv:2606.00815*, 2026.

Dennis J. McFarland and Jonathan R. Wolpaw. Brain-computer interfaces for communication and control. *Commun. ACM*, 54(5):60–66, May 2011. ISSN 0001-0782. doi: 10.1145/1941487.1941506. URL https://doi.org/10.1145/1941487.1941506.

Apolline Mellot, Antoine Collas, Pedro LC Rodrigues, Denis Engemann, and Alexandre Gramfort. Harmonizing and aligning M/EEG datasets with covariance-based techniques to enhance predictive regression modeling. *Imaging Neuroscience*, 1:imag–1, 2023.

Apolline Mellot, Antoine Collas, Sylvain Chevallier, Denis Engemann, and Alexandre Gramfort. Physics-informed and unsupervised riemannian domain adaptation for machine learning on heterogeneous eeg datasets. In *2024 32nd European Signal Processing Conference (EUSIPCO)*, pp. 1367–1371. IEEE, 2024.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

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

Vangelis P. Oikonomou, Georgios Liaros, Kostantinos Georgiadis, Elisavet Chatzilari, Katerina Adam, Spiros Nikolopoulos, and Ioannis Kompatsiaris. Comparative evaluation of state-of-the-art algorithms for ssvep-based bcis, 2016.

OpenAI, Josh Achiam, Steven Adler, et al. GPT-4 technical report, 2024. URL `https://arxiv.org/abs/2303.08774`.

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 *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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

Steven M Peterson, Zoe Steine-Hanson, Nathan Davis, Rajesh PN Rao, and Bingni W Brunton. Generalized neural decoders for transfer learning across participants and recording modalities. *Journal of Neural Engineering*, 18(2):026014, 2021.

Cédric Rommel, Joseph Paillard, Thomas Moreau, and Alexandre Gramfort. Data augmentation for learning predictive models on EEG: a systematic comparison. *Journal of Neural Engineering*, 19 (6):066020, 2022.

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.

Aaqib Saeed, David Grangier, Olivier Pietquin, and Neil Zeghidour. Learning from heterogeneous EEG signals with differentiable channel reordering. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 1255–1259. IEEE, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. BCI2000: a general-purpose brain-computer interface (BCI) system. *IEEE Transactions on biomedical engineering*, 51(6):1034–1043, 2004.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human Brain Mapping*, aug 2017. ISSN 1097-0193. doi: 10.1002/hbm.23730.

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, 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. ISSN 1558-0210. doi: 10.1109/TNSRE.2022.3230250.

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

Thomas Tveitstøl, Mats Tveter, Ana S Pérez T, Christoffer Hatlestad-Hall, Anis Yazidi, Hugo L Hammer, and Ira RJ Hebold Haraldsen. Introducing region based pooling for handling a varied number of eeg channels for deep learning models. *Frontiers in Neuroinformatics*, 17:1272791, 2024.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity: festschrift for alexey chervonenkis*, pp. 11–30. Springer, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Fang Wang, Sheng-hua Zhong, Jianfeng Peng, Jianmin Jiang, and Yan Liu. Data augmentation for EEG-based emotion recognition with deep convolutional neural networks. In *International conference on multimedia modeling*, pp. 82–93. Springer, 2018.

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

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *The Thirteenth International Conference on Learning Representations*, 2025.

Dongrui Wu, Yifan Xu, and Bao-Liang Lu. Transfer learning for EEG-based brain–computer interfaces: A review of progress made since 2016. *IEEE Transactions on Cognitive and Developmental Systems*, 14(1):4–19, 2022.

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

Wei Xiong, Jiangtong Li, Jie Li, Kun Zhu, and Changjun Jiang. EEG-FM-Bench: A comprehensive benchmark for the systematic evaluation of EEG foundation models. *arXiv preprint arXiv:2508.17742*, 2025.

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

Ke Yi, Yansen Wang, Kan Ren, and Dongsheng Li. Learning topology-agnostic EEG representations with geometry-aware modeling. *Advances in Neural Information Processing Systems*, 36:53875–53891, 2023.

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. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=agcXjEHmyW.

A Additional Related Work

A.1 Spatial Alignment and Positional Representations

Methods for heterogeneous EEG configurations have long sought to establish common representations across electrode layouts. Classical spherical spline interpolation reconstructs scalp potentials at specified locations (Perrin et al., 1989). Physics-informed field interpolation can be combined with Riemannian covariance alignment to handle heterogeneous montages (Mellot et al., 2024). In deep learning, HTNet projects electrode-level spectral features onto common brain regions (Peterson et al., 2021), while Region Based Pooling aggregates channels into fixed regional representations (Tveitstøl et al., 2024). CHARM instead learns an input-dependent mapping from heterogeneous channels to a canonical order (Saeed et al., 2021). These approaches facilitate transfer by reconstructing, aggregating, or aligning observed signals and their derived representations.

A related but distinct direction incorporates spatial structure into positional encodings. In fMRI, Brain-JEPA uses functional connectivity gradients to encode relationships among brain regions (Dong et al., 2024). SPR geometrically extends and reuses electrode-specific position vectors learned jointly with EEG representations during pretraining. Rather than remapping target observations, our method adapts positional parameters to the target montage while leaving the measured EEG signals unchanged. This enables existing receiving backbones to reuse pretrained positional knowledge beyond its original layout and architecture.

B Theoretical Analysis

We provide the derivations underlying the positional parameterization comparison and the SPR initialization properties in Section 3. The analyses follow the order of the main text and use the same notation.

B.1 Local Updates Under Different Positional Parameterizations

Comparison Setting. Fix the signal backbone, task head, training batch, source coordinates $\mathcal{R}$, and temporal conditions. Any stochastic operations in the prediction and loss computation are also held fixed. Both positional schemes enter the same model computation through a common positional output $\mathbf{p} \in \mathbb{R}^{C_s d_s}$, giving a differentiable loss $\mathcal{L}(\mathbf{p})$. For the coordinate generator, let $\vartheta \in \mathbb{R}^m$ collect its m learnable parameters and let $\mathbf{f}_\vartheta(\mathcal{R})$ be its complete positional output, including any fixed feature components. At the comparison point, set the direct table to $\mathbf{p} = \mathbf{f}_\vartheta(\mathcal{R})$ and define $\mathbf{g} = \nabla_{\mathbf{p}}\mathcal{L}$ and $J_\vartheta = \partial \mathbf{f}_\vartheta(\mathcal{R})/\partial\vartheta \in \mathbb{R}^{C_s d_s \times m}$. We compare one Euclidean gradient descent step with the same learning rate $\eta > 0$.

Direct Table Update. When $\mathbf{p}$ is itself the trainable parameter, the update is exactly

$$\mathbf{p}^+ = \mathbf{p} - \eta\mathbf{g}, \qquad \Delta\mathbf{p}_{\text{table}} = \mathbf{p}^+ - \mathbf{p} = -\eta\mathbf{g}. \tag{8}$$

Shared Generator Update. For the coordinate parameterization, the chain rule gives

$$\nabla_\vartheta \mathcal{L}(\mathbf{f}_\vartheta(\mathcal{R})) = J_\vartheta^\top \mathbf{g}, \qquad \Delta\vartheta = \vartheta^+ - \vartheta = -\eta J_\vartheta^\top \mathbf{g}. \tag{9}$$

Assume that $\mathbf{f}_\vartheta(\mathcal{R})$ is twice continuously differentiable with respect to ϑ in a neighborhood of the current parameters. For a sufficiently small η such that the update segment remains in this neighborhood, Taylor expansion yields

$$\begin{aligned} \Delta\mathbf{p}_{\text{coord}} &= \mathbf{f}_{\vartheta+\Delta\vartheta}(\mathcal{R}) - \mathbf{f}_\vartheta(\mathcal{R}) \\ &= J_\vartheta \Delta\vartheta + \mathbf{r}_\vartheta \\ &= -\eta J_\vartheta J_\vartheta^\top \mathbf{g} + \mathbf{r}_\vartheta, \end{aligned} \tag{10}$$

where $\mathbf{r}_\vartheta$ is the Taylor remainder. Local twice continuous differentiability gives a finite constant M_f bounding the operator norm of the second derivative on a sufficiently small closed neighborhood, so

$$\|\mathbf{r}_\vartheta\|_2 \le \frac{M_f}{2}\|\Delta\vartheta\|_2^2 = \frac{M_f\eta^2}{2}\|J_\vartheta^\top \mathbf{g}\|_2^2. \tag{11}$$

At the fixed comparison point, J_ϑ and $\mathbf{g}$ do not depend on η, hence $\mathbf{r}_\vartheta = O(\eta^2)$ in Euclidean norm. Together with Equation 8, this establishes Equation 1.

Cross-Location Coupling. Let $J_i \in \mathbb{R}^{d_s \times m}$ be the Jacobian block for source electrode i, and let $\mathbf{g}_i \in \mathbb{R}^{d_s}$ be the corresponding block of $\mathbf{g}$. The updates for this electrode are

$$\begin{aligned} \Delta \mathbf{p}_{\text{table},i} &= -\eta \mathbf{g}_i, \\ \Delta \mathbf{p}_{\text{coord},i} &= -\eta \sum_{a=1}^{C_s} J_i J_a^\top \mathbf{g}_a + O(\eta^2). \end{aligned} \tag{12}$$

For $a \neq i$, a nonzero block $J_i J_a^\top$ allows the positional gradient at electrode a to affect the first-order update at electrode i through shared parameters. The independent table has no such generator-induced coupling, although each $\mathbf{g}_i$ can still depend on all electrodes through the common model and loss. This local comparison describes the effect of parameterization under Euclidean gradient descent, without imposing a performance ordering between the two schemes.

