DISCERN, THEN COMBINE: PROBE-GUIDED MULTI-DATASET FINE-TUNING FOR EEG FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Electroencephalography (EEG) foundation models pretrained on large-scale heterogeneous corpora provide general-purpose representations, yet the standard practice of fine-tuning them on a single target dataset, while direct and widely used, has proven suboptimal in many settings. While incorporating auxiliary datasets can provide effective multi-domain regularization, indiscriminate data mixing frequently induces negative transfer due to incompatible characteristics. Crucially, dataset compatibility in EEG is highly counter-intuitive and cannot be reliably predicted from surface-level task labels or metadata. We propose POSE, a data-centric, target-conditioned framework for multi-dataset EEG fine-tuning. POSE uses early target-only adaptation as a dynamic probe, combining endpoint discrepancies and temporal trends to select and weight auxiliary datasets. A phased training schedule then regulates auxiliary participation to balance joint learning with target-focused refinement. Comprehensive experiments show that POSE achieves higher average performance than all evaluated baselines and improves over target-only fine-tuning on most targets, while maintaining stable and robust performance across heterogeneous settings. Beyond performance improvements, our selection analyses and case studies provide empirical guidance for deciding when and how to incorporate auxiliary data for a given target and backbone. More broadly, this work advances a compatibility-aware perspective on EEG foundation-model adaptation, encouraging further exploration of how heterogeneous datasets can complement one another in target-specific learning.

1 INTRODUCTION

Electroencephalography (EEG) is a widely used noninvasive measure of brain activity with applications in brain-computer interfaces and clinical diagnostics (Roy et al., 2019; Craik et al., 2019). The growing availability of public EEG datasets offers diverse experimental paradigms and participant populations for learning generalizable neural representations (Xiong et al., 2025; Wu et al., 2025). However, differences in electrode montages, recording hardware, and participant characteristics introduce substantial heterogeneity that complicates data integration (Mellot et al., 2023). Inspired by the success of foundation models in other domains (Bommasani et al., 2021; Vaswani et al., 2017; Brown et al., 2020; OpenAI et al., 2024), EEG foundation models such as LaBraM (Jiang et al., 2024) and CBraMod (Wang et al., 2025) learn transferable representations through large-scale self-supervised pretraining (Lee et al., 2025; Yang et al., 2026). Their ability to accommodate different channel configurations allows a shared backbone with dataset-specific prediction heads to support joint fine-tuning across heterogeneous datasets (Xiong et al., 2025). This flexibility creates an opportunity to use existing EEG datasets to improve adaptation to a specific target task.

Despite this potential, multi-dataset EEG fine-tuning remains at an early stage, with initial studies primarily exploring multi-task training that directly mixes available datasets (Xiong et al., 2025). While the additional supervision provided by such training can improve target generalization through regularization, incorporating more data does not guarantee better performance. For example, in our experiments, unrestricted mixing underperforms target-only fine-tuning on average, whereas ratio-controlled mixing improves average performance but still degrades some targets. These results indicate that regulating the amount of auxiliary training is insufficient without also

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.

considering which datasets participate. Task categories and electrode configurations offer a natural basis for this choice, yet selection based on such metadata does not reliably identify beneficial sources. Effective auxiliary selection therefore requires compatibility estimates that reflect what the target task needs and what the pretrained model captures.

Such compatibility estimates depend not only on dataset characteristics but also on the representation in which their relationship is assessed. EEG signals contain diverse neural and acquisition-related information, only some of which is relevant to a given target (Biasiucci et al., 2019; Wu et al., 2022). Proximity in a generic pretrained space may therefore reflect task-irrelevant similarities rather than useful auxiliary supervision. As early target-specific fine-tuning reshapes this space, it provides an opportunity to observe how relationships between candidates and the target evolve during adaptation. Yet a distance measured at a single checkpoint describes only the current relationship, not its evolution. To capture this evolution, we use early target adaptation as a dynamic probe, combining the relative discrepancies at the end of probing with their temporal trends.

We introduce POSE (**PrO**be-guided **S**election and Schedul**E**), a target-conditioned framework for multi-dataset EEG fine-tuning built around this dynamic probe. During a short target-only training stage, POSE measures Maximum Mean Discrepancy (MMD) and Fréchet Distance (FD) between target and candidate feature distributions at successive checkpoints. It combines the final relative distances and their trends to select candidates and assign sampling weights. A phased schedule then gradually introduces, sustains, and reduces auxiliary participation to support joint learning followed. All candidates are evaluated along the same target adaptation trajectory, without training a separate model for each source or enumerating auxiliary combinations. Because EEG heterogeneity and backbone-dependent transfer make the combined contribution of auxiliary datasets difficult to predict across a combinatorial search space, our objective is to identify reliable compositions for stable and robust target adaptation rather than to recover a globally optimal subset.

Across comprehensive experiments, covering different models and diverse tasks, POSE achieves higher average performance than all evaluated baselines. It also improves over target-only fine-tuning on most targets. Ablation studies and evaluations further assess the effectiveness, stability and robustness of the framework. Beyond aggregate performance, selection patterns and case studies reveal preferences not fully explained by task or electrode metadata and show how the benefits of auxiliary training vary across models and targets. These analyses provide empirical insight into the use of heterogeneous EEG data and the conditions under which auxiliary training is beneficial.

Our main contributions are as follows.

- We introduce a shared target-adaptation probe that evaluates all candidates along one target-only trajectory, combining endpoint discrepancies and temporal trends to capture how candidate-target relationships evolve rather than relying on single-checkpoint similarity.
- We develop POSE, integrating probe-guided selection, compatibility-based weighting, and phased auxiliary participation for heterogeneous EEG fine-tuning without separate source-specific model training or changes to the target model's inference architecture.
- We conduct extensive experiments, demonstrating higher average performance than the evaluated baselines and improvements over target-only fine-tuning on most targets.
- Through cross-dataset analyses and case studies, we identify backbone-dependent auxiliary benefits and selection patterns not fully explained by task or electrode metadata, supporting our design and providing empirical guidance for heterogeneous EEG data integration.

2 Related Work

EEG Foundation Models. EEG foundation models use large-scale pretraining to learn reusable representations for downstream decoding (Yang et al., 2023; Wang et al., 2024), and their flexible input interfaces and shared representations provide a basis for multi-dataset fine-tuning.

Multi-Dataset and Multi-Source EEG Training. Cross-dataset EEG transfer and multi-source adaptation have predominantly been studied within individual task families, including motor imagery (Guetschel & Tangermann, 2023; Miao et al., 2024), emotion recognition (Cao et al., 2022; Li et al., 2024), and Parkinson's disease detection (Liu et al., 2025a). More recent work explores joint fine-tuning across heterogeneous tasks using shared foundation models (Xiong et al., 2025). These

initial efforts illustrate the potential of auxiliary supervision, but mainly rely on direct dataset mixing rather than explicit target-conditioned source selection. This motivates our focus on selecting auxiliary data for a given pretrained backbone and target task, including sources with different label spaces and electrode configurations.

Gradient-Based Multi-Task Optimization. Another line of work addresses interference among tasks by modifying how their losses or gradients contribute to shared parameters. GradNorm (Chen et al., 2018) adjusts task loss weights based on gradient magnitudes and relative training rates. PCGrad (Yu et al., 2020) modifies conflicting task gradients through pairwise projection to mitigate destructive interference. ConFIG (Liu et al., 2025b), developed for physics-informed neural networks, constructs joint update directions by seeking positive alignment with individual loss gradients. These approaches coordinate task updates, whereas POSE focuses on auxiliary data selection.

Transferability Estimation and Auxiliary Selection. Task similarity and model transferability offer related perspectives on identifying useful sources of supervision. Task2Vec (Achille et al., 2019) represents tasks using estimates of the Fisher information, enabling comparisons between tasks and selection of pretrained feature extractors. LEEP (Nguyen et al., 2020) estimates transferability from a source classifier's predictions on labeled target examples by relating source predictions to target labels. LogME (You et al., 2021) instead evaluates the maximum Bayesian evidence of target labels given features extracted by a candidate model. These methods provide a basis for transferability-driven source selection, although selecting a pretrained model and selecting auxiliary data for joint fine-tuning are distinct decisions. POSE instead evaluates candidate datasets along a shared trajectory of early target adaptation, which is a dynamic process. It combines endpoint discrepancies with temporal trends to select and weight auxiliary data, using the evolution of distributional relationships rather than a single evaluation point.

3 METHOD

3.1 PROBLEM FORMULATION

Consider a target training dataset $\mathcal{D}_T = \{(x_n^T, y_n^T)\}$ and a pool of N candidate auxiliary training datasets $\mathcal{A} = \{\mathcal{D}_1, \ldots, \mathcal{D}_N\}$. Each dataset may differ in task type, label space, electrode montage, and distributional characteristics. We adopt a multi-head architecture in which all datasets share a pretrained backbone Φ_θ but use dataset-specific heads ψ_i, so the prediction for dataset i is

$$f_i(x) = \psi_i(\Phi_\theta(x)).$$

This design decouples representation learning from task-specific decoding and allows auxiliary datasets to contribute to the backbone without requiring a shared label space. Our objective is to determine a subset $\mathcal{S} \subseteq \{1, \ldots, N\}$ with $|\mathcal{S}| \leq K$, nonnegative per-dataset sampling weights $\{w_j\}_{j\in\mathcal{S}}$ that sum to one for a nonempty subset, and an epoch-dependent schedule $\lambda(e) \geq 0$, with the aim of improving target generalization while controlling auxiliary participation.

During training, the backbone receives batches from the target dataset and, when auxiliary participation is active, from the selected auxiliary datasets. Each dataset loss is computed independently through its own head. For clarity, we express the effective per-epoch objective as

$$\mathcal{L}(\theta, \psi_T, \{\psi_j\}_{j\in\mathcal{S}}, e) = \mathcal{L}_T(\theta, \psi_T) + \lambda(e) \sum_{j\in\mathcal{S}} w_j \, \mathcal{L}_j(\theta, \psi_j), \tag{1}$$

where $\mathcal{L}_T$ and $\mathcal{L}_j$ denote the empirical losses on the target and the j-th auxiliary dataset, respectively. In practice, $\lambda(e)$ and w_j are realized by controlling the number and allocation of auxiliary batches per epoch rather than by explicitly scaling loss values. Equation (1) is therefore a conceptual summary of relative training exposure, with the batch-level updates specified below. We must decide which auxiliaries to include, how much each should contribute, and when and at what intensity they participate.

3.2 DYNAMIC TARGET PROBING

As illustrated in Figure 1, POSE first probes candidate relationships during target-only adaptation, then uses the selected data in phased joint training.

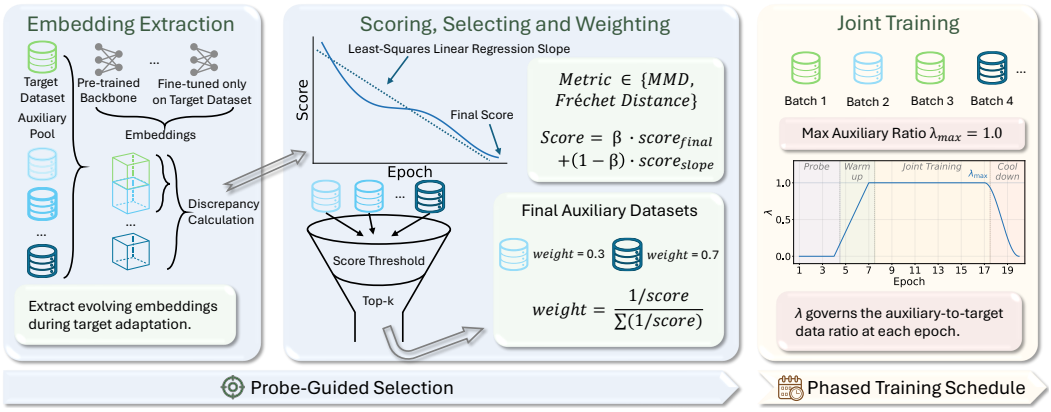

Figure 1: Overview of POSE. **Left and Middle: Probe-Guided Selection.** We extract embeddings from the early target-adapted checkpoints obtained during probe training, compute distributional discrepancies between the target and each candidate auxiliary dataset, and use the resulting trajectories to select a compatible auxiliary subset with inverse-score weighting. **Right: Phased Training Schedule.** The selected auxiliaries are then incorporated through a warm-up, joint training, and cool-down schedule that controls λ, the overall auxiliary participation ratio. Together, these components realize a data-centric and target-conditioned framework for multi-dataset EEG fine-tuning.

The first phase of our framework POSE determines which auxiliary datasets to include and how to weight them. Rather than evaluating compatibility in a static pretrained feature space, we assess how it evolves as the backbone adapts to the target task. This phase consists of a short target-only probe period, followed by compatibility scoring and selection. During probing, we fine-tune the backbone Φ_θ and the target head ψ_T on $\mathcal{D}_T$ for T_1 epochs, which are included in the total training length E.

Feature sampling. For each dataset i, we sample a probe subset $\widetilde{\mathcal{D}}_i$ from its training partition using a default budget of 1,000 examples. These subsets remain fixed across probe checkpoints so that the trajectories track representation changes on the same examples. At the end of each probe epoch $t \in \{1, \ldots, T_1\}$, we extract mean-pooled features from the target and each candidate auxiliary dataset using the current backbone, before the dataset-specific prediction heads. Denoting this representation by $z_{\theta_t}(x)$, the feature collection for dataset i is

$$Z_i^{(t)} = \{z_{\theta_t}(x) \mid x \in \widetilde{\mathcal{D}}_i\}. \tag{2}$$

All candidates are evaluated using the same target-adapted backbone, and discrepancy computation does not require auxiliary labels or auxiliary head training.

