SAIL: STABLE ANCHORING WITH INDIVIDUAL-CONDITIONED LEARNING FOR CROSS-SUBJECT TIME SERIES CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-subject classification of biomedical time series is challenging because the same task semantics can be expressed differently across individuals, causing models trained on seen subjects to generalize poorly to unseen subjects. Existing methods typically suppress subject-specific information to learn invariant representations, which may impair the preservation of task-relevant semantics. In this paper, we rethink the relationship between task semantics and individual variation, viewing semantic representations as shared task semantics modulated by individual-specific variation. We further find that cross-subject variation can be decomposed into reusable shift modes, which can be recombined to adapt semantic representations to unseen individuals. Based on these observations, we propose SAIL, which extracts a physiology-inspired semantic anchor and employs a mixture-of-experts (MoE) architecture to model reusable variation patterns. Individual context conditionally selects and combines these patterns, while feature-wise gating controls their contribution to the final representation. Across five medical datasets, SAIL improves the average macro-F1 over the lightweight TeCh baseline by 5.12% , while achieving approximately 9× faster inference.

1 INTRODUCTION

Cross-subject classification of biomedical time series (BTS) aims to learn task-relevant representations from observed individuals and generalize them to previously unseen subjects. This setting is particularly important for real-world applications of physiological signals such as electroencephalography (EEG) and electrocardiography (ECG), including medical diagnosis Hannun et al. (2019), emotion recognition Shen et al. (2023), and brain–computer interfaces Lawhern et al. (2018), where robust generalization across individuals is essential for practical deployment.

Cross-subject generalization is challenging because the same task semantics can exhibit substantially different signal distributions across individuals due to physiological variability Mahmood et al. (2019). As shown in Figure 1 (a), subjects exhibit broadly consistent spectral patterns, summarized by the solid lines. These shared spectral patterns reflect physiological rhythms consistently associated with the same task, which we refer to as a *semantic anchor*. However, individual subjects still deviate substantially from this anchor, reflecting subject-specific variations in how the same semantics are expressed Demuru & Fraschini (2020). These variations induce cross-subject distribution shifts and make it difficult for models trained on seen subjects to generalize to unseen individuals.

In this paper, we revisit Cross-subject classification from the perspective of how individual variation interacts with task semantics. Existing methods typically seeks to recover shared task semantics by reducing subject-specific variation. Although implemented differently, many distribution alignment Long et al. (2015), adversarial suppression Ganin et al. (2016) and explicit disentanglement Yang et al. (2026) approaches follow this subject-invariant learning principle. However, task semantics and individual variation can be coupled: individual differences may influence how task-relevant patterns are expressed, rather than merely introduce irrelevant noise. Consequently, indiscriminately suppressing individual information may also remove task-relevant components and impair the semantic representation Zhao et al. (2019) (Sec 4.1).

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

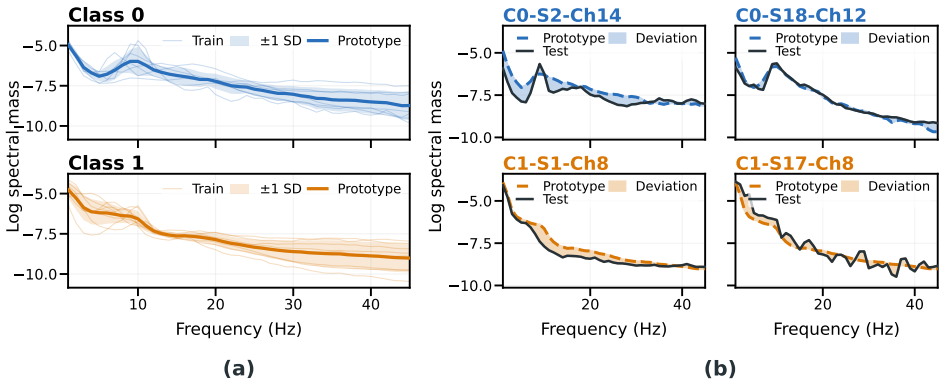

Figure 1: **Cross-subject semantics are shared in support but individualized in expression. (a)** Channel-averaged log spectral-mass curves from individual training subjects (thin lines) exhibit class-dependent shared structure. Solid lines denote class spectral prototype computed exclusively from training subjects. **(b)** Spectra from held-out subjects (black) are compared with the corresponding prototypes (dashed). Here, C, S, and Ch denote class, subject, and channel, respectively.