B.2 Anchor Preservation and Perturbation Stability

Consider the remapping matrix W in Equation 3, with fixed source and target coordinates, $1 \leq K \leq C_s$, and $\tau > 0$. For a matching target location, its row of W places unit weight on the selected source anchor and zero weight elsewhere. For an unmatched location, the exponential weights are positive on its nonempty neighborhood and are normalized by their sum, with zero weight outside that neighborhood. Thus every row satisfies

$$W_{ji} \geq 0, \qquad \sum_{i=1}^{C_s} W_{ji} = 1. \tag{13}$$

Preservation and Convexity. The initialization $P_T^{(0)} = W P_S$ gives the target vector $P_T^{(0)}[j] = \sum_i W_{ji} P_S[i]$. The nonnegative weights and unit row sum make this a convex combination of its supporting source vectors. For a matching location, the one-hot row gives $P_T^{(0)}[j] = P_S[i_j]$ exactly, including when matching is determined by the numerical tolerance ε. When the ordered source and target layouts coincide and every match is unique, $i_j = j$ for every row, so $W = I$ and $P_T^{(0)} = P_S$.

Stability to Source-Vector Perturbations. For a perturbation ΔP_S to the source table, fixing W gives the target perturbation $W(P_S + \Delta P_S) - W P_S = W \Delta P_S$. For each target electrode j, the triangle inequality and Equation 13 imply

$$\begin{aligned} \|(W \Delta P_S)[j]\|_2 &= \left\| \sum_i W_{ji} \Delta P_S[i] \right\|_2 \\ &\leq \sum_i W_{ji} \|\Delta P_S[i]\|_2 \\ &\leq \sum_i W_{ji} \|\Delta P_S\|_{\text{row},\infty} \\ &= \|\Delta P_S\|_{\text{row},\infty}. \end{aligned} \tag{14}$$

Taking the maximum over target rows establishes

$$\|W \Delta P_S\|_{\text{row},\infty} \leq \|\Delta P_S\|_{\text{row},\infty}, \tag{15}$$

which is Equation 4. The result controls perturbations in source-vector values under the maximum row norm for fixed geometry, rather than changes in the remapping weights caused by coordinate perturbations.

B.3 Local Extension Error

Fix a target electrode t_j and let $\hat{p}_j = P_T^{(0)}[j]$. Let g_S and g_T be source and target reference positional fields with values in the same source embedding space $\mathbb{R}^{d_s}$. Write $P_S[i] = g_S(s_i) + e_i$ and assume Equation 5 holds for each source electrode with $W_{ji} > 0$. Recall the weighted neighborhood radius $r_W(t_j) = \sum_i W_{ji} d(s_i, t_j)$ and the source-target discrepancy $\delta(t_j) = \|g_T(t_j) - g_S(t_j)\|_2$.

Substituting the source-vector decomposition into the remapping and using $\sum_i W_{ji} = 1$ gives the exact identity

$$\begin{aligned}\hat{p}_j - g_T(t_j) &= \sum_i W_{ji}\big(g_S(s_i) + e_i\big) - g_T(t_j) \\ &= \sum_i W_{ji}\big(g_S(s_i) - g_S(t_j)\big) \\ &\quad + \sum_i W_{ji} e_i + g_S(t_j) - g_T(t_j).\end{aligned} \tag{16}$$

Applying the triangle inequality to these three contributions yields

$$\begin{aligned}\|\hat{p}_j - g_T(t_j)\|_2 &\le \sum_i W_{ji}\|g_S(s_i) - g_S(t_j)\|_2 \\ &\quad + \left\|\sum_i W_{ji} e_i\right\|_2 + \delta(t_j) \\ &\le L\sum_i W_{ji} d(s_i, t_j) \\ &\quad + \left\|\sum_i W_{ji} e_i\right\|_2 + \delta(t_j) \\ &= L r_W(t_j) + \left\|\sum_i W_{ji} e_i\right\|_2 + \delta(t_j).\end{aligned} \tag{17}$$

The second inequality uses local regularity only at source locations carrying positive weight. This proves Equation 6 without requiring independence or zero mean of the anchor errors. The reference fields specify the conditional approximation being analyzed and are not additional functions fitted by SPR. Geometric coverage controls the first term, while anchor quality and source-target discrepancy enter separately. Both this bound and the preservation and stability properties concern the remapped initialization, before the target table is independently fine-tuned.

C Details of Auxiliary Adaptation Module

The name RADAR is derived from Spatial Positional **R**em**A**pping, **D**ata **A**ugmentation through STR, and Distributionally **R**obust Optimization.

C.1 Details of the STR module

Stochastic Channel Dropout for Spatial Redundancy. To improve robustness to channel-level sparsity and prevent over-reliance on specific sensors during fine-tuning, we introduce Stochastic Channel Dropout (SCD). Rather than simulating explicit changes in electrode layouts, this augmentation regularizes the model to exploit distributed spatial information by randomly removing subsets of channels within a fixed configuration. As a result, the model is encouraged to rely on redundant and globally informative neural patterns across the scalp.

Given an input trial $\mathbf{X} \in \mathbb{R}^{C\times T}$, where C denotes the number of channels and T the number of time points, we sample a binary channel mask $\mathbf{m} \in \{0,1\}^C$ according to

$$m_i \sim \text{Bernoulli}(1 - P_{\text{drop}}), \tag{18}$$

where P_{drop} controls the dropout probability and i is the channel index. The augmented signal $\mathbf{X}'_i$ is computed as

$$\mathbf{X}'_i = \mathbf{X}_i \cdot m_i \cdot \frac{C}{\sum_{j=1}^{C} m_j + \epsilon}, \tag{19}$$

where the rescaling factor preserves the expected signal magnitude across different numbers of active channels, and ϵ is a small constant for numerical stability. To avoid degenerate cases, we enforce a safety constraint ensuring that at least one channel remains active for each trial.

From a geometric perspective, SCD acts as a regularization mechanism that encourages stable spatial representations under partial channel observations, which is beneficial for improving robustness in downstream EEG fine-tuning.

Additive Gaussian Noise for Temporal Robustness. While Stochastic Channel Dropout primarily regularizes spatial over-reliance, we further employ Additive Gaussian Noise to reduce overfitting to precise temporal waveforms. Such a strategy encourages the model to focus on stable neural dynamics rather than incidental fluctuations that may arise from recording noise or hardware variability.

For each input trial selected with probability P_{noise}:

$$\mathbf{X}'' = \mathbf{X} + \mathbf{N}, \quad \mathbf{N} \in \mathbb{R}^{C\times T},\ N_{i,t} \sim \mathcal{N}(0, \sigma^2), \tag{20}$$

where σ controls the noise intensity and $\mathbf{N}$ is sampled independently across channels and time points. By perturbing the input signals with temporally unstructured noise, this operation acts as a regularization mechanism that discourages sensitivity to spurious high-frequency correlations. Such perturbation is particularly relevant for Transformer-based backbones, as it regularizes the model against over-sensitivity to sample-specific, high-frequency temporal patterns that are unlikely to generalize across subjects.

Both channel dropout and additive Gaussian noise are applied only during training and are disabled at standard inference. Taken together, these regularizations create a challenging yet coherent training environment, facilitating the learning of resilient spatiotemporal representations during fine-tuning.

C.2 Details and Interpretation of the DRO Module

The final component of RADAR focuses on the optimization objective. Conventional EEG models are typically trained under Empirical Risk Minimization (ERM, Vapnik, 2013; Vapnik & Chervonenkis, 2015), which minimizes the average loss across all training samples. However, in the presence of data heterogeneity, ERM can bias the model toward easy patterns while under-serving harder sub-populations or under-represented distributions. Thus, we adopt a batch-partitioned loss-reweighting strategy inspired by Distributionally Robust Optimization (DRO) (Mohajerin Esfahani & Kuhn, 2018; Sagawa et al., 2020). The objective assigns greater weight to higher-loss groups within each training batch.

Group Partitioning. Explicit group annotations may be unavailable or costly to obtain, and defining meaningful groups requires a justified criterion rather than arbitrary assignments. Even when subject or domain labels are available, their boundaries need not align with differences in the model's prediction difficulty. We therefore randomly partition each shuffled batch of B samples into G disjoint groups $\{\mathcal{B}_g\}_{g=1}^{G}$, each containing B/G samples, with G dividing B. These groups serve as units for loss aggregation and reweighting without requiring subject or domain annotations or assuming that they represent coherent latent sub-populations. The mean loss for group g is

$$L_g \;=\; \frac{1}{|\mathcal{B}_g|} \sum_{n\in\mathcal{B}_g} \ell\big(\hat{y}_n, y_n\big), \tag{21}$$

where $\ell(\cdot)$ denotes the per-sample loss and $\hat{y}_n$ is the model prediction for input $\tilde{X}_n$.

Softmax Reweighting of Group Risks. To prioritize groups with higher empirical loss, we assign a dynamic weight β_g to each group based on its current mean loss. The training objective is a differentiable softmax-weighted sum of group losses

$$\beta_g \;=\; \frac{\exp\big(\lambda\, L_g\big)}{\sum_{a=1}^{G} \exp\big(\lambda\, L_a\big)}, \quad \mathcal{J} \;=\; \sum_{g=1}^{G} \beta_g\, L_g, \tag{22}$$

where $\lambda \geq 0$ governs the reweighting strength. The dependence of β_g on the group losses is retained during backpropagation.

Rationale and Stabilization. The parameter λ interpolates between ERM ($\lambda = 0$, uniform weighting) and a near worst-case emphasis ($\lambda \to +\infty$, concentrating weight on the highest-loss groups). To avoid destabilizing gradients during fine-tuning of large pre-trained backbones, we apply a linear warmup schedule to λ. Concretely, λ is initialized at zero and then linearly increased to its target value over N_{warm} optimization steps, which mitigates gradient noise introduced by aggressive reweighting early in training.