Distance estimation. Using these features, we compute two distributional distances, Maximum Mean Discrepancy (MMD) and Fréchet Distance (FD), as complementary proxies of source-target discrepancy. MMD provides a nonparametric measure of distributional discrepancy through kernel mean embeddings (Gretton et al., 2012; Long et al., 2015). We use an RBF kernel with bandwidth chosen by the median heuristic. For FD, we fit Gaussian distributions to the extracted features using their empirical means and covariance matrices, and compute the Fréchet distance between them (Heusel et al., 2017; Dowson & Landau, 1982). We use the two metrics jointly because they emphasize different aspects of distributional mismatch. For each metric $m \in \mathcal{M} = \{\mathrm{MMD}, \mathrm{FD}\}$,

$$d_j^m(t) = D_m\big(Z_T^{(t)}, Z_j^{(t)}\big), \qquad t = 1, \ldots, T_1. \tag{3}$$

The resulting metric-specific trajectories are normalized separately before scoring.

Local response interpretation. The discrepancy value describes the current relationship between feature distributions, whereas its change measures how that relationship responds to target adaptation. For fixed probe samples and a fixed discrepancy functional, write $D_j(\theta)$ for one such discrep-

ancy and $\Delta\theta$ for a backbone update. If ∇D_j is L_D-Lipschitz along the update segment, then

$$D_j(\theta + \Delta\theta) - D_j(\theta) = \langle \nabla D_j(\theta), \Delta\theta \rangle + R_j, \qquad |R_j| \leq \frac{L_D}{2} \|\Delta\theta\|^2. \tag{4}$$

For a target gradient descent step, the leading term becomes $-\eta \langle \nabla D_j(\theta), \nabla_\theta \mathcal{L}_T \rangle$. Thus, candidates with similar discrepancies can respond differently to target learning, providing information that a single distance value does not determine. This motivates combining current proximity with observed changes during probing. Appendix B.1 gives the derivation and distinguishes this local interpretation from the normalized trajectory score and downstream transfer performance. Classical domain adaptation bounds also motivate considering distributional discrepancy (Ben-David et al., 2010), with their connection to our setting discussed in Appendix B.2.

3.3 Trajectory-Based Selection and Weighting

To account for both static proximity and temporal evolution, we define a trend-aware compatibility score for each candidate dataset. For each metric, we first apply epoch-wise min-max normalization across all candidates to obtain a distance trajectory $\bar{d}_j^m(t)$ on a common $[0, 1]$ scale. From this trajectory, we extract a final value $v_j^m = \bar{d}_j^m(T_1)$ that reflects the distance at the end of probing and a trend term k_j^m given by the slope of an ordinary least-squares linear fit. To match the scale of endpoint values, we apply a second round of min-max normalization across candidates within each metric, yielding $\hat{k}_j^m \in [0, 1]$. The score of metric m and the final candidate score are

$$s_j^m = \beta v_j^m + (1 - \beta)\hat{k}_j^m, \qquad s_j = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} s_j^m, \tag{5}$$

where $\beta \in [0, 1]$ balances current proximity and relative temporal evolution. Lower scores indicate stronger estimated compatibility relative to the candidate pool. Because normalization is performed across candidates, the trend describes relative changes rather than absolute distance contraction. The normalization and slope definitions are given in Appendix A.1.

We retain candidates with $s_j < \tau$, subject to a maximum of K datasets selected in ascending score order. If no candidate satisfies this criterion, subsequent training uses only target-domain batches. Selected auxiliaries are assigned weights inversely proportional to their scores

$$w_j = \frac{(s_j + \epsilon)^{-1}}{\sum_{i \in \mathcal{S}} (s_i + \epsilon)^{-1}}, \qquad j \in \mathcal{S}, \tag{6}$$

where $\epsilon = 10^{-8}$ prevents division by zero. This weighting assigns greater auxiliary participation to candidates with lower compatibility scores. The selected subset and weights remain fixed after probing, while the total auxiliary participation changes during subsequent training.

3.4 Phased Multi-Dataset Fine-Tuning

Once the auxiliary subset $\mathcal{S}$ and weights $\{w_j\}$ are determined, training proceeds from the final probe checkpoint through three additional phases that control $\lambda(e)$, the ratio of auxiliary to target batches at epoch e. This schedule transitions smoothly from target-only optimization to joint multi-dataset training and then back to target-focused refinement.

During target-only probing ($1 \leq e \leq T_1$), $\lambda(e) = 0$. For a nonempty auxiliary subset, **warm-up** ($T_1 < e \leq T_2$) uses $\lambda(e) = \lambda_{\max}(e - T_1)/(T_2 - T_1)$ to introduce auxiliary data gradually. **Joint training** ($T_2 < e \leq T_3$) maintains $\lambda(e) = \lambda_{\max}$, providing controlled multi-domain regularization. **Cooldown** ($T_3 < e \leq E$) uses $\lambda(e) = \lambda_{\max}[1 + \cos(\pi(e - T_3)/(E - T_3))]/2$ to smoothly reduce auxiliary participation to zero and re-center the model on the target distribution.

Batch-level training. Let B_T denote the number of target batches in an epoch. The allocation to auxiliary dataset j follows $B_j(e) \approx \lambda(e) w_j B_T$, subject to discrete batch allocation. Each step processes a batch $\mathcal{B}$ from one dataset i and computes

$$\mathcal{L}_i(\mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{(x,y) \in \mathcal{B}} \ell_i\big(\psi_i(\Phi_\theta(x)), y\big), \tag{7}$$

updating the backbone and the active head while leaving other heads inactive. Algorithm 1 summarizes the full procedure, with the training configuration provided in Appendix A.2.

Table 1: Performance comparison on LaBraM. Entries report the mean over five random seeds, with standard deviations in parentheses. Average excludes CCD regression task. Superscripts in the average row indicate two-tailed paired t-tests against POSE ($^{*}p < 0.05$, $^{**}p < 0.01$).

Dataset	Target Only	Ratio Mixed	Full Mixed	Random Select	Meta Select	ConFIG	GradNorm	LogME Based	POSE
LaBraM									
BCIC-IV-2a	0.437 (.038)	0.430 (.017)	0.445 (.008)	0.423 (.048)	0.448 (.016)	**0.462** (.022)	0.431 (.021)	0.424 (.010)	0.457 (.039)
FACED	0.552 (.020)	0.562 (.016)	0.526 (.015)	0.552 (.011)	0.563 (.016)	0.532 (.016)	0.541 (.012)	0.555 (.011)	**0.564** (.014)
HGD	0.746 (.024)	0.728 (.014)	0.705 (.011)	0.717 (.043)	0.727 (.026)	0.700 (.029)	0.724 (.023)	0.745 (.047)	**0.757** (.015)
MAMEM1	0.603 (.019)	0.556 (.032)	0.579 (.044)	0.546 (.057)	0.562 (.039)	0.582 (.032)	0.594 (.038)	0.606 (.020)	**0.611** (.019)
PhysioNet-MI	0.576 (.021)	0.587 (.015)	0.562 (.011)	0.589 (.018)	**0.596** (.014)	0.549 (.013)	0.551 (.017)	0.564 (.016)	0.595 (.009)
SEED	0.568 (.037)	0.616 (.016)	0.593 (.014)	0.590 (.039)	0.611 (.025)	0.551 (.024)	0.616 (.027)	0.602 (.050)	**0.618** (.019)
SEED-IV	0.338 (.024)	0.350 (.033)	0.354 (.030)	0.358 (.017)	0.350 (.025)	0.342 (.032)	0.303 (.022)	0.339 (.015)	**0.359** (.011)
SEED-V	0.270 (.016)	0.285 (.020)	0.284 (.032)	**0.291** (.023)	0.283 (.018)	**0.291** (.017)	0.284 (.027)	0.254 (.006)	**0.291** (.038)
SHU-MI	0.557 (.020)	0.569 (.017)	0.554 (.014)	0.561 (.012)	0.556 (.018)	0.543 (.018)	0.555 (.017)	0.562 (.010)	**0.571** (.007)
TUAB	0.806 (.011)	0.797 (.005)	**0.807** (.009)	0.802 (.018)	0.802 (.007)	0.798 (.011)	0.790 (.009)	0.800 (.008)	0.801 (.011)
TUEV	**0.537** (.044)	0.529 (.025)	0.488 (.056)	0.524 (.048)	0.532 (.045)	0.493 (.051)	0.525 (.040)	0.523 (.003)	0.532 (.044)
CCD (↓)	0.996 (.012)	**0.983** (.015)	1.046 (.005)	1.013 (.031)	0.984 (.012)	1.156 (.016)	1.149 (.014)	1.038 (.006)	0.988 (.016)
Average	0.545^{**}	0.546^{*}	0.536^{**}	0.541^{*}	0.548^{*}	0.531^{**}	0.538^{**}	0.543^{**}	**0.560**

4 Experiments

4.1 Experiment Setups

We evaluate our framework on two widely adopted EEG foundation models, LaBraM (Jiang et al., 2024) and CBraMod (Wang et al., 2025). Our evaluation spans 12 EEG datasets that vary in electrode configuration from 22 to 256 channels and cover motor imagery, emotion recognition, steady-state visually evoked potentials, clinical EEG analysis, and behavioral prediction. We report balanced accuracy for the 11 classification tasks and normalized root mean square error (NRMSE) for the CCD regression task. Higher balanced accuracy and lower NRMSE indicate better performance.

Baselines. We compare POSE against eight baselines covering data mixing, heuristic selection, multi-task optimization, and transferability-based selection. Target-only performs standard single-dataset fine-tuning without auxiliary data. Ratio-mixed includes all auxiliary datasets with uniform weights under a fixed participation ratio $\lambda_{\max}$. Full-mixed mixes all datasets in each epoch without ratio control. Random-select randomly selects one to three auxiliary datasets with uniform weights at a fixed ratio $\lambda_{\max}$. Meta-select groups auxiliary datasets based on task paradigm or electrode layout and applies uniform weighting within the selected group at a fixed ratio $\lambda_{\max}$. ConFIG (Liu et al., 2025b) coordinates task gradients through conflict-free update directions, while GradNorm (Chen et al., 2018) adaptively balances task losses based on gradient magnitudes and relative training rates. Both are adapted to the heterogeneous-label setting using a shared backbone and dataset-specific heads. LogME-based selection uses the LogME transferability criterion (You et al., 2021) to rank candidate auxiliary datasets for joint fine-tuning.

Evaluation protocol. All datasets are partitioned using a cross-subject protocol, with the same training, validation, and test splits used across methods. The main comparison uses one fixed split per dataset and reports the mean and standard deviation over five random seeds, 42 to 46. Average performance is computed over the classification datasets, excluding CCD. Table 1 presents the LaBraM results, and Table 5 in Appendix D.1 provides the complete results for both backbones. Dataset and implementation details are provided in Appendix C.

4.2 Main Results

Overall performance and robustness. Table 1 and Appendix Table 5 show that POSE achieves the highest average classification performance on both backbones. The aggregate improvements over all eight baselines are statistically significant under two-tailed paired t-tests ($p < 0.05$). At the dataset level, POSE attains the best mean on seven of the twelve tasks for LaBraM and eight for CBraMod. This breadth of competitive results supports the robustness of the framework across

heterogeneous EEG tasks and backbone architectures. Results aggregated over three subject partitions are reported in Appendix D.3, providing additional evidence of overall robustness. POSE also achieves the highest average performance in a supplementary six-dataset evaluation on REVE (Appendix D.2), providing further evidence that its benefits extend beyond the two main backbones.

Gains over target-only fine-tuning. Compared with standard target-only fine-tuning, POSE increases average balanced accuracy from 0.545 to 0.560 on LaBraM and from 0.528 to 0.550 on CBraMod. Across evaluated tasks, POSE improves target performance on 10 of the 12 datasets with LaBraM and nine with CBraMod. Even where gains are absent, performance remains comparable to Target-only, with no significant degradation. Together, these results highlight the overall robustness and stability of POSE, characteristics that are particularly important in practice and support reliable EEG foundation model adaptation.

Limitations of naive and heuristic selection. Full-mixed reduces average classification performance relative to Target-only on both backbones, showing that unrestricted data integration can offset the benefits of additional supervision. Random-select, included as a control for arbitrary auxiliary composition, also fails to provide consistent gains across the two backbones. In contrast, Ratio-mixed and Meta-select modestly improve average performance, but these gains mask negative transfer on multiple targets. For example, on MAMEM1 with LaBraM, their balanced accuracies fall to 0.556 and 0.562 from the Target-only value of 0.603. In contrast, the same Meta-select grouping improves balanced accuracy on MAMEM1 from 0.487 to 0.542 with CBraMod, highlighting the backbone dependence of auxiliary utility. Thus, controlling auxiliary volume or selecting sources by coarse metadata alone does not provide sufficiently reliable integration across heterogeneous targets. POSE instead combines target-conditioned compatibility assessment with controlled auxiliary participation, achieving more consistent improvements and smaller observed performance losses than these mixing and heuristic selection strategies.