Motivated by this coupling, we further examine how task semantics are expressed across individuals. As shown in Figure 1 (b), different subjects preserve similar spectral patterns while exhibiting systematic subject-specific deviations across frequencies. This suggests that cross-subject representations can be viewed as a relatively stable semantic anchor modulated by individual-dependent shifts. **Rather than treating cross-subject semantics as the common information remaining after individual variation is removed, we view them as shared semantic structure whose expression varies across individuals.**

This perspective suggests two key principles for cross-subject modeling. First, physiological signals contain structured task-related information in their spectral organization Buzsáki & Draguhn (2004); Donoghue et al. (2020). We therefore construct a physiology-inspired semantic anchor from these stable spectral patterns to capture shared task semantics across individuals. Second, we explore *what transferable knowledge individual variation contributes to cross-subject generalization*. Although prior work has either suppressed or exploited subject-specific information Bui et al. (2021), its functional role in generalizing to unseen subjects remains unclear. **We find that individual variation is not purely random, but contains structured and reusable components across subjects.** Moreover, in our residual analysis (Sec. 4.2) shows that cross-subject differences can be represented by a compact set of variation directions learned from training subjects that also characterize residuals from unseen subjects.

Finally, we propose SAIL, a cross-subject classification model that preserves stable task semantics while adapting their expression to individual variation. Specifically, to extract stable task semantics, we introduce Spectral Anchor–Context Pooling, which summarizes rhythmic spectral activity using band-wise mean power and power dispersion, yielding a compact task-oriented semantic anchor. We retain complementary spectral information as subject-rich context and employ shared MoE transformations Jacobs et al. (1991); Shazeer et al. (2017) to model reusable semantic shifts. A subject-conditioned router then selects and combines these shifts according to both the semantic anchor and the aggregated context of the current subject. Finally, a feature-wise gate controls the contribution of the resulting shift, which is incorporated into the semantic anchor through a residual connection to form the final representation for classification. **Across seven datasets, SAIL exceeds TeCh Yu et al. (2026) by 3.88%. On five cross-subject clinical datasets, it improves the average macro-F1 by 5.12%, while achieving 9× faster inference.** (Due to space limitations, we highlight our contributions in bold.)

2 Related Work

Existing cross-subject classification methods can be broadly divided into implicit and explicit approaches according to whether individual variation is explicitly modeled in the task representation.

Implicit Individual-Invariant Learning. Some studies rely on stronger backbones to extract transferable temporal or latent representations without explicitly modeling subject-specific factors Wang et al. (2024a); Liu et al. (2025); Yang et al. (2023). Other approaches introduce representation-level constraints Jeon et al. (2023), such as distribution matching, correlation alignment (Sun & Saenko, 2016), or contrastive objectives, to reduce the discrepancy among subjects and encourage a shared semantic space Long et al. (2015); Shen et al. (2023); Wang et al. (2024b); Tan et al. (2025). However, substantial inter-individual variation makes it difficult to learn robust shared task semantics, leading to degraded performance on unseen subjects Mahmood et al. (2019).

Explicit Individual Suppression. Another line of work explicitly utilizes subject information to suppress individual-specific representations, as shown in Figure 2. Adversarial approaches typically introduce a subject discriminator and gradient reversal layer (GRL) to encourage the encoder to produce subject-invariant representations Ganin et al. (2016); Rayatdoost et al. (2021); Huang et al. (2025). Disentanglement-based methods further decompose the latent representation into task-related and subject-related components, and impose independence or orthogonality constraints between them Bousmalis et al. (2016); Yang et al. (2026). These objectives encourage invariance or separation, but their success need not require perfect disentanglement. Their benefit depends on whether the suppressed variation is irrelevant to the task Zhao et al. (2019); Bui et al. (2021). Unlike prior suppression-based approaches, we redefine individual variation as a conditional modifier of shared task semantics rather than nuisance information to be removed.