Batch partitioning and reweighting provide a computationally efficient way to emphasize higher-loss groups without explicit subject or domain annotations.

C.3 Interpretation of DRO.

Random partitioning forms stochastic groups for loss reweighting rather than recovering true physiological or subject-level sub-populations. It avoids requiring a predefined grouping criterion for heterogeneous EEG data.

The softmax weighting rule can be motivated by a KL-regularized DRO problem. For fixed model parameters θ and $\lambda > 0$, consider an adversarial distribution q over the G groups, regularized toward the uniform distribution $u_g = 1/G$,

$$\max_{q \in \Delta_G} \left\{ \sum_{g=1}^{G} q_g L_g(\theta) - \frac{1}{\lambda} \mathrm{KL}(q \| u) \right\}, \tag{23}$$

where Δ_G is the probability simplex. The optimal group weights are

$$q_g^* = \frac{\exp(\lambda L_g(\theta))}{\sum_{a=1}^{G} \exp(\lambda L_a(\theta))} = \beta_g(\theta). \tag{24}$$

Substituting these weights into the regularized objective yields the log-sum-exp (LSE) risk

$$R_\lambda(\theta) = \frac{1}{\lambda} \log \left(\frac{1}{G} \sum_{g=1}^{G} \exp(\lambda L_g(\theta)) \right). \tag{25}$$

This risk is also related to tilted empirical risk minimization (Li et al., 2021).

We use these weights in the training objective $\mathcal{J} = \sum_{g=1}^{G} \beta_g L_g$, differentiating through both the losses and the weights rather than directly optimizing R_λ.

D Dataset Details

Table 2 summarizes the key properties of the three high-density and five native low-density datasets used in this work, including task type, number of channels, sample duration, and dataset scale. These datasets span motor execution, motor imagery, SSVEP, emotion recognition, clinical EEG analysis, and behavioral prediction, covering a wide range of recording setups and subject populations.

Table 2: Statistical properties of the datasets. SSVEP stands for steady-state visually evoked potentials. For CCD, 1/2 denotes one continuous regression target and a two-class classification task, respectively.

Dataset	Task	Classes	Bandpass	Channels	Window Length	Subjects	Total Samples
MAMEM1	SSVEP	5	5–45 Hz	256	3 s	11	1,081
HGD	Motor Execution	4	0.5–75 Hz	128	4 s	14	26,968
CCD	Behavior Prediction	1/2	0.5–50 Hz	128	2 s	1,964	116,086
BCIC-IV-2a	Motor Imagery	4	0.5–45 Hz	22	4 s	9	2,592
FACED	Emotion Recognition	9	0.05–47 Hz	30	10 s	123	10,332
PhysioNet-MI	Motor Imagery	4	0.5–45 Hz	64	3 s	109	19,022
TUAB	Abnormal Detection	2	0.1–75 Hz	23	10 s	2,328	409,083
TUEV	Event Type Classification	6	0.1–75 Hz	23	5 s	370	112,237

Preprocessing. We apply a notch filter at either 50 Hz or 60 Hz depending on the power-line frequency of the data acquisition location to remove line noise. All datasets are band-pass filtered as listed in Table 2 and resampled to 200 Hz.

Dataset Subsampling. We construct fixed subsets of TUAB and TUEV and use them consistently in all experiments, because TUAB and TUEV are substantially larger than the other low-density datasets. For TUEV, we randomly sample approximately 20% of the data from each subject, resulting in 22,302 samples. For TUAB, we sample approximately 5% per subject, yielding 19,538 samples.

D.1 Multimedia Authoring & Management using your Eyes and Mind (MAMEM)

The Multimedia Authoring & Management using your Eyes and Mind (MAMEM) dataset is a public steady-state visually evoked potential (SSVEP) dataset (Oikonomou et al., 2016; Goldberger et al., 2000). Although the dataset contains three distinct experimental configurations, our paper focuses exclusively on MAMEM1. We exclude MAMEM2 from our main analysis because our preliminary evaluations indicate that the performance of various models on MAMEM2 only marginally exceeds chance levels, offering limited discriminative value for further analysis. MAMEM3 uses a 14-channel low-density EEG headset and is not included in our high-density evaluation.

The MAMEM1 and MAMEM2 datasets involve 11 healthy subjects (8 males and 3 females) aged between 24 and 39 years. Data were acquired using a high-density EEG system with 256 channels. The visual stimuli consist of five magenta boxes flickering at distinct frequencies of 6.66 Hz, 7.5 Hz, 8.57 Hz, 10 Hz, and 12 Hz. In the MAMEM1 paradigm, subjects are exposed to non-overlapping flickering lights, where only one box flickers at a single isolated frequency during each stimulus window. MAMEM2 utilizes a simultaneous flickering paradigm where all five boxes flicker at their respective frequencies concurrently while the subject focuses on a target marked by a yellow arrow.

In MAMEM1, each participant completed up to five sessions. Every session begins with a 100-second baseline rest period to stabilize the subject's physiological state. Following this, an adaptation period is conducted, consisting of eight 5-second flickering windows (randomly selected from the five target frequencies) separated by 5-second rest intervals. The main experimental phase comprises 15 trials per session, where each target frequency is presented three times in ascending order. Each trial consists of a 5-second stimulus window followed by a 5-second rest period. Including the adaptation trials, each session yields 23 flickering windows per subject.

We apply a 50-Hz notch filter to suppress power-line interference, followed by a band-pass filter with a frequency range of 5–45 Hz to retain the relevant SSVEP components while removing low-frequency drifts and high-frequency artifacts. Following the standard procedure implemented in the Mother of All BCI Benchmarks (MOABB) open-source repository (Aristimunha et al., 2026), we extract a 3-second temporal window from the 1st to the 4th second of each original 5-second trial. Subsequently, the signals are resampled to 200 Hz to maintain consistency.

D.2 High Gamma Dataset (HGD)

The High-Gamma Dataset (HGD) (Schirrmeister et al., 2017) is a motor execution benchmark recorded in an environment optimized for the detection of high-frequency movement-related EEG components. The dataset comprises recordings from 14 healthy subjects (27.2 ± 3.6 years old, 6 female, 2 left-handed). The original data were acquired using 128 electrodes. The experimental paradigm involves four classes comprising sequential finger-tapping of the left or right hand, repetitive clenching of both feet (toes), and a rest state (no movement). Each subject performed approximately 1,000 four-second trials divided into 13 runs. During each trial, subjects were presented with a gray arrow on a black background for 4 seconds, indicating the type of movement to perform. The order of stimuli was pseudo-randomized, with a continuous random inter-trial interval of 3 to 4 seconds. Subjects performed the repetitive movements at their own pace for the duration of the visual cue.

We apply a band-pass filter between 0.5 and 75 Hz to the raw signals. For temporal segmentation, following the official implementation provided in the HGD GitHub repository[1], we extract two 4-second windows from each trial. The first window covers $[-0.5, 3.5]$ seconds and the second covers

[1]The official repository of the HGD dataset is available at `https://github.com/robintibor/high-gamma-dataset`.

$[0.0, 4.0]$ seconds relative to stimulus onset. Finally, all signals are resampled to 200 Hz to maintain consistency.

D.3 Contrast Change Detection (CCD)

The Healthy Brain Network EEG (HBN-EEG) dataset is a large-scale, curated collection of high-resolution EEG recordings that comprises 11 dataset releases from over 3,000 participants aged 5–21 years, contributed by the Child Mind Institute (Shirazi et al., 2024; Alexander et al., 2017). Our paper specifically utilizes the Contrast Change Detection (CCD) task, an active experimental paradigm designed to measure visual attention and decision-making. In this task, participants monitor two co-centric flickering grated disks (one left-leaning and one right-leaning) and are required to identify which disk exhibits an increase in contrast. Subjects receive immediate feedback (e.g., a smiley or sad face) based on their response accuracy. We evaluate a binary classification task and a regression task that predicts participants' reaction times measured from stimulus onset. For classification, we retain only correctly answered trials and use 2-s EEG epochs spanning $[-1, 1)$ seconds relative to stimulus onset to predict whether the left- or right-tilted grating increased in contrast. For regression, following the official implementation guidelines[2] (Aristimunha et al., 2025), we extract a 2-second window from each trial, covering the interval of $[0.5, 2.5]$ seconds relative to stimulus onset. This interval is selected to capture the stable evoked response following the initial stimulus transition. We apply a 60-Hz notch filter to remove power-line interference, followed by a band-pass filter between 0.5 and 50 Hz to isolate task-relevant neural activity. Finally, all signals are resampled to 200 Hz to maintain consistency.

E Detailed Experiment Setups

E.1 Cross-Subject Dataset Partitions

We use five predefined train, validation, and test partitions for each dataset. Partitions are constructed at the subject or release level, with disjoint subject sets for training, validation, and testing. For MAMEM1, each partition assigns nine of the eleven subjects to training, one to validation, and one to testing. For HGD, twelve of the fourteen subjects are used for training, with one subject each for validation and testing. For CCD, which comprises eleven releases with different subject counts, each partition uses nine releases for training, one for validation, and one for testing. The same five partitions are used across models and adaptation configurations. Each partition is repeated with random seeds 42, 43, and 44, and results are averaged over the resulting 15 runs. Reported standard deviations are computed across the five partition-level means after averaging the three seeds within each partition. Validation data are used for hyperparameter and checkpoint selection.

E.2 Evaluation Metrics

We report balanced accuracy, defined as the unweighted mean of per-class recall, for all EEG classification tasks, including the five native low-density benchmarks. Higher balanced accuracy indicates better performance.