Comparison with optimization and transferability baselines. POSE also achieves higher average classification performance than ConFIG and GradNorm on both backbones, supporting explicit control of auxiliary data composition beyond gradient coordination or task loss balancing alone. Against LogME-based selection, POSE improves average balanced accuracy by 1.7 percentage points on LaBraM and 3.3 percentage points on CBraMod. This comparison supports assessing auxiliary datasets for joint target adaptation rather than relying solely on a source model's transferability as the selection criterion. Together, these results favor the target-conditioned design of POSE, which evaluates evolving distributional relationships and uses them to guide auxiliary selection, weighting, and phased participation.

4.3 Ablation Study

We examine the contribution of each component and the choice of discrepancy metric, with LaBraM results in Table 2 and results for both backbones in Appendix D.4. The main ablation focuses on eight classification datasets where POSE improves over target-only fine-tuning on both backbones, making differences between variants easier to examine. Results for the remaining targets are reported in Appendix D.5. No Probe selects auxiliaries once from the pretrained checkpoint without target-adapted probing. No Warm-up and No Cool-down remove the gradual introduction and final reduction of auxiliary participation, respectively, while Uniform Weight replaces score-based sampling weights with equal weights. FD-only and MMD-only use a single discrepancy metric within the same selection procedure. All variants retain the remaining hyperparameters of the complete framework.

Removing target-adapted probing produces the largest average degradation among the component ablations on both backbones, supporting the importance of evaluating auxiliary data in a representation shaped by the target task. Complementary probe analyses show that adding temporal trends changes the top-three candidate set in over 80% of cases on both backbones relative to endpoint-only selection (Appendix D.9). Warm-up, cooldown, and compatibility-based weighting also improve average performance, supporting the contribution of each component. Using FD or MMD alone yields lower average performance than their combination, supporting the use of complementary discrepancy signals. Together, these ablations support the integration of target-conditioned selection with controlled auxiliary participation.

Table 2: Component and metric ablations on LaBraM. Entries report mean balanced accuracy over five random seeds, with standard deviations in parentheses. POSE is shared by the component comparison (left) and the metric comparison (right). Boldface marks the best mean in the component comparison. Underlining marks the best mean in the metric comparison.

Dataset	No Probe	No Warm-up	No Cool-down	Uniform Weight	POSE	FD-only	MMD-only
LaBraM							
BCIC-IV-2a	0.422 (.026)	0.427 (.033)	0.431 (.030)	0.432 (.036)	**0.457** (.039)	0.433 (.019)	0.419 (.024)
FACED	0.540 (.029)	0.534 (.022)	0.536 (.023)	0.543 (.029)	**0.564** (.014)	0.547 (.019)	0.557 (.016)
MAMEM1	0.554 (.040)	0.576 (.041)	0.570 (.030)	0.548 (.030)	**0.611** (.019)	0.604 (.044)	0.543 (.055)
PhysioNet-MI	0.587 (.011)	0.573 (.022)	0.574 (.014)	0.582 (.022)	**0.595** (.009)	0.584 (.013)	0.589 (.015)
SEED	0.612 (.009)	0.613 (.016)	**0.619** (.022)	0.606 (.022)	0.618 (.019)	0.596 (.032)	0.588 (.031)
SEED-IV	0.332 (.031)	0.346 (.017)	0.358 (.019)	0.345 (.023)	**0.359** (.011)	0.334 (.028)	0.338 (.026)
SEED-V	0.264 (.025)	0.275 (.029)	0.278 (.010)	0.272 (.051)	**0.291** (.038)	0.287 (.017)	0.282 (.050)
SHU-MI	0.557 (.019)	0.566 (.013)	**0.574** (.009)	0.567 (.007)	0.571 (.007)	0.553 (.017)	0.569 (.014)
Average	0.484	0.489	0.493	0.487	**0.508**	0.492	0.485

4.4 Cross-Dataset Affinity Matrix

To better understand the selection behavior of our method, we construct a cross-dataset affinity matrix that aggregates the normalized auxiliary weights determined at the end of probing. Each entry averages the final weight over five random seeds, assigning zero to runs in which the candidate is not selected. Unlike a simple count-based frequency matrix, this weighted formulation reflects both how often a dataset is selected and how much weight it receives when selected. Figure 3 in Appendix D.6 shows the resulting selection patterns for both backbones.

The observed affinities are only partially aligned with metadata-based intuition. Some expected relationships emerge, such as the mutual affinity between SEED and SEED-IV across both backbones. However, SEED-V shows weaker affinity to these datasets despite belonging to the same emotion-recognition family. The relationship between TUAB and TUEV is also asymmetric, with TUAB receiving greater aggregate weight as an auxiliary for TUEV than in the reverse direction. Cross-paradigm selections are also present, including SHU-MI as an auxiliary for SEED-V on LaBraM and bidirectional selection between CCD and BCIC-IV-2a on CBraMod. These patterns indicate that compatibility is not determined solely by task labels or electrode layouts, but also depends on finer-grained dataset characteristics and the backbone representation. Together with the Meta-select comparison, they support evaluating auxiliary data through target-conditioned representations rather than relying on task labels or electrode layouts alone.

4.5 Case Study

We next examine two representative cases to characterize the benefits and limitations of auxiliary training. Figure 4 in Appendix D.7 presents the two cases.

A successful case on MAMEM1. For illustration, we show one representative run with a single random seed. MAMEM1 is a small dataset with roughly 1,000 samples in total. In the illustrated run, the target-only model reaches a high validation peak that does not translate into similarly strong test performance on held-out subjects. POSE instead achieves stronger test performance at its validation-selected checkpoint, consistent with an auxiliary regularization benefit in this setting. The five-seed comparison supports this observation, with balanced accuracy increasing from 0.487 for Target-only to 0.562 for POSE, compared with 0.542 for Meta-select. This case illustrates the value of compatible auxiliary supervision for improving generalization beyond the available target training subjects.

A limited-gain case on TUAB. The target-only and POSE validation trajectories are already stable and closely aligned in the illustrated TUAB run, with little difference in test performance. The five-seed means are also identical at the reported precision, with both methods achieving 0.809 balanced accuracy. Within the evaluated setting, target supervision appears sufficient to reach a competitive solution, leaving limited room for improvement from the available auxiliary pool.

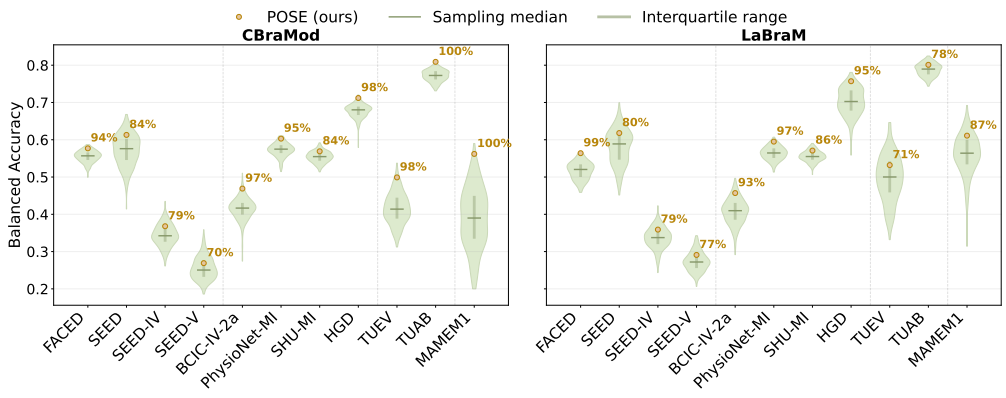

Figure 2: Sampling distribution analysis for CBraMod (left) and LaBraM (right) across 11 datasets. Each violin shows the distribution of test performance obtained by randomly sampled auxiliary dataset subsets. CCD is omitted (NRMSE), and achieves the 98th percentile on both models.

In practice, we recommend comparing probe-guided auxiliary training with a target-only baseline for the chosen target and backbone. Both settings should be evaluated on the same held-out validation subjects, with the comparison repeated across subject partitions where feasible. The decision to incorporate auxiliary data should be based on consistent gains that are meaningful for the intended application, rather than dataset size or a single validation peak. This focused comparison avoids the need to exhaustively evaluate auxiliary combinations. Together, these analyses and recommendations provide empirical guidance for deciding when and how to incorporate auxiliary data into EEG foundation-model adaptation.

4.6 Comparison with Naive Dataset Mixing

We also compare our selected compositions against a broad space of naive dataset mixtures. Since there are $2^{12} - 1 = 4{,}095$ possible non-empty combinations, exhaustive evaluation is infeasible. This number would grow exponentially as more datasets are included. We therefore sample 500 combinations per backbone and use them to approximate the performance distribution of naive mixing. Our selected compositions consistently rank near the top of this distribution (Figure 2), with an average percentile of around 90%. More details are provided in Appendix D.8.

Auxiliary datasets can interact during joint training, making their combined contribution an important consideration in practical data integration. We emphasize that our goal is not to recover a globally optimal subset, since EEG heterogeneity and differing backbone inductive biases make the effects of joint training context-dependent and difficult to predict across a combinatorial search space. Instead, we focus on reliably identifying high-quality auxiliary compositions that support stable and robust target adaptation without requiring combinatorial exploration. The competitive results among sampled mixtures, together with the repeated-seed and subject-partition evaluations, provide empirical support for this practical objective across the evaluated settings.

5 Conclusion

We presented POSE, a target-conditioned framework that uses early adaptation dynamics to guide auxiliary selection, weighting, and phased participation in multi-dataset EEG fine-tuning. Experiments across heterogeneous EEG tasks and multiple backbones demonstrate the stability and robustness of POSE, which achieves higher average performance than the evaluated baselines and improves over target-only fine-tuning on most targets. Beyond aggregate performance, our analyses show that auxiliary preferences are not fully explained by task or electrode metadata and that the benefits of auxiliary training vary across targets and backbones. Together, the framework and empirical analyses provide practical guidance for using heterogeneous auxiliary data without exhaustive combination search. Ultimately, we hope that this work encourages a deeper understanding of dataset compatibility, opening avenues for harnessing the complementary strengths of heterogeneous EEG data in target-conditioned foundation-model adaptation.

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 additional algorithmic details in Appendices B and A, respectively. Dataset descriptions and preprocessing procedures are provided in Appendix C.1. Evaluation metrics are specified in Section 4.1, while data partitions, training hyperparameters, and implementation details are documented in Appendix C.3. The additional subject and release partitions used in the three-fold evaluation are described in Appendix D.3. 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

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona. Task2Vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6430–6439, 2019.

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.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

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, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. 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 P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher

R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram 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 A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *ArXiv*, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020.

Jiangsheng Cao, Xueqin He, Chenhui Yang, Sifang Chen, Zhangyu Li, and Zhanxiang Wang. Multi-source and multi-representation adaptation for cross-domain electroencephalography emotion recognition. *Frontiers in Psychology*, 12:809459, 2022.

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.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pp. 794–803. PMLR, 2018.

Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. Deep learning for electroencephalogram (EEG) classification tasks: a review. *Journal of neural engineering*, 16(3):031001, 2019.

David C Dowson and B. V. Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. Differential entropy feature for EEG-based emotion classification. In *6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 81–84. IEEE, 2013.

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.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

Pierre Guetschel and Michael Tangermann. Transfer learning between motor imagery datasets using deep learning–validation of framework and comparison of datasets. *arXiv preprint arXiv:2311.16109*, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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.

Na Lee, Konstantinos Barmpas, Yannis Panagakis, Dimitrios Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. Are large brainwave foundation models capable yet? insights from fine-tuning. In *Forty-second International Conference on Machine Learning*, 2025.

Wei Li, Siyi Wang, Shitong Shao, and Kaizhu Huang. Distillation-based domain generalization for cross-dataset EEG-based emotion recognition. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 9(3):2474–2490, 2024.

Hongjun Liu, Chao Yao, Yalan Zhang, Xiaokun Wang, and Xiaojuan Ban. Spatial imputation drives cross-domain alignment for EEG classification. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 8586–8595, 2025a.

Qiang Liu, Mengyu Chu, and Nils Thuerey. ConFIG: Towards conflict-free training of physics informed neural networks. In *International Conference on Learning Representations*, volume 2025, pp. 59531–59566, 2025b.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 2021.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105. PMLR, 2015.

Jun Ma, Banghua Yang, Wenzheng Qiu, Yunzhe Li, Shouwei Gao, and Xinxing Xia. A large EEG dataset for studying cross-session variability in motor imagery brain-computer interface. *Scientific Data*, 9(1):531, 2022.

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.

Minmin Miao, Zhong Yang, Zhenzhen Sheng, Baoguo Xu, Wenbin Zhang, and Xinmin Cheng. Multi-source deep domain adaptation ensemble framework for cross-dataset motor imagery EEG transfer learning. *Physiological Measurement*, 45(5):055024, 2024.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In *International conference on machine learning*, pp. 7294–7305. PMLR, 2020.

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.

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.

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.

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.

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.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 39249–39280. Curran Associates, Inc., 2024. doi: 10.52202/079017-1239.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBramod: A criss-cross brain foundation model for EEG decoding. In *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. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 78240–78260. Curran Associates, Inc., 2023.

Liuyin Yang, Qiang Sun, Ang Li, and Marc M. Van Hulle. Are EEG foundation models worth it? comparative evaluation with traditional decoders in diverse BCI tasks. In *The Fourteenth International Conference on Learning Representations*, 2026.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical assessment of pre-trained models for transfer learning. In *International conference on machine learning*, pp. 12133–12143. PMLR, 2021.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

W. Zheng, W. Liu, Y. Lu, B. Lu, and A. Cichocki. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics*, pp. 1–13, 2018. ISSN 2168-2267. doi: 10.1109/TCYB.2018.2797176.