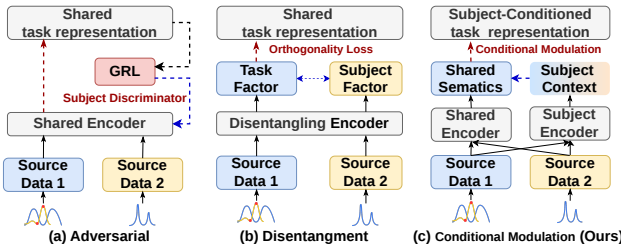

Figure 2: **Cross-subject learning paradigms.** Adversarial suppression, disentanglement, and conditional modulation.

Although prior work has incorporated subject information Nørskov et al. (2023), its role in cross-subject generalization remains insufficiently understood. Beyond SAIL, this offers a broader perspective for cross-subject learning: individual differences contain structured and reusable modes that can support controlled semantic adaptation.

3 PROBLEM FORMULATION

Cross-subject classification aims to learn a task model from a set of training subjects and generalize it to previously unseen subjects. Formally, let $\mathcal{S}_{\text{tr}}$ and $\mathcal{S}_{\text{te}}$ denote the sets of training and test subjects, respectively, with

$$\mathcal{S}_{\text{tr}} \cap \mathcal{S}_{\text{te}} = \varnothing. \tag{1}$$

Given labeled physiological signals $\{(x_i, y_i, s_i)\}$ from subjects $s_i \in \mathcal{S}_{\text{tr}}$, the goal is to learn a predictor f_θ that correctly predicts task labels for samples from subjects in $\mathcal{S}_{\text{te}}$, none of whom are observed during training.

4 UNDERSTANDING INDIVIDUAL VARIATION: COUPLED SEMANTICS AND REUSABLE STRUCTURE

4.1 TASK SEMANTICS AND INDIVIDUAL VARIATION ARE SPECTRALLY COUPLED

We first examine whether task-relevant semantics and individual variation can be separated by frequency. We conduct a probe study on APAVA for Alzheimer's disease (AD) versus healthy control (HC) classification, dividing the 1–45 Hz spectrum into six non-overlapping bands: δ (1–4 Hz), θ (4–8 Hz), α (8–13 Hz), β_{L} (13–20 Hz), β_{H} (20–30 Hz), and γ (30–45 Hz). We investigate whether task-relevant semantics and individual variation occupy separate spectral regions. As shown in Figure 3 (a), multiple frequency bands exhibit substantial task predictability. In particular, β_{L} and β_{H} achieve Macro-F1 scores of 71.85% and 75.63%, respectively. This indicates that task-relevant information is broadly distributed across the spectrum rather than being confined to a single band.

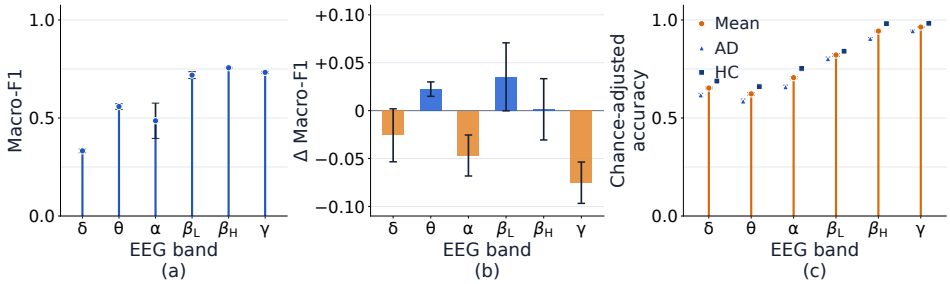

Figure 3: **Coupling Between Individual Variation and Task Semantics.** (a) Task predictability of individual frequency bands. (b) Incremental task contribution measured by removing each frequency band from the full-spectrum representation. (c) Subject identifiability within each frequency band.