For CCD reaction time regression, NRMSE is the RMSE normalized by the standard deviation of the evaluated targets. Given N evaluated trials with targets y_n and predictions $\hat{y}_n$, we compute

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n - y_n)^2}}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - \bar{y})^2}}, \qquad \bar{y} = \frac{1}{N}\sum_{n=1}^{N} y_n. \tag{26}$$

E.3 General Hyperparameters

Table 3 summarizes the shared downstream training settings and model-specific hyperparameter search spaces. Positional variants within each backbone and task use the same training budget and

[2]The official start-kit for the EEG Foundation Challenge, which utilizes the CCD dataset, can be found at `https://github.com/eeg2025/startkit`.

Table 3: Shared downstream training settings and model-specific hyperparameter search spaces. STR parameters apply only when the module is enabled.

Hyperparameter	Value
Optimizer	AdamW
Training epochs	50
Batch size	64
Learning-rate candidates	$\{1 \times 10^{-4}, 5 \times 10^{-4}\}$
Foundation-model weight decay	1×10^{-2}
Traditional-model weight-decay candidates	$\{1 \times 10^{-3}, 1 \times 10^{-2}, 1 \times 10^{-1}\}$
Learning-rate scheduler	Warmup cosine
Learning-rate warmup epochs	5
Minimum learning rate	1×10^{-6}
Channel dropout probability P_{drop}	0.1
Gaussian noise application probability P_{noise}	0.1
Gaussian noise standard deviation	0.02

shared defaults. The STR probabilities apply when STR is enabled, and the five-epoch learning-rate warmup is separate from the DRO reweighting warmup.

Additional Details on Hyperparameter Tuning. For each foundation model on each task, we select the learning rate from $\{1 \times 10^{-4}, 5 \times 10^{-4}\}$ using validation performance. When DRO is enabled, its group count G, reweighting coefficient λ, and warmup steps N_{warm} are also tuned using validation data. The core SPR configuration remains fixed across experiments without dataset-specific tuning, while STR uses $P_{\text{drop}} = P_{\text{noise}} = 0.1$ by default.

For each traditional model on each task, we jointly search learning rates in $\{1 \times 10^{-4}, 5 \times 10^{-4}\}$ and weight decay values in $\{1 \times 10^{-3}, 1 \times 10^{-2}, 1 \times 10^{-1}\}$. The optimal hyperparameters are selected based on validation performance. For foundation models, weight decay remains fixed at 1×10^{-2}. For dropout, we follow the default architecture settings of 0.5 for traditional models, as implemented in the `braindecode` library. For foundation models, we use the default dropout settings from their respective official implementations. Other general training hyperparameters remain fixed except in the explicitly reported sensitivity analyses.

E.4 Hardware and Software

Model training and fine-tuning are performed on $1\times$ NVIDIA A100 GPU with 40 or 80 GB of memory. Our implementation is based on Python 3.10 and PyTorch 2.7.1. The full list of dependencies and their specific versions will be documented in our accompanying code repository.

E.5 Pretrained Foundation Model Checkpoints

We use the official publicly released pretrained checkpoints for all five foundation models. LaBraM-Base is obtained from the official GitHub checkpoint directory[3]. LUNA provides Base, Large, and Huge checkpoints, and we use LUNA-Base from its official Hugging Face release[4]. REVE provides Base and Large checkpoints, and we use REVE-Base[5]. Since LaBraM's official pretrained release is limited to Base, we also select the Base variants of LUNA and REVE to maintain a consistent variant-selection policy rather than compare LaBraM against larger variants of the other models.

For CBraMod, we use the official pretrained weights hosted on Hugging Face[6]. For CSBrain, we use the pretrained weights in the `pth` directory of the Google Drive release linked from its official

[3] `https://github.com/935963004/LaBraM/tree/main/checkpoints`.

[4] `https://huggingface.co/PulpBio/LUNA/tree/main`.

[5] The Base and Large checkpoints are available at `https://huggingface.co/brain-bzh/reve-base` and `https://huggingface.co/brain-bzh/reve-large`, respectively.

[6] `https://huggingface.co/weighting666/CBraMod/tree/main`.

GitHub repository[7]. CBraMod and CSBrain each provide a single pretrained model size in these releases, so we retain their official names without a size suffix.

SPR modifies only the positional encoding components and their backbone-specific interfaces. All other backbone parameters are initialized from the official checkpoints and remain unaltered prior to fine-tuning.

E.6 Pretraining Coverage of Target Datasets and Layouts

We distinguish exposure to a dataset from exposure to its electrode layout or a related sensor-net family, based on the pretraining sources disclosed for the five foundation models.

Reported Dataset Coverage. LaBraM's disclosed pretraining corpus spans multiple public and self-collected EEG datasets but does not list HGD, MAMEM1, or HBN (CCD is a subset of HBN). CBraMod and CSBrain report pretraining on 19-channel TUEG recordings, while LUNA uses TUEG and Siena with 20/22 bipolar and 29 unipolar channels, respectively. These sources likewise do not include the three target datasets. In contrast, REVE lists HGD under the MOABB name `Schirrmeister2017` and HBN releases R1–R9 among its pretraining sources. MAMEM1 is not named in the disclosed pretraining sources of any of the five models.

HGD and Standard Electrode Locations. HGD uses a dense 128-electrode configuration with standard electrode names rather than a GSN montage. Some scalp locations can therefore overlap with standard montages represented during pretraining, even when the complete target layout is absent. For REVE, HGD is a disclosed pretraining source, although dataset inclusion alone does not specify which channels were retained during pretraining.

GSN-Family Coverage. MAMEM1 uses a 256-channel HydroCel GSN, whereas CCD was acquired with a 128-channel GSN system. REVE's pretraining on HBN provides exposure to GSN-128 layouts relevant to CCD. This experience may also provide transferable spatial knowledge for the GSN-256 layout used in MAMEM1, even though MAMEM1 itself is not listed among its pretraining sources. The other four models' disclosed sources do not establish comparable GSN-family coverage, while still providing spatial priors from standard scalp locations.

Interpretation of the Evaluation. SPR draws its positional prior from LaBraM-Base, whose reported pretraining sources include neither HGD nor GSN-family recordings. REVE, by contrast, has pretraining exposure to HGD and GSN layouts. The comparison with REVE-derived positional priors therefore does not give SPR an advantage in target-data or layout exposure and may instead favor REVE.

E.7 Source Electrode Layout for Remapping

The electrode remapping procedure requires a reference source layout to establish spatial correspondence. In this work, we adopt the predefined electrode configuration used in LaBraM as the source distribution. To ensure consistency and remove potential ambiguities, we further filter this layout using `mne.channels.make_standard_montage("standard_1020")`, retaining only electrodes that match the standard 10-20 system. The coordinate lookup yields 94 electrode names, including four pairs of aliases at identical locations. We retain T7, T8, P7, and P8 and exclude their respective aliases T3, T4, T5, and T6, leaving 90 unique electrode positions for SPR. Non-standard channels, such as bipolar montages, are excluded from the source layout.

E.8 Backbone-Specific Positional Integration

Each receiving model uses its own pretrained signal encoder, while SPR supplies 200-dimensional electrode vectors from the remapped LaBraM table. The table is constructed before fine-tuning as described in Section 3.2, with rows aligned to the target input channels. Spatial vectors are broadcast across temporal patches and incorporated through the backbone-specific interfaces described below.

[7]Official repository at `https://github.com/yuchen2199/CSBrain` and checkpoint release at `https://drive.google.com/drive/folders/1-GsVVewRM0B93H08yts5m53yU2whxYvj`.

Table 4: Cross-subject LaBraM results with frozen positional embeddings and a fine-tuned backbone and task head. Entries are unnormalized means, with balanced accuracy for HGD and MAMEM1 and NRMSE for CCD-reg regression. CCD-reg uses a subset of 20 subjects from each of the 11 releases, totaling 220 subjects. N/A denotes insufficient source vectors for MAMEM1's 256 electrodes.

Positional Encoding	HGD ↑	MAMEM1 ↑	CCD-reg ↓
Random PE	0.659	0.449	0.957
Unaligned Pretrained PE	0.683	N/A	0.961
SPR	**0.727**	**0.552**	**0.950**

Rationale for Cross-Backbone Evaluation. Our cross-backbone evaluation examines whether pretrained positional representations remain useful beyond the architecture in which they were learned. LaBraM supplies an electrode-indexed spatial prior, while different receiving backbones provide distinct settings for testing its reuse. We assess this utility through performance changes within each receiving backbone rather than its ranking relative to LaBraM. In practice, the choice of foundation model depends on task requirements and deployment constraints, as no single EEG foundation model has been established as uniformly superior across tasks. Reusing positional priors offers an additional way to improve these existing checkpoints without replacing the signal encoder or repeating large-scale pretraining.

Common Settings. SPR uses learned projections where needed to match the receiving width. For non-REVE backbones, transferred REVE-PE evaluates the pretrained 512-dimensional Fourier-plus-MLP position function at $(x, y, z, 0)$ in head-centered coordinates measured in meters, retaining its normalization and using fixed dimension selection to match the receiving width. Unless explicitly frozen, positional parameters and interfaces are fine-tuned with the receiving backbone.

Backbone Integration. LaBraM replaces its spatial position contribution with SPR vectors or REVE-PE outputs while retaining temporal embeddings and CLS. CBraMod removes its positional convolution and concatenates 100-dimensional spatial and temporal components, with the same anchor-derived temporal initialization for both schemes. CSBrain removes its positional convolution and directly adds 200-dimensional spatial vectors from either SPR or REVE-PE, without an additional temporal table. LUNA bypasses its original coordinate-encoding branch and replaces its output with projected SPR vectors or REVE-PE outputs before channel cross-attention, retaining temporal RoPE. REVE adds projected SPR vectors to its original positional output evaluated at zero spatial coordinates and the current time index, retaining positional LayerNorm.