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

A Additional Method Details

A.1 Normalization and Trajectory Scoring

For nonconstant candidate values $a_1, \ldots, a_N$, define $a_{\min} = \min_i a_i$ and $a_{\max} = \max_i a_i$. We use the min-max operator

$$\mathrm{MM}_j(a) = \frac{a_j - a_{\min}}{a_{\max} - a_{\min}}. \tag{8}$$

For each metric m, we first apply epoch-wise min-max normalization across all candidates to place the values on a common $[0, 1]$ scale, yielding $\bar{d}_j^m(t)$. With $T_1 \geq 2$, let $\bar{t} = (T_1 + 1)/2$ and $\bar{d}_j^m = T_1^{-1} \sum_{t=1}^{T_1} \bar{d}_j^m(t)$. The endpoint and fitted slope are

$$v_j^m = \bar{d}_j^m(T_1), \qquad k_j^m = \frac{\sum_{t=1}^{T_1}(t - \bar{t})(\bar{d}_j^m(t) - \bar{d}_j^m)}{\sum_{t=1}^{T_1}(t - \bar{t})^2}. \tag{9}$$

To match the scale of endpoint values, we apply a second round of min-max normalization across candidates within each metric, yielding $\hat{k}_j^m \in [0, 1]$. The per-metric scores and their average are then computed using Equation (5). This second normalization preserves the ordering of slopes while placing both score components on a common scale. Setting $\beta = 1$ gives endpoint-only scoring in the target-adapted space without using the trend term. An empty selected subset skips the weight computation in Equation (6).

A.2 Training Configuration and Procedure

We use $E = 20$ total epochs with $T_1 = 4$, $T_2 = 7$, and $T_3 = 17$, so the probe, auxiliary warm-up, joint training, and cooldown phases last 4, 3, 10, and 3 epochs, respectively. The default settings are $K = 3$ and $\beta = 0.7$. POSE uses AdamW with batch size 32, fixed weight decay 10^{-2}, and gradient clipping at norm 1.0. The learning rate follows a cosine schedule with five warm-up epochs and a minimum value of 10^{-6}. This learning-rate warm-up is separate from the auxiliary participation warm-up in Section 3.4. For POSE, the base learning rate is fixed for each backbone, while $\lambda_{\max}$ and τ are selected using target validation performance, following the protocol in Appendix C.3.4. No test data are used for parameter or checkpoint selection.

Equation (1) describes relative exposure rather than simultaneous loss aggregation at each update. Actual updates use the single-dataset batch loss in Equation (7). When no auxiliary is selected, training uses only target batches, but the existing implementation retains a learning-rate horizon initialized before selection and the random state affected by probing. Such runs therefore need not be numerically identical to an independently executed target-only baseline.

A.3 Summary

Algorithm 1 summarizes the procedure. Feature extraction uses the fixed training subsets defined in Section 3.2, and all training stages continue the current model state without reinitialization. Each training instruction uses the batch-level loss in Equation (7), with auxiliary batch allocation skipped when $\mathcal{S} = \varnothing$. The target validation metric is evaluated after every training epoch, and the best checkpoint is retained for the final output.

B Theoretical Motivation

B.1 Local Response to Target Adaptation

Consider a discrepancy $D_j(\theta)$ evaluated on fixed probe samples with a fixed measurement rule. Assume that it is differentiable on the parameter update segment and that its gradient is L_D-Lipschitz there. The fundamental theorem of calculus gives

$$\begin{aligned} D_j(\theta + \Delta\theta) - D_j(\theta) &= \int_0^1 \langle \nabla D_j(\theta + u\Delta\theta), \Delta\theta \rangle \, du \\ &= \langle \nabla D_j(\theta), \Delta\theta \rangle + R_j, \end{aligned} \tag{10}$$

Algorithm 1 Probe-Guided Multi-Dataset Fine-Tuning

Require: Target training data $\mathcal{D}_T$, candidates $\mathcal{A} = \{\mathcal{D}_1, \ldots, \mathcal{D}_N\}$, target validation data, pretrained backbone Φ_θ, heads $\{\psi_i\}$
Require: Phase boundaries T_1, T_2, T_3, E, hyperparameters $\lambda_{\max}, \beta, \tau, K$, stabilizer ϵ

Probe-Guided Selection
for $e = 1$ **to** T_1 **do**
  Train Φ_θ, ψ_T on $\mathcal{D}_T$
  Extract target features $Z_T^{(e)}$
  for each $\mathcal{D}_j \in \mathcal{A}$ **do**
    Extract candidate features $Z_j^{(e)}$
    $d_j^m(e) \leftarrow D_m(Z_T^{(e)}, Z_j^{(e)})$ for each $m \in \mathcal{M}$
  end for
end for
for each metric $m \in \mathcal{M}$ **do**
  for each probe epoch $t = 1, \ldots, T_1$ **do**
    $\bar{d}_j^m(t) \leftarrow \mathrm{MM}_j(\{d_i^m(t)\}_{i=1}^N)$ for all candidates j
  end for
  $v_j^m \leftarrow \bar{d}_j^m(T_1)$ for all candidates j
  $k_j^m \leftarrow \text{OLS-slope}(\bar{d}_j^m(1), \ldots, \bar{d}_j^m(T_1))$ for all candidates j
  $\hat{k}_j^m \leftarrow \mathrm{MM}_j(\{k_i^m\}_{i=1}^N)$ for all candidates j
  $s_j^m \leftarrow \beta v_j^m + (1-\beta)\hat{k}_j^m$ for all candidates j
end for
$s_j \leftarrow |\mathcal{M}|^{-1} \sum_{m \in \mathcal{M}} s_j^m$ for all candidates j
$\mathcal{S} \leftarrow \{j \mid s_j < \tau\}$, truncated to the K lowest scores
if $\mathcal{S} \neq \varnothing$ **then**
  $w_j \leftarrow (s_j + \epsilon)^{-1} / \sum_{i \in \mathcal{S}} (s_i + \epsilon)^{-1}$ for each $j \in \mathcal{S}$
end if
Warm-up
for $e = T_1 + 1$ **to** T_2 **do**
  $\lambda(e) \leftarrow \lambda_{\max}(e - T_1)/(T_2 - T_1)$
  Train on $\mathcal{D}_T$ with auxiliary batches from $\mathcal{S}$ at ratio $\lambda(e)$ and weights $\{w_j\}_{j \in \mathcal{S}}$
end for
Joint Training
for $e = T_2 + 1$ **to** T_3 **do**
  $\lambda(e) \leftarrow \lambda_{\max}$
  Train on $\mathcal{D}_T$ with auxiliary batches from $\mathcal{S}$ at ratio $\lambda(e)$ and weights $\{w_j\}_{j \in \mathcal{S}}$
end for
Cool-down
for $e = T_3 + 1$ **to** E **do**
  $\lambda(e) \leftarrow \frac{\lambda_{\max}}{2}\big(1 + \cos(\pi(e - T_3)/(E - T_3))\big)$
  Train on $\mathcal{D}_T$ with auxiliary batches from $\mathcal{S}$ at ratio $\lambda(e)$ and weights $\{w_j\}_{j \in \mathcal{S}}$
end for
return Φ_θ, ψ_T from the best target validation checkpoint

where

$$R_j = \int_0^1 \langle \nabla D_j(\theta + u\Delta\theta) - \nabla D_j(\theta), \Delta\theta \rangle \, du,$$
$$|R_j| \le \int_0^1 L_D u \|\Delta\theta\|^2 \, du = \frac{L_D}{2}\|\Delta\theta\|^2. \tag{11}$$

This establishes Equation (4) for any update satisfying the local assumptions. The gradient descent specialization follows by substituting $\Delta\theta = -\eta \nabla_\theta \mathcal{L}_T$. For AdamW, the general expression instead uses the actual backbone update, including adaptive scaling, momentum, and weight decay. Across multiple updates, discrepancy changes accumulate as the sum of these local response terms and their remainders.

The discrepancy value and its response to target updates describe different aspects of the relationship between two feature distributions. In particular, equal discrepancy values do not constrain their directional derivatives to be equal. This supports observing changes during adaptation in addition to the endpoint value. The argument concerns a fixed discrepancy functional, whereas the POSE slope summarizes epoch-level changes after candidate-wise normalization. Changes in measurement settings, such as re-estimated kernel bandwidths, can also contribute to the observed trajectory. The normalized slope is therefore a summary of relative evolution, not an exact estimate of the unnormalized directional derivative.

Connection to transfer utility. Let $g_T = \nabla_\theta \mathcal{L}_T$ and $g_j = \nabla_\theta \mathcal{L}_j$ denote gradients on shared parameters, holding the task-specific heads fixed for a local comparison. For a target loss with an L_T-Lipschitz gradient along an auxiliary update segment,

$$\mathcal{L}_T(\theta - \eta g_j) - \mathcal{L}_T(\theta) = -\eta \langle g_T, g_j \rangle + R_T, \qquad |R_T| \le \frac{L_T \eta^2}{2} \|g_j\|^2. \tag{12}$$

The first-order effect of auxiliary supervision depends on task gradients at the current model state. This is distinct from the response of a feature discrepancy to target updates, which involves ∇D_j rather than g_j. Accordingly, the local response analysis motivates the trajectory signal but does not establish that decreasing discrepancy guarantees positive transfer or derive the particular score mixture in Equation (5).

B.2 Connection to Domain Adaptation Bounds

Setup. For the classical binary classification setting, let all domains share an input space, a label space, and a hypothesis class $\mathcal{H}$. Write $\epsilon_i(h)$ and $\epsilon_T(h)$ for source and target expected classification errors, and P_i and P_T for their input marginals. Define the ideal joint error $\xi_i = \inf_{h \in \mathcal{H}} \big(\epsilon_i(h) + \epsilon_T(h)\big)$. For fixed weights $\alpha_i \ge 0$ with $\sum_{i=1}^N \alpha_i = 1$, define the weighted source error

$$\epsilon_\alpha(h) = \sum_{i=1}^N \alpha_i \epsilon_i(h). \tag{13}$$

Step 1. Pairwise source-target bound. The standard domain adaptation inequality (Ben-David et al., 2010) applies in both directions. Since the divergence and ideal joint error are symmetric in the source and target, for any $h \in \mathcal{H}$ we have

$$\begin{aligned} \epsilon_T(h) &\le \epsilon_i(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(P_i, P_T) + \xi_i, \\ \epsilon_i(h) &\le \epsilon_T(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(P_i, P_T) + \xi_i. \end{aligned} \tag{14}$$

Combining the two inequalities yields the absolute error-gap bound

$$|\epsilon_i(h) - \epsilon_T(h)| \le \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(P_i, P_T) + \xi_i. \tag{15}$$

Step 2. Aggregation over sources. The normalization of the weights gives $\epsilon_T(h) = \sum_{i=1}^N \alpha_i \epsilon_T(h)$. Using this identity, the triangle inequality, and Equation (15), we obtain

$$\begin{aligned} |\epsilon_\alpha(h) - \epsilon_T(h)| &= \left| \sum_{i=1}^N \alpha_i \big(\epsilon_i(h) - \epsilon_T(h)\big) \right| \\ &\le \sum_{i=1}^N \alpha_i \left| \epsilon_i(h) - \epsilon_T(h) \right| \\ &\le \frac{1}{2} \sum_{i=1}^N \alpha_i d_{\mathcal{H}\Delta\mathcal{H}}(P_i, P_T) + \sum_{i=1}^N \alpha_i \xi_i. \end{aligned} \tag{16}$$

Step 3. Upper bound on target error. The target error is at most the weighted source error plus their absolute difference. Substituting Equation (16) therefore gives

$$\begin{aligned}\epsilon_T(h) &\le \epsilon_\alpha(h) + |\epsilon_\alpha(h) - \epsilon_T(h)| \\ &\le \sum_{i=1}^{N} \alpha_i \epsilon_i(h) + \frac{1}{2} \sum_{i=1}^{N} \alpha_i d_{\mathcal{H}\Delta\mathcal{H}}(P_i, P_T) + \sum_{i=1}^{N} \alpha_i \xi_i.\end{aligned} \tag{17}$$

Step 4. Constant term for fixed weights. Writing

$$C(\alpha) = \sum_{i=1}^{N} \alpha_i \xi_i, \tag{18}$$

the bound can equivalently be expressed as

$$\epsilon_T(h) \le \sum_{i=1}^{N} \alpha_i \epsilon_i(h) + \frac{1}{2} \sum_{i=1}^{N} \alpha_i d_{\mathcal{H}\Delta\mathcal{H}}(P_i, P_T) + C(\alpha). \tag{19}$$

The term $C(\alpha)$ is independent of h for fixed mixture weights, but changes with the weights and therefore cannot be discarded when reasoning about source selection.

Interpretation and scope. This derivation gives a population-level inequality, rather than a finite-sample guarantee for estimated discrepancies. The bound identifies distributional discrepancy as one factor in transfer alongside source error and joint task compatibility. In particular, the bound does not prescribe inverse-discrepancy weighting or justify neglecting source risk on the basis of low training loss. POSE uses dataset-specific heads across heterogeneous label spaces and measures empirical feature discrepancies along a changing representation. We therefore use this classical result as motivation for considering distributional relationships, rather than as a direct guarantee for the POSE training procedure or an identification of MMD and FD with $d_{\mathcal{H}\Delta\mathcal{H}}$.