We further examine whether these bands provide complementary information to the full-spectrum representation. As shown in Figure 3 (b), the contribution of different bands is highly non-uniform. Removing θ and β_{L} decreases Macro-F1 by 2.25 and 3.53 percentage points, respectively, whereas removing γ improves performance despite its strong standalone predictability. This suggests that task-predictive frequency components can be complementary, redundant, or even interfering when jointly represented.

Finally, we examine how individual information is distributed across the same frequency bands. As shown in Figure 3 (c), subject identifiability remains substantial across the spectrum and becomes particularly strong in higher-frequency bands. For example, β_{L} and β_{H} achieve subject-identification accuracies of 82.13% and 94.32%, respectively, while γ reaches 96.38%. Notably, these are also frequency regions that exhibit strong task predictability.

Motivation 1: Task-relevant semantics and individual variation share spectral support rather than occupying cleanly separated frequency regions. Individual information therefore cannot be regarded simply as removable nuisance variation.

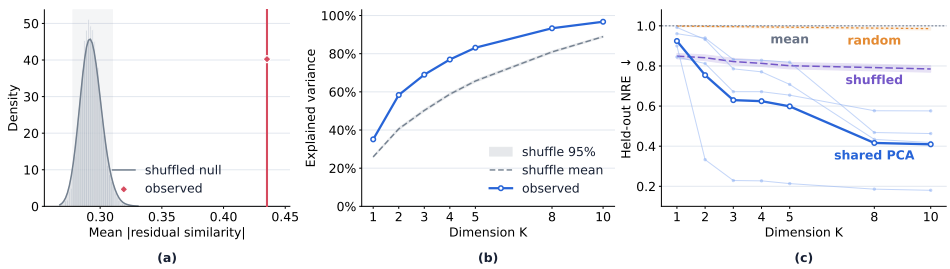

Figure 4: **Individual Variation Exhibits Reusable Structure.** (a) Structured subject variation. Pairwise similarity of subject residual spectra is significantly higher than a frequency-shuffled null distribution. (b) Compact variation space. Training-subject residuals exhibit substantially stronger low-rank structure. (c) Transfer to unseen subjects. A PCA basis learned exclusively from training subjects reconstructs held-out subject residuals better than mean, random, and shuffled baselines.

4.2 Individual Variation Exhibits Reusable Structure

We investigate whether patterns of individual variation observed in training subjects can be reused for unseen subjects. On APAVA, we first construct class-specific shared spectral prototypes using training subjects and subtract the corresponding prototype from each subject's spectrum to obtain residuals that characterize individual variation. We then calculate the mean absolute pairwise similarity between these residuals. As shown in Figure 4 (a), the observed similarity reaches 0.435, significantly exceeding the frequency-shuffled controls. This indicates that individual variation contains systematic spectral patterns rather than merely unstructured noise.

We next apply PCA (Jolliffe & Cadima, 2016) to the training-subject residuals to examine whether such variation can be represented by a compact structure. As shown in Figure 4 (b), the first three principal components explain approximately 69% of the residual variation, which is 18.7% higher

than the shuffled control. This result suggests that individual variation exhibits an approximately low-rank structure and can therefore be represented using a compact set of variation directions.

Finally, we conduct a held-out reconstruction experiment to examine whether this structure can be reused for unseen subjects. Specifically, we learn a PCA basis exclusively from training-subject residuals and use it to reconstruct residuals from held-out subjects. As shown in Figure 4 (c), the reconstruction achieves a normalized reconstruction error (NRE) of 0.41 at $K = 10$, approximately 48% lower than the shuffled-PCA baseline.

Motivation 2: Individual variation contains reusable cross-subject structure, motivating the conditional combination of shared variation modes to accommodate unseen individuals' semantic space.

TAKEAWAY

Cross-subject representation learning should preserve a **stable semantic anchor** and model individual variation through **conditional semantic adaptation**.

- Individual information is **not merely nuisance noise to be removed**; it contains structured and reusable variation across subjects.
- Such reusable individual variation can provide **contextual guidance** for how a stable shared representation should change across different individual expressions.