Control Definitions. Random PE randomizes the target electrode table, whereas Unaligned Pretrained PE assigns source vectors by index without geometric alignment. These controls change the electrode-table component while retaining other spatial and temporal mechanisms. Unaligned Pretrained PE is unavailable for MAMEM1 because its 256 channels exceed the 128 source vectors. Native LaBraM uses a random electrode table on high-density datasets and coincides with SPR on native low-density layouts. Non-REVE Random-Init REVE-PE constructs a new randomly initialized three-dimensional coordinate encoder at the receiving width, while REVE resets its learned positional mapping and normalization but retains fixed Fourier features.

F Additional Results

F.1 Full Positional Encoding Results

Tables 4 and 5 provide the full numerical results for the positional knowledge analysis in Section 4.3. The former compares frozen positional embeddings on LaBraM, while the latter compares trainable positional schemes across all five receiving backbones.

Table 5: Further study of positional encoding variants across backbones. Balanced accuracy is used for classification tasks and NRMSE for regression. Average indicates the mean across classification tasks. Brown text reports the standard deviation across five dataset splits. Light gray rows highlight our SPR method. N/A indicates that MAMEM1's 256 electrodes exceed the number of positional embeddings natively available in LaBraM.

Model	HGD ↑	MAMEM1 ↑	CCD-class ↑	CCD-reg ↓	Average ↑
LaBraM	0.664 ±.062	0.447 ±.133	0.763 ±.036	0.956 ±.015	0.625
w/ Unaligned Pretrained PE	0.701 ±.048	N/A	0.777 ±.030	0.952 ±.016	N/A
w/ SPR	**0.727** ±.041	**0.579** ±.209	**0.781** ±.031	**0.946** ±.013	**0.696**
w/ Random-Init REVE-PE	0.612 ±.050	0.483 ±.192	0.759 ±.024	0.955 ±.013	0.618
w/ REVE-PE	0.636 ±.058	0.495 ±.149	0.770 ±.028	0.960 ±.013	0.634
CBraMod	0.667 ±.059	0.364 ±.117	0.726 ±.034	0.961 ±.014	0.586
w/ Random PE	0.663 ±.038	0.370 ±.081	0.738 ±.029	0.958 ±.008	0.590
w/ Unaligned Pretrained PE	**0.687** ±.036	N/A	0.741 ±.034	0.965 ±.007	N/A
w/ SPR	0.686 ±.025	**0.377** ±.129	**0.746** ±.029	**0.956** ±.012	**0.603**
w/ Random-Init REVE-PE	0.642 ±.058	0.200 ±.000	0.680 ±.096	0.967 ±.009	0.507
w/ REVE-PE	0.642 ±.079	0.242 ±.085	0.630 ±.106	0.962 ±.009	0.505
CSBrain	0.643 ±.048	0.367 ±.131	0.701 ±.028	0.963 ±.014	0.570
w/ Random PE	0.669 ±.067	0.373 ±.117	0.701 ±.035	0.965 ±.013	0.581
w/ Unaligned Pretrained PE	0.672 ±.050	N/A	0.695 ±.032	0.967 ±.012	N/A
w/ SPR	**0.681** ±.072	**0.376** ±.106	**0.702** ±.034	**0.962** ±.010	**0.586**
w/ Random-Init REVE-PE	0.641 ±.047	0.364 ±.096	0.701 ±.025	0.967 ±.013	0.569
w/ REVE-PE	0.640 ±.054	0.328 ±.116	0.698 ±.027	0.973 ±.020	0.555
LUNA	0.607 ±.090	0.272 ±.056	0.625 ±.066	1.061 ±.039	0.501
w/ Random PE	0.614 ±.068	**0.294** ±.107	0.670 ±.035	1.052 ±.049	**0.526**
w/ Unaligned Pretrained PE	**0.626** ±.069	N/A	0.666 ±.040	**1.051** ±.043	N/A
w/ SPR	**0.626** ±.070	0.270 ±.077	**0.673** ±.040	1.059 ±.054	0.523
w/ Random-Init REVE-PE	0.621 ±.085	0.242 ±.079	0.650 ±.076	1.058 ±.039	0.504
w/ REVE-PE	0.625 ±.086	0.234 ±.022	0.564 ±.072	1.060 ±.040	0.474
REVE	0.657 ±.126	**0.573** ±.205	**0.748** ±.027	0.958 ±.026	**0.659**
w/ Random PE	0.608 ±.141	0.547 ±.194	0.715 ±.029	0.955 ±.012	0.623
w/ Unaligned Pretrained PE	0.519 ±.128	N/A	0.725 ±.031	0.954 ±.016	N/A
w/ SPR	**0.665** ±.058	0.569 ±.198	0.734 ±.026	**0.950** ±.013	0.656
w/ Random-Init REVE-PE	0.627 ±.061	0.566 ±.186	0.720 ±.027	**0.950** ±.010	0.637

F.2 Full Component Ablation Results

Table 6 reports the task-level results underlying Section 4.4 for all five backbones, including REVE. It compares native encoding, SPR alone, STR+DRO with native encoding, and full RADAR.

Incremental Contributions of STR and DRO. Table 7 compares native encoding, SPR, SPR+STR, and full RADAR on LaBraM and CBraMod. Adding STR to SPR improves both classification tasks on both backbones. Adding DRO to SPR+STR yields further gains on both tasks, supporting the incremental contribution of each auxiliary module beyond positional adaptation.

F.3 Spatial Coordinate Probing

This experiment complements Section 4.5 by testing whether pretrained positional vectors retain spatial structure that remains linearly decodable after remapping. It uses only frozen positional tables and electrode coordinates, without EEG samples, backbone training, or downstream fine-tuning.

Source-Only Probe Fitting. We extract the 200-dimensional electrode vectors from the pretrained LaBraM table and retain 90 unique source locations with available coordinates, excluding special tokens, bipolar channels, and duplicate-coordinate aliases. We use a fixed split of 70 source locations for training and 20 for validation, shared across all representations and both target layouts. The regression targets are three-dimensional unit-sphere coordinates rather than physical distances in millimeters. For each representation, we standardize embedding features using only the training source locations and fit a linear ridge regression probe. The ridge penalty is selected by coordinate

Table 6: Ablation study of RADAR components across backbones. Balanced accuracy is used for classification tasks and NRMSE for regression. Average indicates the mean across classification tasks. Brown text reports the standard deviation across five dataset splits.

Model	HGD ↑	MAMEM1 ↑	CCD-class ↑	CCD-reg ↓	Average ↑
LaBraM	0.664 ±.062	0.447 ±.133	0.763 ±.036	0.956 ±.015	0.625
w/ SPR	0.727 ±.041	0.579 ±.209	**0.781** ±.031	**0.946** ±.013	0.696
w/ STR+DRO	0.675 ±.057	0.480 ±.152	0.763 ±.012	0.955 ±.017	0.639
w/ RADAR	**0.749** ±.046	**0.597** ±.198	0.780 ±.028	0.947 ±.013	**0.709**
CBraMod	0.667 ±.059	0.364 ±.117	0.726 ±.034	0.961 ±.014	0.586
w/ SPR	0.686 ±.025	0.377 ±.129	0.746 ±.029	**0.956** ±.012	0.603
w/ STR+DRO	0.682 ±.044	0.371 ±.124	0.745 ±.030	0.961 ±.012	0.600
w/ RADAR	**0.709** ±.046	**0.410** ±.135	**0.751** ±.032	0.957 ±.010	**0.623**
CSBrain	0.643 ±.048	0.367 ±.131	0.701 ±.028	0.963 ±.014	0.570
w/ SPR	0.681 ±.072	0.376 ±.106	0.702 ±.034	0.962 ±.010	0.586
w/ STR+DRO	0.675 ±.049	0.346 ±.116	0.720 ±.027	**0.958** ±.014	0.580
w/ RADAR	**0.691** ±.058	**0.377** ±.090	**0.723** ±.026	0.959 ±.016	**0.597**
LUNA	0.607 ±.090	0.272 ±.056	0.625 ±.066	1.061 ±.039	0.501
w/ SPR	0.626 ±.070	0.270 ±.077	**0.673** ±.040	**1.059** ±.054	**0.523**
w/ STR+DRO	**0.640** ±.061	0.276 ±.052	0.557 ±.071	1.060 ±.035	0.491
w/ RADAR	0.635 ±.078	**0.310** ±.058	0.581 ±.102	1.062 ±.029	0.509
REVE	0.657 ±.126	0.573 ±.205	0.748 ±.027	0.958 ±.026	0.659
w/ SPR	0.665 ±.058	0.569 ±.198	0.734 ±.026	0.950 ±.013	0.656
w/ STR+DRO	**0.703** ±.076	**0.621** ±.224	**0.750** ±.031	**0.946** ±.014	**0.691**
w/ RADAR	0.660 ±.054	0.570 ±.180	0.732 ±.026	0.952 ±.016	0.654

Table 7: Incremental cross-subject component ablation on LaBraM and CBraMod.

Model	HGD ↑	MAMEM1 ↑
LaBraM	0.664	0.447
w/ SPR	0.727	0.579
w/ SPR+STR	0.743	0.583
w/ RADAR	**0.749**	**0.597**
CBraMod	0.667	0.364
w/ SPR	0.686	0.377
w/ SPR+STR	0.696	0.387
w/ RADAR	**0.709**	**0.410**

RMSE on the source validation locations, which are not subsequently added to the training set. The fitted transformation and probe are then applied directly to the existing remapped GSN-128 and GSN-256 tables, without a second remapping step or target-layout probe fitting and selection.