C Experimental Details

C.1 Dataset Details

Table 3 summarizes the key properties of all datasets used in this work, including task type, number of channels, sample duration, and dataset scale. These datasets span multiple paradigms such as motor imagery, SSVEP, emotion recognition, clinical EEG analysis, and behavioral prediction, covering a wide range of recording setups and subject populations.

The motor imagery datasets are BCIC-IV-2a (Tangermann et al., 2012), HGD (Schirrmeister et al., 2017), PhysioNet-MI (Schalk et al., 2004), and SHU-MI (Ma et al., 2022). The emotion-recognition datasets are FACED (Chen et al., 2023), SEED (Duan et al., 2013; Zheng & Lu, 2015), SEED-IV (Zheng et al., 2018), and SEED-V (Liu et al., 2021). We use MAMEM1 (Oikonomou et al., 2016; Goldberger et al., 2000) for SSVEP classification and CCD (Alexander et al., 2017; Shirazi et al., 2024) for behavioral prediction. For clinical EEG analysis, we use TUAB for abnormal detection and TUEV for event type classification (Obeid & Picone, 2016).

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 and ensure consistency across datasets. All datasets are band-pass filtered as listed in Table 3, and are resampled to 200 Hz.

Dataset subsampling. As shown in Table 3, TUAB, TUEV, and CCD are substantially larger than the other datasets. To avoid imbalance in multi-dataset training, we construct fixed subsets for these datasets and use them consistently across methods and folds. For TUEV, we randomly sample approximately 20% of data from each subject, resulting in 22,302 samples. For TUAB, we sample approximately 5% per subject, yielding 19,538 samples. For CCD, we use the official mini-releases with 20 subjects per release across all 11 releases, resulting in 13,304 samples.

Table 3: Statistical properties of the datasets. SSVEP denotes steady-state visually evoked potentials. The number of classes is not applicable to CCD, which is a regression task.

Dataset	Task	Classes	Bandpass	Channels	Window Length	Subjects	Total Samples
BCIC-IV-2a	Motor Imagery	4	0.5 to 45 Hz	22	4s	9	2,592
CCD	Behavior Prediction	N/A	0.5 to 50 Hz	128	2s	1,964	116,086
FACED	Emotion Recognition	9	0.05 to 47 Hz	30	10s	123	10,332
HGD	Motor Imagery	4	0.5 to 75 Hz	128	4s	14	26,968
MAMEM1	SSVEP	5	5 to 45 Hz	256	3s	11	1,081
PhysioNet-MI	Motor Imagery	4	0.5 to 45 Hz	64	3s	109	19,022
SEED	Emotion Recognition	3	0.1 to 75 Hz	62	4s	15	37,890
SEED-IV	Emotion Recognition	4	0.1 to 75 Hz	62	4s	15	37,575
SEED-V	Emotion Recognition	5	1 to 75 Hz	62	4s	16	29,168
SHU-MI	Motor Imagery	2	0.1 to 75 Hz	32	4s	25	11,988
TUAB	Abnormal Detection	2	0.1 to 75 Hz	23	10s	2,328	409,083
TUEV	Event Type Classification	6	0.1 to 75 Hz	23	5s	370	112,237

CCD regression metric. 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

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n - y_n)^2}}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - \bar{y})^2}}, \qquad \bar{y} = \frac{1}{N}\sum_{n=1}^{N} y_n. \tag{20}$$

C.2 Pretrained Backbones and Data Coverage

Our main experiments use pretrained LaBraM and CBraMod backbones, both learned through self-supervised objectives without downstream task labels. According to the original pretraining descriptions, LaBraM includes PhysioNet-MI, SEED, and SEED-IV, while TUAB and TUEV are reported to be excluded despite sharing the broader TUH data source (Jiang et al., 2024). CBraMod is pretrained on TUEG, which includes TUAB and TUEV (Wang et al., 2025).

Our evaluation compares fine-tuning strategies from a common pretrained initialization within each backbone, using identical downstream training, validation, and test partitions. Probing and gradient updates use training data only, while tunable hyperparameters and checkpoints are selected using target validation performance.

Among targets not listed in the corresponding reported pretraining corpora, the existing results in Table 5 show improvements over Target-only on seven of nine datasets for LaBraM and nine of ten for CBraMod, using balanced accuracy for classification and NRMSE for CCD. The gains are therefore not confined to targets with reported pretraining overlap. As complementary evidence, the original CBraMod study reports balanced-accuracy decreases of 0.40 and 0.12 percentage points when TUAB and TUEV, respectively, are excluded from pretraining under that study's evaluation protocol (Wang et al., 2025). These results support the effectiveness of POSE for adapting the evaluated pretrained backbones, with performance and selection patterns interpreted in the context of their pretraining histories.

C.3 Implementation Details

C.3.1 Hardware and Software

Experiments were conducted on a single NVIDIA V100 GPU with 32 GB of memory. The code will be open-sourced on GitHub upon acceptance. The full list of dependencies and their specific versions will be documented in our accompanying code repository.

C.3.2 Dataset Splitting Protocol

The EEG classification datasets use a cross-subject evaluation protocol with disjoint training, validation, and test subjects. All samples from the same subject remain in a single partition within each fold. CCD is instead partitioned by data release. Table 4 reports the fixed first-fold allocations used for the main comparison. The two additional partitions are described in Appendix D.3.

Table 4: First-fold dataset splitting strategies and allocations. Numeric ranges indicate positions in the sorted subject list, with counts in parentheses. Partition units are subjects except for CCD, where they are releases denoted by R.

Dataset	Partition Units	Split Allocation		
		Train	Val	Test
BCIC-IV-2a	9	3–9 (7)	1 (1)	2 (1)
CCD	11	R1–R3, R6–R11 (9)	R4 (1)	R5 (1)
FACED	123	25–123 (99)	1–12 (12)	13–24 (12)
HGD	14	3–14 (12)	1 (1)	2 (1)
MAMEM1	11	3–11 (9)	1 (1)	2 (1)
PhysioNet-MI	109	21–109 (89)	1–10 (10)	11–20 (10)
SEED	15	3–15 (13)	1 (1)	2 (1)
SEED-IV	15	3–15 (13)	1 (1)	2 (1)
SEED-V	16	3–16 (14)	1 (1)	2 (1)
SHU-MI	25	5–25 (21)	1–2 (2)	3–4 (2)
TUAB	2328	465–2328 (1864)	1–232 (232)	233–464 (232)
TUEV	370	75–370 (296)	1–37 (37)	38–74 (37)

C.3.3 Meta Selection Protocol

We implement Meta-select as a heuristic baseline that groups datasets based on high-level metadata and restricts auxiliary selection within each group. The grouping follows two principles, grouping datasets that share the same task and using similar acquisition setups when task-based grouping is not applicable. Specifically, we define four groups.

- **Emotion recognition.** FACED, SEED-V, SEED-IV, SEED.
- **Motor imagery.** PhysioNet-MI, BCIC-IV-2a, SHU-MI, HGD.
- **Clinical EEG.** TUAB, TUEV.
- **GSN montage.** MAMEM1, CCD.

The target itself is excluded from its auxiliary group. The first three groups are constructed based on task similarity, under the assumption that datasets with the same paradigm are more likely to be compatible. Datasets in the last group share a common electrode configuration based on the Geodesic Sensor Net system.

C.3.4 Hyperparameters

We use common optimization settings for the main LaBraM and CBraMod experiments, with method-specific search spaces described below. The backbone is trained with a batch size of 32 using AdamW and gradient clipping at norm 1.0. The learning rate follows a cosine schedule with five warm-up epochs and a minimum learning rate of 10^{-6}. The total training length is $E = 20$ epochs. For POSE, the phase boundaries are $T_1 = 4$, $T_2 = 7$, and $T_3 = 17$, giving 4 probe epochs, 3 auxiliary warm-up epochs, 10 joint-training epochs, and 3 cooldown epochs. At most $K = 3$ auxiliary datasets are selected, and the score-mixing coefficient is fixed at $\beta = 0.7$.

For POSE, we fix the base learning rate to 1×10^{-4} for LaBraM and 5×10^{-4} for CBraMod, and the weight decay to 10^{-2} for both backbones. These learning rates and the weight decay are empirically suitable defaults and are held constant across target datasets. We jointly search the maximum auxiliary ratio $\lambda_{\max} \in \{0.5, 1.0, 2.0, 5.0\}$ and selection threshold $\tau \in \{0.1, 0.3\}$, yielding eight configurations for each backbone and target pair.

For baselines that use an auxiliary participation ratio, we jointly search the base learning rate over $\{1 \times 10^{-4}, 3 \times 10^{-4}, 5 \times 10^{-4}\}$ and $\lambda_{\max} \in \{0.5, 1.0, 2.0, 5.0\}$, while keeping weight decay fixed at 10^{-2}. This yields twelve configurations per backbone and target pair. For the remaining baselines

without auxiliary ratio control, we jointly search the same three learning rates and weight decay over $\{10^{-1}, 10^{-2}, 10^{-3}\}$, yielding nine configurations per backbone and target pair.

Tunable hyperparameters are selected separately for each method and target dataset using only target validation performance, maximizing balanced accuracy for classification and minimizing NRMSE for CCD. Checkpoints are likewise selected using the target validation metric, with no test data used for either decision.

C.3.5 Additional Baseline Details

LogME-based selection. For each candidate, we load the backbone from its single-dataset fine-tuned checkpoint, discard the task head, and compute LogME (You et al., 2021) on the same target training examples and labels, using at most 1,000 samples with sampling seed 42. Features are mean-pooled over non-CLS tokens, and all LaBraM candidates use the initial positional embedding table corresponding to the target montage. We use classification LogME for classification targets and regression LogME for CCD. The three highest-scoring datasets among the 11 non-target candidates are selected using training data only and assigned equal auxiliary sampling weights of $1/3$. The final model is reinitialized from the foundation-model pretrained checkpoint rather than a source checkpoint, with a shared backbone and dataset-specific heads. Joint fine-tuning runs for 20 epochs from epoch 1 at a constant auxiliary-to-target batch ratio, with a default ratio of 1, without target-only probing or ratio ramping. The best target-validation checkpoint is evaluated on the target test set.

Shared optimization protocol. ConFIG and GradNorm use all specified tasks without source selection, sharing a backbone with dataset-specific heads. Before each update, one batch per task is used to compute separate losses and gradients, with exhausted data loaders cycled as needed.

ConFIG. ConFIG (Liu et al., 2025b) normalizes task gradients on shared parameters and uses the pseudoinverse of their Gram matrix to obtain an equal-weight composite direction. Projections of the original gradients onto this direction determine the update magnitude. The shared parameters receive the composite gradient, while each task head receives its original gradient, without learned task weights.

GradNorm. GradNorm (Chen et al., 2018) maintains positive task weights initialized to one and adjusts them using shared-parameter gradient norms and loss ratios relative to initial losses, giving slower-learning tasks stronger gradient contributions. We use $\alpha = 1.5$ and a separate Adam optimizer with learning rate 0.025 for the weights, renormalizing their sum to the number of tasks after each update. The shared gradient is the sum of weighted task gradients, and each task head's gradient is scaled by its corresponding weight. These are loss and gradient weights, not sampling probabilities.

D Additional Experiment Results

D.1 Detailed Main Results

Table 5 reports the complete main comparison for LaBraM and CBraMod on the same single-fold protocol used in Section 4.2. Each dataset row reports the mean over five random seeds with the standard deviation in parentheses, with CCD evaluated by NRMSE and all other datasets by balanced accuracy. The average is taken over the 11 classification datasets only.

D.2 Supplementary Evaluation on REVE

To examine whether the benefits extend beyond LaBraM and CBraMod, we additionally evaluate POSE on REVE using three motor-imagery and three emotion-recognition datasets. Table 6 compares POSE with Target-only, Ratio-mixed, and Meta-select on this six-dataset subset.

POSE outperforms Target-only and Meta-select on all six datasets and achieves the best performance on four. It attains the highest average balanced accuracy of 0.576, improving over Target-only and Meta-select by 3.30 and 2.82 percentage points, respectively. These results provide further evidence

Table 5: Performance comparison on LaBraM and CBraMod. Entries report the mean over five random seeds, with standard deviations in parentheses. Average excludes CCD regression task. Superscripts in the average row indicate two-tailed paired t-tests against POSE ($^{*}p < 0.05$, $^{**}p < 0.01$).