These observations can be summarized as,

$$z(c) = \mu + \sum_{k=1}^{K} \alpha_k(c) v_k, \qquad K \ll D, \tag{2}$$

where μ denotes the stable semantic anchor, $\{v_k\}_{k=1}^{K}$ are reusable individual variation modes shared across subjects, and $\alpha_k(c)$ determines how these modes are conditionally combined according to the semantic and individual context c of each sample.

5 METHOD

The overall architecture of SAIL is shown in Figure 5. It comprises four core modules for subject-conditioned semantic refinement. Given an time series segment, the **Temporal–Spectral Encoder (TSE)** first extracts latent features using a temporal, spectral, or dual configuration. The resulting latent spectrum is processed by **Anchor–Context Pooling (ACP)** to construct a task-oriented semantic anchor and a complementary context representation. **Subject Context Aggregation (SCA)** then summarizes the available context from each subject into a subject-level information. Guided by this prototype and the anchor's task prediction, **Conditional Expert Mixing (CEM)** combines shared context transformations and regulates their contribution through gated residual refinement. Finally, the resulting correction is added to the semantic anchor, and the refined representation is passed through the classifier shared with the anchor prediction. This design preserves an explicit semantic reference while adapting the task representation to individual context.

5.1 SPECTRAL ANCHOR–CONTEXT POOLING

Our first goal is to construct a stable semantic anchor from the shared spectral structure observed across individuals. Rather than directly using unconstrained latent features, we summarize physiology-inspired spectral quantities, including the strength and dispersion of spectral power, which characterize the rhythmic organization of biomedical signals (Buzsáki & Draguhn, 2004; Donoghue et al., 2020). These structured statistics provide the basis for the semantic anchor while complementary spectral information is retained for modeling individual variation.

Given an input window x, the Temporal–Spectral Encoder produces a latent sequence $h \in \mathbb{R}^{T \times D}$, where D is the latent feature dimension. We compute its spectrum $H = \mathcal{F}(h)$ and denote the power at frequency bin f by $P_f = |H_f|^2 \in \mathbb{R}^D$, which collects the spectral power across all latent channels.

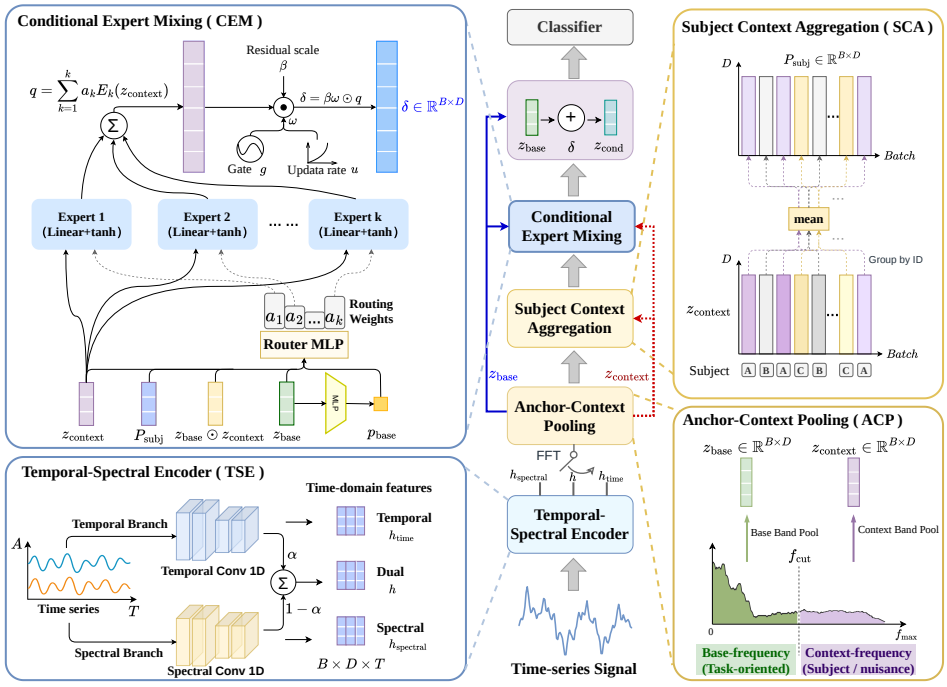

Figure 5: **Overview of SAIL**. The Temporal–Spectral Encoder (TSE) extracts latent features, and Anchor–Context Pooling (ACP) constructs a semantic anchor and complementary context. Subject Context Aggregation (SCA) forms subject-level prototypes. Conditional Expert Mixing (CEM) combines shared transformations under subject- and task-conditioned routing, producing a gated residual correction to the anchor before classification.