Matched Representation Controls. Random replaces pretrained source vectors with independently sampled Gaussian vectors whose per-dimension means and standard deviations are estimated from the training source vectors. Shuffled instead permutes the pretrained vector-electrode assignments while keeping source coordinates and the training-validation split fixed. Each control source table is mapped to the target layouts using the same geometric remapping as SPR, and a separate probe is fitted and selected under the same source-only protocol. The reported Random and Shuffled results are averaged over 20 random tables and 20 fixed permutations, respectively. These repetitions measure variation in the control representations rather than variation across EEG training runs or source-location splits. The 70/20 split governs probe supervision, not the construction of the target tables, whose remapping can also use source rows outside the probe training subset.

Evaluation and Interpretation. Table 8 reports coordinate NRMSE, mean angular error, and R^2 over all electrodes in each target layout. Coordinate errors are computed from the raw ridge predictions, while angular error compares their normalized directions with the true unit-sphere directions. Since geometric remapping itself introduces spatial relationships, the matched controls distinguish this contribution from the additional information carried by pretrained vectors and their electrode

Table 8: Coordinate prediction from frozen positional embeddings using linear ridge probes trained on 70 source electrodes and selected on 20 source validation electrodes. Targets are unit-sphere coordinates, and evaluation on GSN-128 and GSN-256 uses no target-layout fitting or model selection. Random uses random source embeddings under matched remapping, while Shuffled disrupts pretrained vector-electrode correspondences. Angle denotes mean angular error in degrees. Bold indicates the best result for each metric and layout.

Representation	**GSN-128**			**GSN-256**		
	NRMSE ↓	**Angle** (°) ↓	R^2 ↑	**NRMSE** ↓	**Angle** (°) ↓	R^2 ↑
SPR	**0.3636**	**7.18**	**0.8676**	**0.4096**	**8.36**	**0.8312**
Random	0.8571	49.46	0.2498	0.8721	55.41	0.2252
Shuffled	0.9301	62.67	0.1267	0.9429	69.24	0.1039

correspondences. SPR achieves lower coordinate and angular errors than both controls on both layouts, supporting the preservation of linearly decodable spatial structure through positional transfer.

F.4 Low-Density Evaluation

We further evaluate SPR on native low-density benchmarks and channel-reduced high-density recordings to examine its applicability across electrode configurations. The two evaluations assess positional knowledge reuse across additional datasets and under reduced target-channel availability, respectively.

F.4.1 Native Low-Density Benchmarks

Setup. We evaluate BCIC-IV-2a (Tangermann et al., 2012), FACED (Chen et al., 2023), PhysioNet-MI (Schalk et al., 2004), and the two Temple University benchmarks, TUAB and TUEV (Obeid & Picone, 2016), using their native electrode configurations. Table 9 compares positional schemes within each receiving backbone. We compare native encoding, Random PE, and SPR, and additionally evaluate REVE-PE on the four non-REVE backbones. Random PE randomly initializes independent target electrode vectors while retaining the pretrained receiving backbone. SPR directly reuses the corresponding LaBraM vectors without interpolation on these layouts, so the LaBraM native row also represents SPR. We use balanced accuracy for all five native low-density benchmarks. Each dataset entry reports the mean balanced accuracy and the standard deviation across five dataset splits, and Average is the unweighted mean across the five datasets.

Results. SPR exceeds Random PE in average balanced accuracy on all five backbones and in 24 of the 25 backbone-dataset comparisons, with CSBrain on TUAB as the sole exception. Since these layouts require no interpolation of the pretrained vectors, this advantage supports the value of pretrained positional content beyond independent electrode-wise parameterization. Including LaBraM's SPR-equivalent native encoding, SPR outperforms REVE-PE on all five datasets across the four non-REVE backbones. SPR also improves average performance over native encoding on CBraMod, CSBrain, and LUNA, with gains on four, three, and all five datasets, respectively. REVE's overall preference for its native encoding is consistent with the expected compatibility between its backbone and the positional mechanism used during pretraining. Even in this setting, SPR improves TUEV and matches TUAB at the displayed precision. Together, these results demonstrate that SPR's value extends beyond high-density inputs, enabling pretrained positional knowledge to benefit multiple receiving architectures.

F.4.2 Channel-Reduced High-Density Benchmarks

Setup. We retain 16, 32, or 64 EEG channels from MAMEM1 and the CCD classification data, yielding six combinations of task and channel count. For each configuration, we evaluate LaBraM with its native positional encoding, REVE-PE, and SPR. For SPR, we extract the corresponding electrode rows from the positional table remapped to the full target montage, without recomputing interpolation for each subset. Table 10 reports mean balanced accuracy and the standard deviation

Table 9: Positional encoding comparisons on native low-density EEG datasets. For LaBraM, native encoding coincides with SPR. Average denotes the mean across the five datasets, and higher values indicate better performance. Brown text reports the standard deviation across five dataset splits. Bold indicates the best mean within each backbone for each column. Light gray rows highlight our SPR method.

Model	BCIC-IV-2a ↑	FACED ↑	PhysioNet-MI ↑	TUAB ↑	TUEV ↑	Average ↑
LaBraM	**0.437** ±.021	**0.552** ±.050	**0.576** ±.048	**0.831** ±.015	**0.607** ±.097	**0.600**
w/ REVE-PE	0.343 ±.007	0.273 ±.028	0.537 ±.054	0.788 ±.010	0.490 ±.044	0.486
w/ Random PE	0.352 ±.049	0.413 ±.038	0.536 ±.051	0.825 ±.014	0.571 ±.050	0.540
CBraMod	0.435 ±.028	0.570 ±.084	0.556 ±.054	0.804 ±.006	**0.504** ±.101	0.574
w/ REVE-PE	0.410 ±.032	0.529 ±.128	0.528 ±.052	0.792 ±.015	0.446 ±.081	0.541
w/ Random PE	0.429 ±.035	0.626 ±.083	0.552 ±.053	0.802 ±.012	0.473 ±.088	0.576
w/ SPR	**0.439** ±.018	**0.648** ±.087	**0.558** ±.050	**0.806** ±.012	0.477 ±.099	**0.586**
CSBrain	0.429 ±.027	0.625 ±.077	0.561 ±.060	0.814 ±.012	**0.514** ±.065	0.589
w/ REVE-PE	0.432 ±.023	0.586 ±.082	0.543 ±.053	0.796 ±.007	0.448 ±.096	0.561
w/ Random PE	0.431 ±.023	0.634 ±.075	0.556 ±.055	**0.815** ±.012	0.491 ±.087	0.585
w/ SPR	**0.437** ±.015	**0.638** ±.084	**0.563** ±.051	0.813 ±.015	0.506 ±.080	**0.592**
LUNA	0.267 ±.012	0.400 ±.027	0.529 ±.033	0.792 ±.011	0.513 ±.046	0.500
w/ REVE-PE	0.261 ±.015	0.266 ±.052	0.518 ±.045	0.792 ±.006	0.523 ±.052	0.472
w/ Random PE	0.256 ±.013	0.376 ±.111	0.522 ±.053	0.796 ±.010	0.516 ±.072	0.493
w/ SPR	**0.271** ±.009	**0.421** ±.042	**0.542** ±.053	**0.801** ±.009	**0.548** ±.039	**0.516**
REVE	**0.492** ±.084	**0.610** ±.081	**0.623** ±.047	**0.846** ±.011	0.638 ±.077	**0.642**
w/ Random PE	0.405 ±.030	0.582 ±.061	0.561 ±.050	0.831 ±.008	0.605 ±.084	0.597
w/ SPR	0.442 ±.053	0.596 ±.067	0.570 ±.051	**0.846** ±.012	**0.663** ±.124	0.623

across five dataset splits for each configuration. This evaluation varies channel availability within the same datasets and task paradigms, complementing the separate native low-density benchmarks.

Electrode Selection. We construct spatially distributed channel subsets using deterministic farthest-point sampling on the unit sphere. Using the full target montage coordinates, we restrict candidates to electrodes with $z \geq 0$ in the existing head coordinate system, yielding 137 candidates for MAMEM1 and 86 for CCD classification. This geometric restriction focuses sampling on the upper scalp rather than peripheral facial or neck locations. We normalize each candidate coordinate to unit length and initialize selection with the electrode having the largest normalized z coordinate. At each subsequent step, we select the candidate with the largest minimum chord distance to the selected directions, with ties resolved by original channel index. We generate one ordered sequence of 64 electrodes and use its first 16 and 32 entries for the smaller configurations, producing nested subsets. Selection uses only geometry, without labels or decoding scores, and the subsets remain fixed across positional schemes, folds, and training seeds. Before model input, retained channels are sorted by their original indices, with EEG signals, coordinates, and SPR table rows aligned to the same order. Each configuration uses the same subset for training, validation, and testing, while preserving cached preprocessing, including the original reference, and the existing sample and subject splits.

Results. SPR improves native LaBraM in all six configurations and achieves the highest mean performance in each. The gains over native encoding range from 2.9 to 4.7 percentage points on MAMEM1 and from 0.6 to 2.1 points on CCD classification. The advantage persists from 16 to 64 channels, showing that SPR continues to exploit pretrained positional knowledge when fewer target electrodes are available. Together with the native low-density results, these findings establish the practical value of SPR across both different recording configurations and reduced-channel inputs.

F.5 Signal-Level Remapping to LaBraM Electrode Positions

Motivation. An alternative to adapting positional representations is to remap observed EEG signals to electrode locations already represented in a pretrained positional table. We evaluate this route with LaBraM, comparing nearest-neighbor signal assignment (Signal-NN) and four-neighbor signal interpolation (Signal-KNN4) against native adaptation and SPR.