Dataset	Target Only	Ratio Mixed	Full Mixed	Random Select	Meta Select	ConFIG	GradNorm	LogME Based	POSE
LaBraM									
BCIC-IV-2a	0.437 (.038)	0.430 (.017)	0.445 (.008)	0.423 (.048)	0.448 (.016)	**0.462** (.022)	0.431 (.021)	0.424 (.010)	0.457 (.039)
FACED	0.552 (.020)	0.562 (.016)	0.526 (.015)	0.552 (.011)	0.563 (.016)	0.532 (.016)	0.541 (.012)	0.555 (.011)	**0.564** (.014)
HGD	0.746 (.024)	0.728 (.014)	0.705 (.011)	0.717 (.043)	0.727 (.026)	0.700 (.029)	0.724 (.023)	0.745 (.047)	**0.757** (.015)
MAMEM1	0.603 (.019)	0.556 (.032)	0.579 (.044)	0.546 (.057)	0.562 (.039)	0.582 (.032)	0.594 (.038)	0.606 (.020)	**0.611** (.019)
PhysioNet-MI	0.576 (.021)	0.587 (.015)	0.562 (.011)	0.589 (.018)	**0.596** (.014)	0.549 (.013)	0.551 (.017)	0.564 (.016)	0.595 (.009)
SEED	0.568 (.037)	0.616 (.016)	0.593 (.014)	0.590 (.039)	0.611 (.025)	0.551 (.024)	0.616 (.027)	0.602 (.050)	**0.618** (.019)
SEED-IV	0.338 (.024)	0.350 (.033)	0.354 (.030)	0.358 (.017)	0.350 (.025)	0.342 (.032)	0.303 (.022)	0.339 (.015)	**0.359** (.011)
SEED-V	0.270 (.016)	0.285 (.020)	0.284 (.032)	**0.291** (.023)	0.283 (.018)	**0.291** (.017)	0.284 (.027)	0.254 (.006)	**0.291** (.038)
SHU-MI	0.557 (.020)	0.569 (.017)	0.554 (.014)	0.561 (.012)	0.556 (.018)	0.543 (.018)	0.555 (.017)	0.562 (.010)	**0.571** (.007)
TUAB	0.806 (.011)	0.797 (.005)	**0.807** (.009)	0.802 (.018)	0.802 (.007)	0.798 (.011)	0.790 (.009)	0.800 (.008)	0.801 (.011)
TUEV	**0.537** (.044)	0.529 (.025)	0.488 (.056)	0.524 (.048)	0.532 (.045)	0.493 (.051)	0.525 (.040)	0.523 (.003)	0.532 (.044)
CCD (↓)	0.996 (.012)	**0.983** (.015)	1.046 (.005)	1.013 (.031)	0.984 (.012)	1.156 (.016)	1.149 (.014)	1.038 (.006)	0.988 (.016)
Average	0.545^{**}	0.546^{*}	0.536^{**}	0.541^{*}	0.548^{*}	0.531^{**}	0.538^{**}	0.543^{**}	**0.560**
CBraMod									
BCIC-IV-2a	0.433 (.033)	0.448 (.022)	0.424 (.026)	0.458 (.056)	0.456 (.020)	0.406 (.032)	0.396 (.034)	0.451 (.024)	**0.469** (.028)
FACED	0.570 (.005)	0.561 (.012)	0.558 (.011)	0.565 (.015)	0.570 (.019)	0.491 (.011)	0.494 (.013)	0.563 (.003)	**0.577** (.018)
HGD	**0.719** (.017)	0.700 (.008)	0.680 (.006)	0.690 (.008)	0.682 (.013)	0.638 (.012)	0.693 (.012)	0.673 (.009)	0.712 (.013)
MAMEM1	0.487 (.041)	0.537 (.034)	0.423 (.059)	0.505 (.051)	0.542 (.021)	0.556 (.039)	0.556 (.038)	0.528 (.010)	**0.562** (.021)
PhysioNet-MI	0.597 (.012)	0.593 (.007)	0.552 (.022)	0.594 (.007)	0.602 (.011)	0.559 (.014)	0.562 (.010)	0.598 (.016)	**0.603** (.005)
SEED	0.565 (.049)	0.583 (.009)	0.595 (.030)	0.574 (.033)	0.605 (.033)	0.506 (.040)	0.573 (.043)	0.538 (.029)	**0.613** (.018)
SEED-IV	0.325 (.017)	0.361 (.038)	0.353 (.008)	0.353 (.020)	0.352 (.032)	0.316 (.027)	0.358 (.025)	0.321 (.013)	**0.368** (.025)
SEED-V	0.244 (.031)	0.281 (.015)	0.265 (.024)	0.269 (.027)	**0.285** (.016)	0.276 (.029)	0.236 (.024)	0.200 (.007)	0.269 (.041)
SHU-MI	0.556 (.009)	0.563 (.009)	**0.569** (.011)	0.563 (.007)	0.559 (.015)	0.556 (.010)	0.557 (.009)	0.543 (.004)	**0.569** (.009)
TUAB	0.809 (.004)	0.797 (.004)	0.778 (.006)	0.799 (.010)	**0.810** (.007)	0.761 (.008)	0.768 (.007)	0.778 (.004)	0.809 (.007)
TUEV	**0.507** (.016)	0.482 (.030)	0.418 (.042)	0.456 (.049)	0.451 (.026)	0.426 (.035)	0.367 (.037)	0.498 (.016)	0.499 (.018)
CCD (↓)	0.974 (.021)	0.978 (.027)	1.061 (.100)	0.990 (.011)	0.963 (.005)	1.106 (.015)	1.246 (.043)	0.990 (.012)	**0.962** (.028)
Average	0.528^{*}	0.537^{**}	0.510^{**}	0.530^{**}	0.538^{*}	0.499^{**}	0.505^{**}	0.517^{**}	**0.550**

Table 6: Supplementary performance comparison on REVE. Entries report balanced accuracy, and Average is computed over the six evaluated datasets. Boldface indicates the best result in each row.

Dataset	Target-only	Ratio-mixed	Meta-select	POSE
REVE				
BCIC-IV-2a	0.479	**0.514**	0.465	0.493
PhysioNet-MI	0.621	0.648	0.640	**0.654**
FACED	0.642	**0.673**	0.643	0.661
SHU-MI	0.574	0.617	0.621	**0.628**
SEED	0.587	0.577	0.572	**0.638**
SEED-IV	0.355	0.379	0.347	**0.382**
Average	0.543	0.568	0.548	**0.576**

for the effectiveness of target-conditioned auxiliary integration on an additional EEG foundation model across the evaluated tasks.

D.3 Evaluation Across Subject Partitions

To assess sensitivity to data partitioning, we extend the main evaluation with two additional splits, giving three folds in total. Fold 1 retains the original allocations in Table 4, while only Folds 2 and 3 are generated using the rules below. For the classification datasets, subject identifiers are sorted before partitioning, and all samples from the same subject are assigned together to one partition within each fold.

For datasets with at least 20 subjects, we divide the sorted subject list into ten consecutive, nearly equal-sized groups $G_1, \ldots, G_{10}$ using `np.array_split(subjects, 10)`. Fold 2 uses G_3 for validation and G_4 for testing, while Fold 3 uses G_5 for validation and G_6 for testing. The remaining eight groups form the training set in each fold, giving approximately 80% training, 10% validation, and 10% test subjects.

For datasets with fewer than 20 subjects, each additional fold reserves one subject for validation and one for testing, using all remaining subjects for training. With N_{subj} subjects and zero-based fold index i, the validation and test subjects occupy zero-based sorted positions $(2i) \bmod N_{\text{subj}}$ and $(2i+1) \bmod N_{\text{subj}}$, respectively. We use $i = 1$ for Fold 2 and $i = 2$ for Fold 3. For example, with nine subjects, Fold 2 uses the third and fourth subjects for validation and testing, respectively, while Fold 3 uses the fifth and sixth. These positions refer to the ordering of the sorted subject identifiers rather than their literal values.

For CCD, the validation and test releases are R4 and R5 in Fold 1, R1 and R2 in Fold 2, and R6 and R7 in Fold 3, respectively. The other nine releases form the training set in each fold, with all samples from a release kept in the same partition.

We compare Target-only, Ratio-mixed, Meta-select, and POSE using the same split for all methods within each fold. Table 7 reports the mean performance and sample standard deviations of the three fold-level scores.

POSE achieves the highest average classification performance among the compared methods in each of the three folds for both backbones, yielding fold-averaged scores of 0.570 on LaBraM and 0.552 on CBraMod. These consistent aggregate gains show that the main findings are not confined to the original subject partition, providing further evidence of robust target adaptation under changes in the held-out subjects.

D.4 Component and Metric Ablations

Table 8 extends the component and metric comparisons in Section 4.3 to both LaBraM and CBraMod. Each variant changes the specified component while retaining the remaining settings of the complete framework. The POSE entries match the main comparison, and each dataset reports the mean and standard deviation over five random seeds. The eight classification datasets are used to examine component differences in settings where POSE improves over Target-only on both backbones, with the other four targets evaluated separately below.

The complete framework achieves the highest average performance in both comparison groups for each backbone. No Probe produces the largest average decrease among the component variants, indicating that target-adapted compatibility estimates offer a stronger selection basis than the initial pretrained representation. The gains from warm-up, cooldown, and weighting vary across datasets. Nevertheless, the full design provides the strongest average performance, while the FD+MMD combination also outperforms either metric alone on average.

D.5 Ablations on the Remaining Datasets

Table 9 reports additional ablations on CCD, HGD, TUAB, and TUEV under a fixed exploratory configuration. We set $\tau = 0.3$ and $\lambda_{\max} = 1.0$, with learning rates of 1×10^{-4} for LaBraM and 5×10^{-4} for CBraMod. No Probe and Uniform Weight follow the definitions in Section 4.3, while No Ramp removes both the auxiliary warm-up and cooldown. Each setting is evaluated over five random seeds.

Probing retains the most consistent benefit in these settings, with the full framework improving on No Probe in seven of the eight backbone and dataset combinations. The effects of ramp scheduling

Table 7: Performance across three data partitions for LaBraM and CBraMod, using subject-based splits for classification and release-based splits for CCD. Entries report the mean across folds, with the sample standard deviation across the three fold-level scores in parentheses. Classification tasks use balanced accuracy (↑), while CCD uses NRMSE (↓). Average excludes CCD and summarizes the three fold-level classification averages.

Dataset	Target-only	Ratio-mixed	Meta-select	POSE
		LaBraM		
BCIC-IV-2a	0.386 (.044)	0.428 (.031)	0.436 (.026)	**0.442** (.025)
FACED	0.563 (.072)	0.561 (.072)	**0.578** (.066)	0.577 (.064)
HGD	0.745 (.047)	0.742 (.052)	0.733 (.043)	**0.776** (.043)
MAMEM1	0.552 (.064)	0.518 (.114)	0.514 (.067)	**0.595** (.086)
PhysioNet-MI	0.584 (.029)	0.606 (.019)	**0.619** (.030)	0.610 (.020)
SEED	0.540 (.036)	0.549 (.059)	0.548 (.061)	**0.556** (.060)
SEED-IV	0.368 (.042)	0.346 (.019)	0.332 (.016)	**0.387** (.054)
SEED-V	0.239 (.035)	0.237 (.042)	0.223 (.055)	**0.256** (.032)
SHU-MI	0.597 (.088)	0.604 (.088)	0.598 (.091)	**0.605** (.086)
TUAB	**0.819** (.012)	0.818 (.018)	0.814 (.010)	0.818 (.016)
TUEV	0.600 (.058)	0.582 (.050)	0.611 (.070)	**0.643** (.097)
CCD (↓)	1.003 (.017)	1.007 (.025)	1.017 (.036)	**0.997** (.008)
Average	0.545 (.008)	0.545 (.005)	0.546 (.002)	**0.570** (.009)
		CBraMod		
BCIC-IV-2a	0.474 (.051)	0.469 (.058)	0.466 (.034)	**0.498** (.041)
FACED	0.564 (.067)	0.555 (.070)	0.559 (.066)	**0.581** (.061)
HGD	0.711 (.053)	0.699 (.031)	0.698 (.048)	**0.723** (.026)
MAMEM1	0.406 (.114)	**0.461** (.148)	0.415 (.127)	0.441 (.131)
PhysioNet-MI	0.601 (.023)	0.608 (.017)	0.594 (.008)	**0.617** (.018)
SEED	0.519 (.039)	0.527 (.050)	0.539 (.057)	**0.547** (.058)
SEED-IV	0.378 (.047)	0.398 (.037)	0.391 (.037)	**0.406** (.043)
SEED-V	0.222 (.041)	0.220 (.054)	0.228 (.056)	**0.259** (.042)
SHU-MI	0.554 (.020)	0.550 (.019)	**0.578** (.055)	0.562 (.018)
TUAB	0.811 (.002)	0.808 (.017)	0.813 (.006)	**0.814** (.011)
TUEV	0.612 (.093)	0.592 (.096)	0.596 (.128)	**0.627** (.114)
CCD (↓)	0.985 (.010)	1.070 (.085)	0.997 (.030)	**0.979** (.019)
Average	0.532 (.006)	0.535 (.005)	0.534 (.004)	**0.552** (.006)

and weighting are smaller or more variable, particularly on TUAB, where the variants remain close. These observations broadly agree with the main ablation, although differences between variants are generally smaller.

D.6 Cross-Dataset Affinity Matrix

For each backbone, target T, and seed r, let $\mathcal{S}_T^{(r)}$ denote the selected auxiliary subset and $w_{Tj}^{(r)}$ its normalized sampling weights. The affinity matrix averages the final weights obtained after probing over $R = 5$ seeds.

$$A_{Tj} = \frac{1}{R}\sum_{r=1}^{R} \widetilde{w}_{Tj}^{(r)}, \qquad \widetilde{w}_{Tj}^{(r)} = \begin{cases} w_{Tj}^{(r)}, & j \in \mathcal{S}_T^{(r)}, \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

Rows correspond to targets and columns to candidate auxiliaries, with the target itself excluded from its auxiliary pool. The entries are not averages of discrepancy scores across probe epochs.

Table 8: Component and metric ablations on both backbones. Entries report mean balanced accuracy over five random seeds, with standard deviations in parentheses. POSE is shared by the component comparison (left) and the metric comparison (right). Boldface marks the best mean in the component comparison. Underlining marks the best mean in the metric comparison.

Dataset	No Probe	No Warm-up	No Cool-down	Uniform Weight	POSE	FD-only	MMD-only
LaBraM							
BCIC-IV-2a	0.422 (.026)	0.427 (.033)	0.431 (.030)	0.432 (.036)	**0.457** (.039)	0.433 (.019)	0.419 (.024)
FACED	0.540 (.029)	0.534 (.022)	0.536 (.023)	0.543 (.029)	**0.564** (.014)	0.547 (.019)	0.557 (.016)
MAMEM1	0.554 (.040)	0.576 (.041)	0.570 (.030)	0.548 (.030)	**0.611** (.019)	0.604 (.044)	0.543 (.055)
PhysioNet-MI	0.587 (.011)	0.573 (.022)	0.574 (.014)	0.582 (.022)	**0.595** (.009)	0.584 (.013)	0.589 (.015)
SEED	0.612 (.009)	0.613 (.016)	**0.619** (.022)	0.606 (.022)	0.618 (.019)	0.596 (.032)	0.588 (.031)
SEED-IV	0.332 (.031)	0.346 (.017)	0.358 (.019)	0.345 (.023)	**0.359** (.011)	0.334 (.028)	0.338 (.026)
SEED-V	0.264 (.025)	0.275 (.029)	0.278 (.010)	0.272 (.051)	**0.291** (.038)	0.287 (.017)	0.282 (.050)
SHU-MI	0.557 (.019)	0.566 (.013)	**0.574** (.009)	0.567 (.007)	0.571 (.007)	0.553 (.017)	0.569 (.014)
Average	0.484	0.489	0.493	0.487	**0.508**	0.492	0.485
CBraMod							
BCIC-IV-2a	0.457 (.023)	0.438 (.024)	0.417 (.023)	0.433 (.038)	**0.469** (.028)	0.440 (.012)	0.428 (.032)
FACED	**0.580** (.007)	0.566 (.004)	0.576 (.006)	0.563 (.008)	0.577 (.018)	0.574 (.011)	0.571 (.006)
MAMEM1	0.494 (.043)	0.499 (.004)	0.496 (.026)	0.518 (.019)	**0.562** (.021)	0.488 (.075)	0.539 (.037)
PhysioNet-MI	0.583 (.013)	**0.603** (.001)	0.591 (.008)	0.585 (.010)	**0.603** (.005)	0.591 (.008)	0.593 (.009)
SEED	0.576 (.054)	0.581 (.028)	0.593 (.029)	0.607 (.021)	**0.613** (.018)	0.592 (.031)	0.587 (.054)
SEED-IV	0.346 (.011)	0.359 (.036)	**0.368** (.024)	**0.368** (.025)	**0.368** (.025)	0.338 (.043)	0.367 (.024)
SEED-V	0.248 (.026)	0.237 (.023)	0.249 (.023)	0.259 (.024)	**0.269** (.041)	0.268 (.017)	0.267 (.029)
SHU-MI	0.554 (.008)	0.563 (.004)	0.555 (.012)	0.561 (.013)	**0.569** (.009)	0.556 (.007)	0.569 (.006)
Average	0.480	0.481	0.481	0.487	**0.504**	0.481	0.490

Table 9: Fixed-configuration ablations on the four remaining datasets. Entries report means over five random seeds, with standard deviations in parentheses. Boldface marks the best mean per dataset and backbone.

Dataset	No Probe	No Ramp	Uniform Weight	POSE
LaBraM				
CCD (↓)	1.037 (.029)	**1.036** (.030)	1.046 (.050)	**1.036** (.038)
HGD	0.731 (.021)	0.735 (.026)	**0.758** (.010)	0.752 (.012)
TUAB	**0.798** (.009)	**0.798** (.010)	0.797 (.011)	0.797 (.009)
TUEV	0.514 (.047)	0.493 (.075)	0.536 (.051)	**0.551** (.059)
CBraMod				
CCD (↓)	1.036 (.030)	**0.981** (.019)	0.988 (.035)	**0.981** (.025)
HGD	0.679 (.007)	0.690 (.015)	0.692 (.006)	**0.695** (.008)
TUAB	0.788 (.014)	0.787 (.009)	**0.794** (.010)	0.792 (.009)
TUEV	0.435 (.052)	0.433 (.029)	0.430 (.036)	**0.438** (.032)

Because different seeds may select different subsets, the aggregated row can contain more than K nonzero entries even though each individual run selects at most K sources. An empty selected subset contributes an all-zero row, so the averaged row sum can be below one. The matrix therefore summarizes selection frequency and conditional sampling weight rather than the marginal performance benefit of an individual source.

D.7 Case Study

Figure 4 presents the CBraMod runs on MAMEM1 and TUAB discussed in Section 4.5.

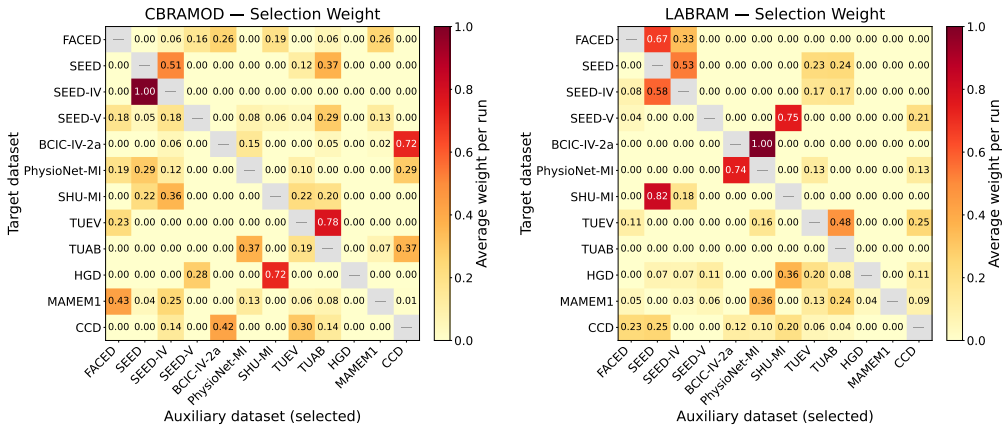

Figure 3: Cross-dataset affinity matrices for CBraMod on the left and LaBraM on the right. Each cell reports the mean final sampling weight assigned to an auxiliary dataset in a column for a target dataset in a row, averaged over five seeds with unselected candidates assigned zero. Higher values indicate greater average allocation within the auxiliary batches, not directly measured transfer gains.

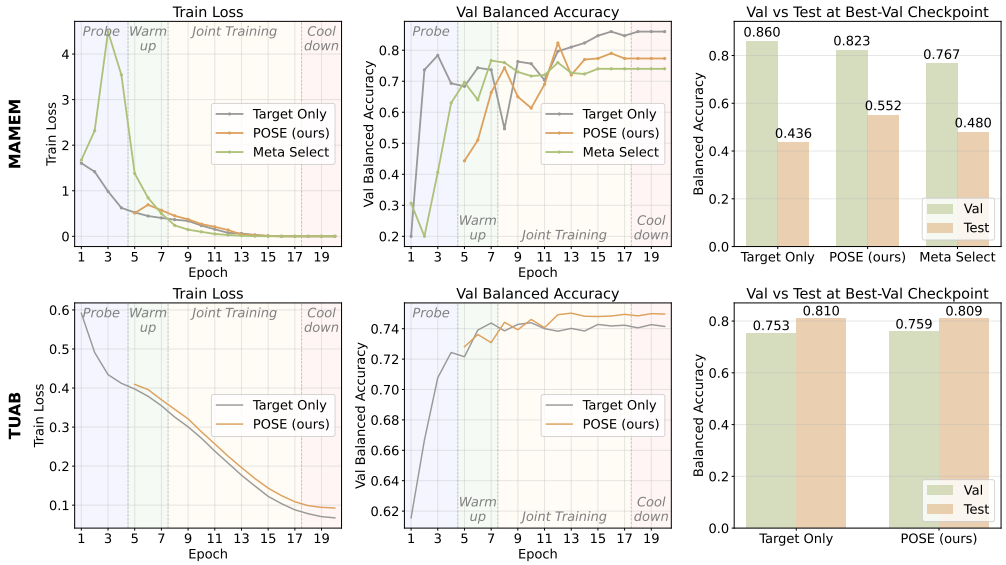

Figure 4: Case study on two representative datasets using CBraMod. **Top row (MAMEM):** A comparison among Target Only, POSE (ours), and Meta Select, showing training loss (left), validation balanced accuracy (middle), and validation vs. test performance at the best-validation checkpoint (right). **Bottom row (TUAB):** Comparison between Target Only and POSE.

D.8 Detailed Comparison with Naive Dataset Mixing

Motivation. While our method aims to select a compatible subset of auxiliary datasets, an important question is how competitive the selected composition is compared to naive mixing strategies. In particular, given the large number of possible dataset combinations, it is unclear whether a simple random choice could achieve comparable performance. We therefore evaluate where our selected composition ranks within the space of naive dataset mixtures.

Experimental Setup. Given 12 datasets, there are $2^{12} - 1 = 4{,}095$ possible non-empty combinations. Exhaustively evaluating all combinations across two backbones, five random seeds, and searching for learning rates is computationally infeasible. Instead, for each backbone, we randomly sample 500 dataset combinations and assign each combination a random seed and learning rate,

thereby approximating the performance distribution of naive mixing. Importantly, each sampled result corresponds to a single-seed run, whereas our method reports the average performance over five seeds.

Results. We report the percentile rank of our method within the sampled distribution for each target dataset, as shown in Figure 2. Across both backbones, our selected compositions consistently lie near the top of the distribution, with an average of near 90%. These results indicate that the selected compositions are consistently among the strongest candidates within a large space of naive mixtures.

Discussion. EEG datasets differ in task, acquisition conditions, and participant populations, while backbone inductive biases shape how these differences affect transfer. The joint value of auxiliary sources can also depend on their interactions, making a globally optimal subset difficult to identify without evaluating a large number of training combinations. Our objective is instead to reliably identify compositions that support stable and robust target adaptation without exhaustive search. The competitive results in this analysis, together with the main baseline comparisons and repeated-seed and subject-partition evaluations, support this practical objective.

D.9 Probe Analysis

We further analyze the probing stage from three complementary perspectives to characterize how target adaptation and temporal trends affect auxiliary selection.

Auxiliary rankings evolve during target-only probing. We first measure the consistency between the auxiliary-dataset rankings obtained at probe epoch 1 and epoch 4 using Kendall's τ. Across all target datasets, seeds, and both distance metrics, the average τ is 0.509 on CBraMod and 0.564 on LaBraM. These moderate values indicate that the relative ordering of candidate auxiliaries changes noticeably during probing, rather than remaining fixed throughout optimization. The target-only probe stage therefore captures changes in relative compatibility that are not described by a single-checkpoint ranking.

Probe-guided and No Probe selections overlap only weakly. To further quantify the difference between target-adapted probing and selection from the pretrained checkpoint, we compute the Jaccard overlap between the auxiliary sets selected by POSE and No Probe. The average Jaccard scores are 0.226 on CBraMod and 0.256 on LaBraM, showing that the two strategies often select largely different auxiliary subsets. This low overlap indicates that target-adapted probing materially changes the selection outcome.

The slope signal changes selection decisions. Finally, we examine whether incorporating the distance slope affects the choice of auxiliaries. We compare the composite score, which combines the normalized slope and endpoint value, against a simulated endpoint-only selection using $\beta = 1$ in Equation (5). The reported top-3 change rate is 84.2% on CBraMod and 81.2% on LaBraM, indicating that the slope term frequently changes candidate prioritization. This result shows that the temporal signal is not redundant with the final probe value. It measures changes in selection decisions rather than isolating the effect of the slope term on downstream performance.

Overall, these analyses support the design of our dynamic probing mechanism. Auxiliary rankings evolve during target adaptation, probe-guided selection differs markedly from pretrained-space selection, and the slope signal frequently changes the final ranking. These observations complement the performance ablations in Section 4.3 and are consistent with assessing compatibility along the target adaptation trajectory rather than at a single checkpoint.

D.10 Hyperparameter Sensitivity Analysis

We further analyze the sensitivity of our framework to key hyperparameters and implementation choices, focusing on robustness and practical usability. This supplementary sweep evaluates the same auxiliary-ratio and threshold grid used for POSE in Appendix C.3.4 and is reported separately from the validation-selected main comparison.