Table 10: LaBraM performance on channel-reduced MAMEM1 and CCD classification tasks. Columns specify the number of retained EEG channels. Balanced accuracy is reported. Brown text reports the standard deviation across five dataset splits. Bold indicates the best mean for each task and channel count, including ties at the displayed precision.

Model	MAMEM1 ↑			CCD-class ↑		
	16 channels	32 channels	64 channels	16 channels	32 channels	64 channels
LaBraM	0.417 ±.149	0.569 ±.195	0.551 ±.205	0.713 ±.029	0.733 ±.027	0.733 ±.011
w/ REVE-PE	0.413 ±.167	0.487 ±.162	0.516 ±.190	**0.720** ±.018	0.735 ±.030	0.742 ±.031
w/ SPR	**0.457** ±.183	**0.598** ±.235	**0.598** ±.247	**0.720** ±.028	**0.739** ±.022	**0.754** ±.034

Table 11: Signal-level and positional remapping for LaBraM. Signal-NN and Signal-KNN4 map EEG signals to 90 native electrode positions and reuse the corresponding pretrained positional rows, whereas SPR retains the observed signals and remaps positional vectors. Classification columns report balanced accuracy, CCD-reg reports NRMSE, and Average is the mean across the three classification tasks. Brown text reports standard deviations. Bold indicates the best mean in each column.

Method	HGD ↑	MAMEM1 ↑	CCD-class ↑	CCD-reg ↓	Average ↑
Native LaBraM	0.664 ±.062	0.447 ±.133	0.763 ±.036	0.956 ±.015	0.625
Signal-NN	0.697 ±.050	0.516 ±.241	0.776 ±.030	0.953 ±.021	0.663
Signal-KNN4	0.712 ±.032	0.491 ±.218	0.762 ±.027	0.955 ±.015	0.655
SPR	**0.727** ±.041	**0.579** ±.209	**0.781** ±.031	**0.946** ±.013	**0.696**

Target Layout and Positional Encodings. Both variants map recorded EEG to the 90 unique LaBraM electrode positions defined in Appendix E.7. LaBraM retains its original 129-row checkpoint positional table and selects the corresponding 90 electrode rows by their original token indices, together with the CLS row. The pretrained spatial and temporal position embeddings and CLS embedding are neither interpolated nor reinitialized and remain trainable.

Geometric Signal Remapping. Signal-NN assigns each of the 90 target LaBraM positions the nearest recorded channel under unit-sphere angular distance. Signal-KNN4 uses the same four-neighbor distance-softmax weighting and exact-match rule as SPR in Equation 3. The geometry-derived mapping is fixed and applied to EEG signals at each time point rather than to pretrained positional vectors.

Results. Both variants follow the existing full-model fine-tuning and validation-based checkpoint selection procedure. Table 11 reports the signal-level comparisons alongside the native LaBraM and SPR references from Table 1. Signal-NN and Signal-KNN4 improve average balanced accuracy from 0.625 to 0.663 and 0.655, respectively, showing that signal-level alignment provides useful adaptation. SPR achieves 0.696 average balanced accuracy and the best mean performance on all four tasks, including regression NRMSE of 0.946 compared with 0.953 and 0.955 for the signal-level alternatives. Four-neighbor mixing improves HGD over nearest-neighbor assignment but does not improve the other tasks, indicating that neighborhood signal aggregation is not uniformly beneficial. Despite the gains from signal-level alignment, these results favor retaining the observed electrode signals and adapting positional representations under the evaluated configurations.

F.6 Input-Level ROI Remapping

Motivation. Projection to a common regional representation has been used in conventional EEG transfer methods to accommodate heterogeneous electrode layouts (Peterson et al., 2021). Inspired by this strategy, we examine whether geometric ROI remapping at the signal input remains useful for pretrained EEG foundation models. This approach contrasts with SPR in where remapping occurs, mixing electrode signals before the backbone rather than adapting positional embeddings while preserving the measured EEG.

Setup. We project the 128-channel HGD and CCD recordings and the 256-channel MAMEM1 recordings to 64 virtual regions of interest (ROIs). These ROIs denote geometric signal aggregates

Table 12: Input-level ROI remapping across five foundation models. Classification columns report balanced accuracy, and CCD-reg reports NRMSE on the full CCD regression dataset. Brown text denotes the standard deviation across partitions. Bold indicates the best mean within each backbone for each task, including ties at the displayed precision.

Model	HGD ↑	MAMEM1 ↑	CCD-class ↑	CCD-reg ↓
LaBraM	**0.664** ±.062	**0.447** ±.133	**0.763** ±.036	**0.956** ±.015
w/ ROI	0.644 ±.030	0.404 ±.120	0.699 ±.036	0.969 ±.015
CBraMod	**0.667** ±.059	**0.364** ±.117	**0.726** ±.034	**0.961** ±.014
w/ ROI	0.645 ±.066	**0.364** ±.088	0.689 ±.022	0.965 ±.008
CSBrain	**0.643** ±.048	**0.367** ±.131	**0.701** ±.028	**0.963** ±.014
w/ ROI	0.608 ±.033	0.337 ±.119	0.674 ±.027	0.968 ±.012
LUNA	**0.607** ±.090	**0.272** ±.056	**0.625** ±.066	**1.061** ±.039
w/ ROI	0.585 ±.074	0.258 ±.022	0.523 ±.042	1.065 ±.031
REVE	**0.657** ±.126	**0.573** ±.205	**0.748** ±.027	0.958 ±.026
w/ ROI	0.648 ±.053	0.471 ±.143	0.678 ±.026	**0.957** ±.018

rather than anatomically localized brain sources. For each dataset, K-means clustering of the three-dimensional electrode coordinates produces 64 anchors, and an isotropic Gaussian kernel assigns electrode weights to each anchor. The kernel bandwidth is the median nearest-neighbor distance among the anchors, and each row of the projection matrix is normalized to sum to one. The resulting transformation is $X_{\mathrm{ROI}} = AX$, where $A \in \mathbb{R}^{64\times C}$ is fixed and nonlearnable and C is the original channel count. Anchors are constructed using geometry alone and remain fixed across folds and models within each dataset. Unlike channel selection, this transformation mixes electrode signals while preserving the temporal axis. The models are configured for 64 input channels, with ROI anchor coordinates supplied to backbones that require spatial coordinates.

All five pretrained backbones are fine-tuned with their task heads, with STR and DRO disabled. The CBraMod ROI configuration uses its absolute-position variant initialized from the GSN-256 adapted checkpoint. Checkpoints are selected by validation balanced accuracy for classification and validation NRMSE for regression. Table 12 reports the ROI results alongside the native references from Table 1.

Results. The ROI configurations yield lower average balanced accuracy across the three classification tasks on all five backbones, with reductions of approximately 2 to 6 percentage points relative to the native references. A central objective of EEG foundation models is to learn reusable representations across heterogeneous acquisition settings. For a backbone that already accepts the target layout, additional input projection may discard useful electrode-specific information or alter signal relationships that its pretrained representations could otherwise exploit. The observed reductions are consistent with this possibility and motivate SPR's alternative of remapping positional embeddings while leaving the original electrode signals unchanged.

F.7 Comparison of Positional Remapping Rules

To evaluate the local aggregation rule used by SPR, we compare three ways of constructing target positional vectors from pretrained LaBraM embeddings while keeping the remaining training settings unchanged. Nearest assigns each target electrode the vector of its closest source electrode, whereas Uniform averages the four nearest source vectors with equal weights. SPR uses the same four-neighbor neighborhood with distance-based softmax weights. This comparison examines both the use of multiple spatial anchors and the contribution of distance-aware weighting.

Table 13 shows that SPR achieves the best mean performance on both cross-subject classification tasks. On MAMEM1, balanced accuracy reaches 0.579, compared with 0.541 for Nearest and 0.537 for Uniform. SPR also improves HGD classification, providing consistent support for its local aggregation design. Unlike single-anchor assignment, SPR combines information from several nearby source locations, while its distance weights distinguish closer anchors from more distant neighbors.

Table 13: Cross-subject LaBraM performance under different positional remapping rules. Entries report mean balanced accuracy for MAMEM1 and HGD. Bold indicates the best mean in each column.

Remapping	MAMEM1 ↑	HGD ↑
Nearest ($K = 1$)	0.541	0.723
Uniform ($K = 4$)	0.537	0.725
SPR ($K = 4$)	**0.579**	**0.727**

Table 14: Mean performance of LaBraM under data augmentation alone, without SPR or DRO. Balanced accuracy is reported for MAMEM1 and NRMSE for CCD-20.

Augmentation	MAMEM1 ↑	CCD-20 ↓
Gaussian Noise	0.522	**0.983**
TimeReverse	**0.527**	0.991
Channel Dropout + Gaussian Noise	0.527	**0.980**
Channel Dropout + Gaussian Noise + FTSurrogate	**0.536**	0.983

These results favor geometry-guided reuse of pretrained vectors over either direct nearest-neighbor assignment or equally weighted aggregation.

F.8 Augmentation Choices for STR

Prior studies have shown that the effectiveness of data augmentation in EEG is highly task-dependent, with different augmentation strategies being optimal for different paradigms Rommel et al. (2022). Motivated by this observation, we design the STR module to be both simple and flexible. Specifically, we adopt two widely used and easy-to-implement augmentations Wang et al. (2018); Saeed et al. (2021), channel dropout (spatial) and Gaussian noise (temporal), as the default configuration, while keeping the module fully modular so that alternative augmentations can be easily incorporated or replaced.