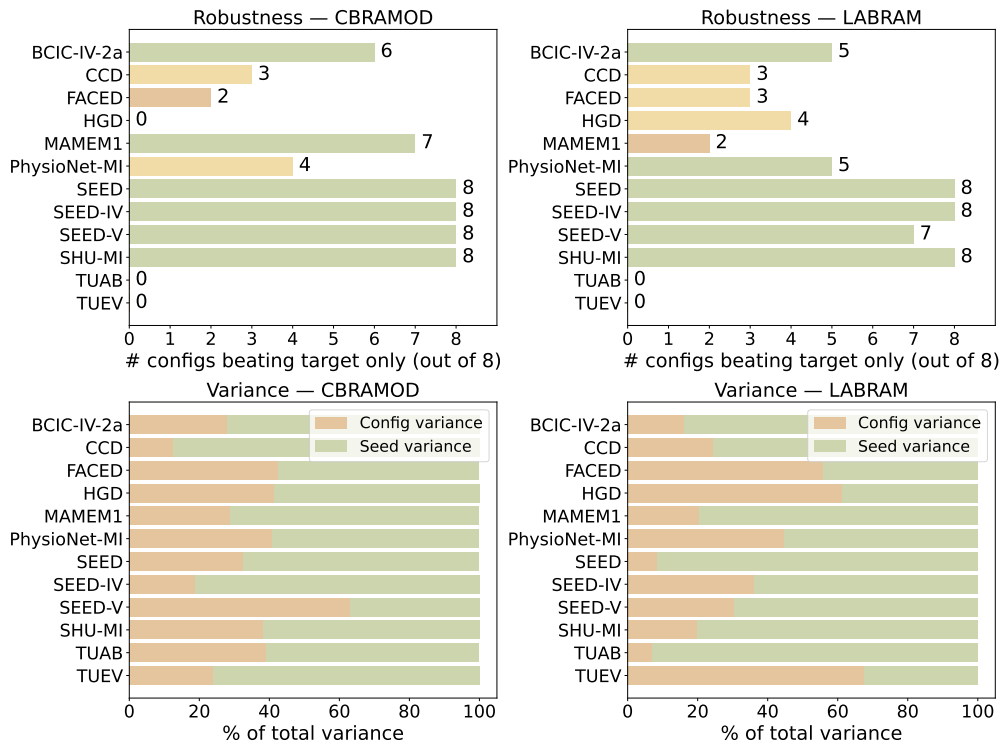

Figure 5: Hyperparameter sensitivity analysis across 12 datasets. The top row shows the number of hyperparameter configurations out of eight, varying the auxiliary ratio and selection threshold, that outperform Target-only for CBraMod on the left and LaBraM on the right. The bottom row shows the relative contributions of hyperparameter configuration and random seed to performance variance within the evaluated sweep.

Improvement across hyperparameter configurations. We evaluate how often the proposed method outperforms the standard target-only baseline under different hyperparameter settings. Specifically, we consider a grid of two thresholds $\tau \in \{0.1, 0.3\}$ and four auxiliary ratios $\lambda_{\max} \in \{0.5, 1.0, 2.0, 5.0\}$, resulting in eight configurations per dataset. Figure 5 reports, for each dataset, the number of configurations that surpass the target-only baseline. We compute a conditional success rate over targets for which at least one configuration improves on Target-only, excluding targets with no improvement in this sweep. The resulting success rates are 66.3% (53/80) for LaBraM and 75.0% (54/72) for CBraMod. Within these targets, improvements occur across a substantial fraction of the explored settings rather than only at a single configuration. At the same time, the results reveal clear dataset-dependent patterns. Several datasets, including SHU-MI, SEED, and SEED-IV, benefit across many configurations, whereas gains are less frequent on targets such as TUAB and TUEV in this sweep. This variation supports evaluating auxiliary participation for the particular target and backbone rather than inferring its value from dataset size alone.

Relative contributions of configuration and seed variation. To further quantify the impact of hyperparameters, we perform a variance decomposition analysis (Figure 5). Within the explored hyperparameter range, the performance variation induced by hyperparameter choices accounts for only $\sim$34% of the total variance on CBraMod and $\sim$33% on LaBraM, while the remaining $\sim$66% and $\sim$67%, respectively, are attributed to randomness across different seeds. This suggests that the method is relatively insensitive to hyperparameter tuning, and that most performance fluctuations arise from inherent stochasticity rather than from unstable design choices.

Effect of probe sample size. Finally, we examine the impact of the sample budget used during probing. To balance computational cost and estimation quality, our default setting uses a budget of 1,000 samples per dataset. Increasing this budget to 5,000 does not lead to systematic improvements,

Table 10: Computational overhead on SHU-MI and FACED. Overhead denotes time spent on distance computation during probing, and the percentage is its share of POSE's total training time.

Dataset	Model	Total Time	Overhead	Percentage
SHU-MI	LaBraM	31,744 s	4,466 s	14%
	CBraMod	25,493 s	4,157 s	16%
FACED	LaBraM	59,959 s	4,843 s	8%
	CBraMod	70,034 s	6,477 s	9%

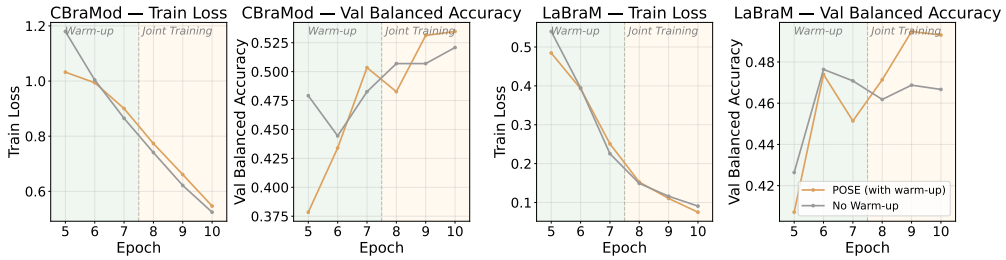

Figure 6: Warm-up phase dynamics on BCIC-IV-2a for CBraMod in the left two panels and LaBraM in the right two panels. Each panel compares POSE with No Warm-up, which directly enters joint training after probing, over epochs 5 to 10.

with reported average performance changes of -0.016 on CBraMod and -0.003 on LaBraM. These observations support the default probe budget as a practical choice in the evaluated setting.

Overall, these results demonstrate that our framework is robust to hyperparameter choices and computational design decisions, further supporting its practicality for real-world EEG foundation model adaptation.

D.11 Computational Overhead Analysis

We analyze the computational overhead introduced by discrepancy estimation during probing, focusing on its practical impact under a representative setting. The timing experiments use $\lambda_{\max} = 2.0$, a total of 20 training epochs, and 4 probe epochs. The evaluated datasets, SHU-MI and FACED, each contain on the order of 10^4 samples.

Table 10 reports the total training time and the time spent on distance computation during probing. Across both backbones and datasets, distance computation accounts for 8% to 16% of the total training time. These percentages are fractions of POSE's runtime rather than slowdowns relative to Target-only. The target-only updates in the probe stage are already included in the 20 training epochs.

The exact overhead depends on the probe duration, auxiliary ratio, and target dataset size. Nevertheless, these measurements show that under the evaluated configurations, discrepancy estimation accounts for a limited share of training time. All candidates are assessed along the same target-only trajectory, avoiding separate source-specific training or exhaustive auxiliary-combination evaluation.

D.12 Training Dynamics

We further analyze the role of the proposed training schedule by examining the dynamics of warm-up and cooldown. For illustration, we show one representative random seed for each setting.

Effect of warm-up. Warm-up provides a gradual transition from target-only optimization to multi-dataset training. Figure 6 illustrates that faster reduction in training loss need not translate into corresponding gains in target validation performance. For example, on BCIC-IV-2a with CBraMod, No Warm-up achieves lower training loss than POSE at early stages, yet its validation accuracy remains

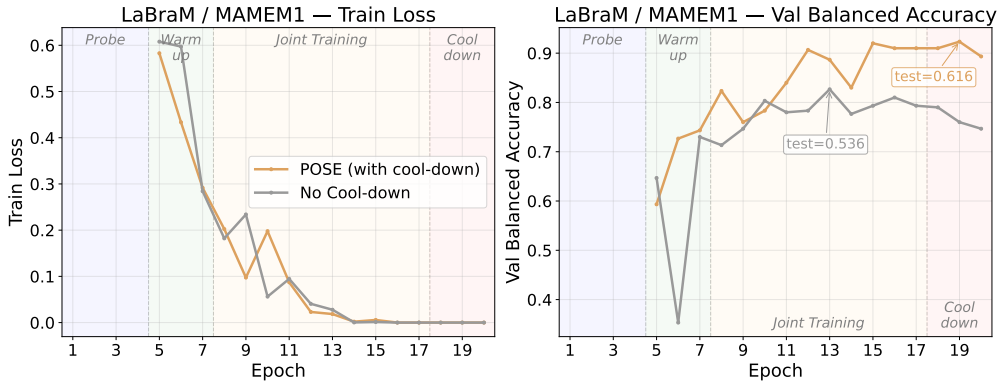

Figure 7: Cooldown phase dynamics on MAMEM1 using LaBraM. Training loss on the left and validation balanced accuracy on the right are compared between POSE and No Cool-down over all 20 epochs. Test balanced accuracy at the best-validation checkpoint is annotated for each method.

lower or more fluctuating. Training losses include changing auxiliary batch compositions and are not directly comparable as measures of target fit. After target-only probing, warm-up gradually increases auxiliary participation, limiting the abrupt change in training composition. The illustrated trajectories are consistent with the benefit of this controlled transition observed in the aggregate ablations.

Effect of cooldown. Cooldown further refines the model by gradually reducing auxiliary influence in the final stage of training. On MAMEM1 with LaBraM, POSE achieves its best validation performance at epoch 19, whereas No Cool-down peaks at epoch 13 and subsequently degrades, as shown in Figure 7. The gradual decay of the auxiliary ratio allows the model to re-focus on the target distribution after benefiting from auxiliary signals. The illustrated run is consistent with improved target adaptation during the final phase.

Takeaway. Together with the aggregate ablations, these trajectories support complementary roles for the two scheduling components. Warm-up controls the transition into auxiliary training, while cooldown restores target-focused training in the final stage. This supports controlling when and how strongly selected sources participate, in addition to deciding which sources to include.

D.13 Discussion on Gradient-Based Approaches

Gradient similarity as a selection metric. A natural alternative to representation-based distance metrics is to measure inter-dataset affinity through gradient similarity, for example by computing the cosine similarity between per-dataset gradient vectors on shared backbone parameters. We conducted preliminary experiments using gradient alignment as a probe metric, replacing MMD and FD with gradient cosine similarity. A separate exploratory diagnostic on MAMEM1 with CBraMod evaluated 13 candidates across 11 checkpoints, including the pretrained initialization and 10 training epochs. This diagnostic uses a different candidate-pool size and checkpoint schedule from the main 12-dataset benchmark and its four-epoch probe. The mean rank of each candidate ranged from 5.2 to 8.3 for gradient similarity, compared with 2.4 to 12.8 for MMD in the same setting. Gradient-based rankings also varied substantially across checkpoints, making candidate prioritization sensitive to the point at which gradients were evaluated. These preliminary observations favor the distributional probe in this diagnostic, but do not establish that gradient-based selection is generally ineffective.

PCGrad during training. We also evaluated PCGrad (Yu et al., 2020), a representative gradient-based multi-task training strategy that resolves inter-task gradient conflicts by projecting each task gradient to remove conflicting components. Unlike POSE, the evaluated PCGrad baseline does not perform explicit dataset selection and instead uses all auxiliary datasets while applying gradient surgery at each coordinated update. We tested PCGrad on several targets across both backbones,

Table 11: Supplementary PCGrad comparison on LaBraM and CBraMod. CCD uses NRMSE (↓), while the other datasets use balanced accuracy (↑).

Dataset	LaBraM			CBraMod		
	Target-only	POSE	PCGrad	Target-only	POSE	PCGrad
BCIC-IV-2a	0.437	**0.457**	0.441	0.433	**0.469**	0.415
CCD (↓)	0.996	**0.988**	1.110	0.974	**0.962**	1.017
FACED	0.552	**0.564**	0.516	0.570	**0.577**	0.504
HGD	0.746	**0.757**	0.684	**0.719**	0.712	0.634
MAMEM1	0.603	**0.611**	0.594	0.487	**0.562**	0.512
SEED	0.568	**0.618**	0.614	0.565	**0.613**	0.550
SEED-V	0.270	**0.291**	0.281	0.244	**0.269**	0.265

with detailed results reported in Table 11. The Target-only and POSE entries match the main comparison in Table 5.

Although PCGrad occasionally exceeds the target-only baseline, its performance varies across datasets and often degrades under the evaluated heterogeneous multi-dataset setting. In this implementation, each coordinated update requires $(N + 1)$ forward and backward passes, where N is the number of auxiliary datasets. POSE instead uses a single-dataset batch per update and controls auxiliary exposure through selection, weighting, and scheduling. These supplementary results are consistent with the ConFIG and GradNorm comparisons in Section 4.2, supporting explicit control of data composition in addition to gradient coordination. They also motivate further investigation of gradient-based optimization strategies tailored to heterogeneous EEG tasks.

E Limitations and Future Work

The highly heterogeneous nature of EEG data makes cross-dataset transfer inherently complex and strongly context-dependent. Differences in task formulations, acquisition conditions, and participant populations interact with backbone inductive biases, so the value of an auxiliary composition is difficult to infer from dataset descriptions or individual source assessments alone. Assessing the combined contribution of sources across a combinatorial search space therefore remains a demanding problem in heterogeneous EEG learning.

POSE assesses the evolving relationships between the target and each auxiliary dataset to guide selection and participation. Building on this target-conditioned assessment, jointly considering relationships among auxiliary datasets is a promising direction for future work. Such modeling could identify complementary sources, anticipate interference within a candidate mixture, and guide their joint selection and weighting. Extending this assessment into joint training could further adapt auxiliary participation as the representation evolves.

These directions extend the scope of compatibility modeling while building on the practical contribution established by POSE. Our objective is to reliably identify auxiliary compositions that support stable and robust target adaptation without exhaustive search, rather than to recover a globally optimal subset, as we emphasized in the main text. Comprehensive experiments across datasets and backbones, together with repeated-seed and subject-partition evaluations, provide empirical support for this objective.