We evaluate data augmentation alone on LaBraM, with SPR and DRO disabled. The regression task uses CCD-20, a subset containing 20 subjects from each of the 11 releases, totaling 220 subjects. To examine alternative augmentation choices, we compare Gaussian noise with TimeReverse and evaluate the addition of FTSurrogate to the default channel dropout and Gaussian noise combination in Table 14. Adding channel dropout to Gaussian noise improves both MAMEM1 classification and CCD-20 regression, with the default combination achieving the lowest regression error among the evaluated choices. FTSurrogate further improves MAMEM1 balanced accuracy from 0.527 to 0.536, whereas the default combination retains better CCD-20 regression performance. TimeReverse also improves MAMEM1 relative to Gaussian noise alone but yields higher CCD-20 regression error. These findings are consistent with prior observations and reinforce the motivation behind STR. Rather than relying on a single fixed augmentation, RADAR keeps these auxiliary modules configurable, allowing augmentation choices and loss reweighting to be tailored to specific EEG tasks. This flexibility preserves SPR's geometric construction and its central role in positional knowledge reuse while accommodating complementary training strategies.

F.9 Sensitivity to Auxiliary Hyperparameters

For the CCD regression hyperparameter search, we use CCD-20, a subset containing 20 subjects from each of the 11 releases, totaling 220 subjects.

F.9.1 Spatio-Temporal Regularization

To examine STR hyperparameters, we perform a grid search over the channel dropout probability P_{drop} and the probability of applying Gaussian noise P_{noise}. The results in Figure 4 favor moderate regularization, while excessive perturbation can reduce decoding performance.

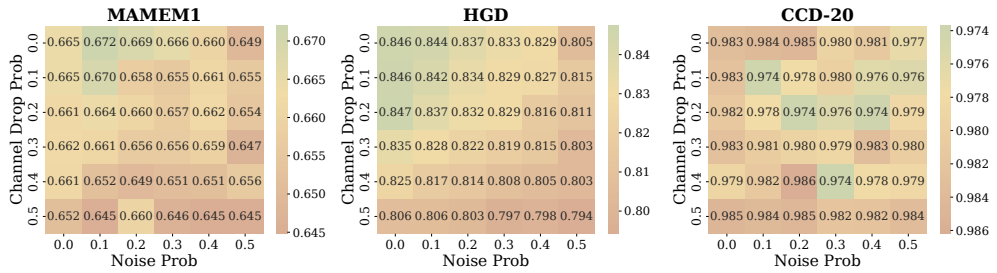

Figure 4: Sensitivity analysis of Spatio-Temporal Regularization hyper-parameters. The heatmaps illustrate the impact of channel dropping probability (y-axis) and Gaussian noise probability (x-axis) on LaBraM's performance on validation set across MAMEM1, HGD, and CCD-20 under cross-subject setting. The results demonstrate that task-specific regularization policies are essential for optimal cross-subject generalization.

Task-Specific Response to Regularization. The impact of these regularization intensities varies with task complexity and information density. For the motor execution task in HGD, we observe a clear diagonal gradient in the performance landscape. While mild regularization provides marginal gains, the accuracy declines monotonically as the combined intensity of spatial and temporal noise increases, suggesting that simpler paradigms require more conservative constraints to preserve signal fidelity. MAMEM1 and CCD-20 exhibit more localized optima, suggesting that their preferred augmentation settings are more task-specific.

Despite these task-specific optima, performance remains stable over a range of moderate hyperparameter values. In particular, $P_{\text{drop}} = P_{\text{noise}} = 0.1$ provides a competitive starting point across the evaluated datasets, supporting practical adaptation with limited validation-based tuning.

F.9.2 Distributionally Robust Optimization

To evaluate the robustness of the DRO module under cross-subject adaptation, we analyze the group number G, the reweighting coefficient λ, and the warmup steps N_{warm} using LaBraM validation performance. Figure 5 summarizes the validation performance distributions across these configurations.

Overall Robustness. Across all three datasets, cross-subject validation performance varies modestly across the evaluated DRO configurations. For instance, under the cross-subject setting, the standard deviation of the validation metric across hyperparameter configurations is 0.00459 for HGD (balanced accuracy), 0.00373 for MAMEM1 (balanced accuracy), and 0.00215 for CCD-20 (NRMSE). Importantly, this stability is observed without relying on carefully tuned hyper-parameter combinations, suggesting that the DRO module in our framework provides reliable performance. As a result, extensive grid search is not required to obtain competitive results, which significantly reduces the practical cost of deployment, especially in cross-subject generalization scenarios.

Group Number and Batch Size. Regarding the group number G, we observed that a finer granularity (e.g., $G = 16$) does not necessarily yield superior performance compared to $G \in \{4, 8\}$. In fact, $G = 4$ and $G = 8$ showed consistent performance parity. We attribute this to the interplay between group size and batch size (B). In our experiments, with batch size of $B = 64$, setting $G = 8$ ensures approximately 8 samples per group, which is sufficient for statistical estimation. However, we hypothesize that for smaller batch sizes (e.g., $B \leq 16$), a smaller group number (e.g., $G = 2$ or 4) would be a more robust choice to guarantee sufficient sample density within each partition.

Impact of Reweighting Strength. For the reweighting coefficient, while the model is generally stable, we observed that an aggressively large λ (e.g., $\lambda = 5$) can be detrimental, particularly on simpler datasets like HGD where the risk of overfitting is lower. In such cases, strong regularization

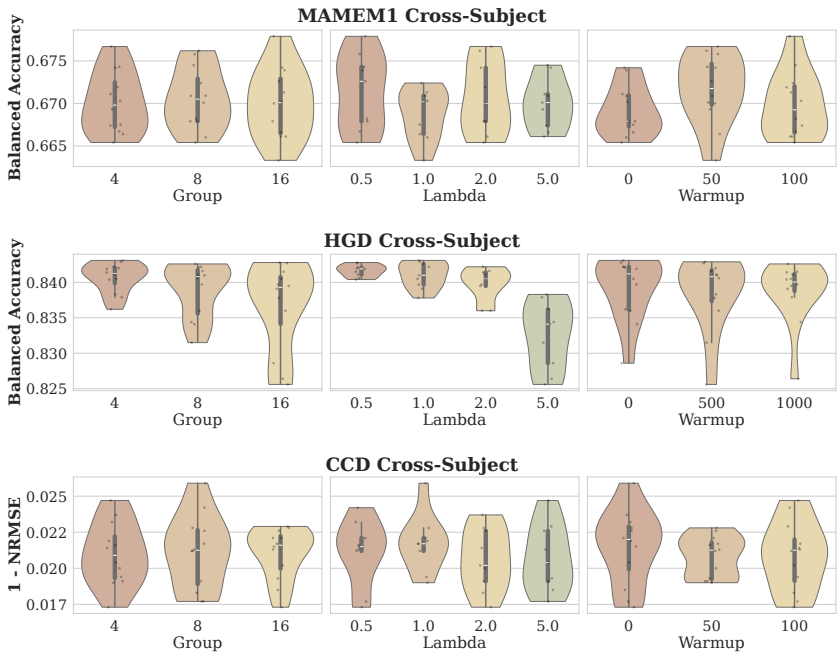

Figure 5: Sensitivity of DRO hyperparameters under cross-subject evaluation. Rows correspond to MAMEM1, HGD, and CCD-20, while columns show the group count G, reweighting strength λ, and warmup steps N_{warm}. Violin plots summarize LaBraM validation performance, with CCD-20 shown as $1 - \text{NRMSE}$ so that higher is better.

may lead to underfitting. Conversely, a moderate λ provides a favorable balance between stable cross-subject validation performance and effective reweighting across tasks.

Warmup Strategy. Finally, we analyzed the impact of the linear warmup schedule. As detailed in the methodology, λ is initialized at zero and linearly increased to mitigate gradient noise introduced by aggressive reweighting early in training. While the overall quantitative impact of warmup appears subtle across aggregated metrics, its benefits become discernible in some challenging scenarios. Specifically, in the MAMEM cross-subject setting, applying warmup consistently improved performance when larger λ values were used. These results suggest that warmup is particularly useful when stronger reweighting is applied.

Implications for Positional Adaptation. Although the preferred STR and DRO settings are task-dependent, both modules offer stable performance over the evaluated ranges of moderate regularization and reweighting. Their flexibility allows auxiliary training to be tailored to each task without changing SPR's geometric construction. Complementing this sensitivity analysis, the matched comparisons in Section 4.4 show that SPR provides additional average classification gains on four evaluated backbones when STR and DRO are already applied. Together, these findings support RADAR's combination of positional knowledge reuse and configurable auxiliary training as a practical approach to more effectively exploiting pretrained EEG representations.

G Limitations and Future Work

One focus of our evaluation is extending pretrained positional knowledge to electrode locations beyond the source layout. Since many EEG foundation models are pretrained primarily on sparse or standard montages, high-density recordings provide a particularly informative setting for examining this capability. Although high-density EEG resources continue to grow, publicly available datasets remain less abundant than their low-density counterparts, limiting the diversity of electrode

configurations available for systematic evaluation. Our native low-density and channel-reduction experiments complement this setting. As richer high-density corpora become available, future studies can examine positional transfer across a broader range of spatial configurations and investigate how more diverse source layouts enrich the positional knowledge available for reuse.

Our experiments examine the transfer of pretrained positional representations and the learning of randomly initialized alternatives during downstream adaptation. A complementary direction is to pretrain the same backbone with different positional encodings under matched data and training budgets, further distinguishing the effects of positional parameterization from its co-adaptation with the backbone. This would extend the present investigation from reusing positional knowledge to understanding how it is acquired. Our focus on existing checkpoints addresses an equally important practical setting, where repeating large-scale pretraining requires time and computational resources beyond those available to many researchers. By combining theoretical analysis with empirical comparisons, RADAR provides guidance for selecting and integrating positional priors while improving downstream performance without additional large-scale pretraining. Such reuse complements advances in pretraining itself, allowing the broader research community to derive greater value from publicly available foundation models.