As illustrated in Figure 5, a cutoff f_{cut} partitions the latent spectrum into a base-frequency region Ω_b and a context-frequency region Ω_c. For each region $i \in \{\Omega_b, \Omega_c\}$, we compute the mean and standard deviation of spectral power:

$$\mu_i = \frac{1}{|i|} \sum_{f \in i} P_f, \qquad \sigma_i = \sqrt{\frac{1}{|i|} \sum_{f \in i} (P_f - \mu_i)^2 + \epsilon}. \tag{3}$$

Here, $\mu_i, \sigma_i \in \mathbb{R}^D$ summarize spectral strength and dispersion across frequency bins within each latent channel. The square and square root are applied elementwise, with ϵ added to each channel for numerical stability.

We then concatenate the log-compressed statistics and project them into a D-dimensional representation:

$$z_i = \phi_i([\log(1 + \mu_i); \log(1 + \sigma_i)]), \qquad i \in \{\Omega_b, \Omega_c\}, \tag{4}$$

where the logarithms are applied elementwise and $\phi_i : \mathbb{R}^{2D} \to \mathbb{R}^D$ is a learned projection with normalization. The resulting $z_{\Omega_b} = z_{\text{base}}$ provides the semantic anchor, while $z_{\Omega_c} = z_{\text{context}}$ retains complementary information for subject-conditioned refinement. The cutoff is dataset-dependent, and the two regions are assigned complementary modeling roles rather than constrained to contain exclusively task or subject information.

5.2 Subject-Conditioned Semantic Refinement

Our goal is to model how the stable semantic anchor should change for a particular individual. Rather than learning subject-specific parameters, we use a bank of shared context transformations. Different individuals and samples can combine these transformations differently; this is an inductive bias motivated by the residual analysis, not an explicit recovery of fixed physiological variation modes. The key is therefore to characterize the current individual's variation and use it to condition the selection of shared transformations.

Subject context aggregation. To provide the router with a more stable description of individual variation, we aggregate window-level context into a subject prototype. During training, let $\mathcal{B}_s = \{j \in \mathcal{B} \mid s_j = s\}$ denote the windows of subject s in mini-batch $\mathcal{B}$. The prototype is computed as

$$P_{\text{subj}}^{(s)} = \text{LN}\left(\frac{1}{|\mathcal{B}_s|}\sum_{j\in\mathcal{B}_s} z_{\text{context}}^{(j)}\right), \tag{5}$$

where $\text{LN}(\cdot)$ denotes layer normalization (Ba et al., 2016). The resulting prototype provides a subject-level reference that complements the local variation captured by z_{context}. Together, these two levels of context enable the router to select shared transformations according to both the individual's characteristics and the current sample. Evaluation-time aggregation is detailed in Appendix C.

Conditional expert mixing. Rather than learning subject-specific transformations, we represent individual variation using a shared bank of reusable transformations (Jacobs et al., 1991; Shazeer et al., 2017). Each expert provides a candidate direction for adapting the semantic anchor, while the router determines how these shared directions should be combined for the current sample and subject. The router combines five complementary factors. The semantic anchor z_{base} specifies the shared task representation to be adapted, while z_{context} captures sample-level variation. Their element-wise interaction $z_{\text{base}} \odot z_{\text{context}}$ explicitly characterizes how individual variation interacts with the current semantic representation. The subject prototype P_{subj} provides a more stable subject-level description of individual variation, and the task prediction p_{base} derived from z_{base} makes the transformation conditional on the current semantic hypothesis. The discussion about five factor can be found in Appendix F.

We learn K shared, bounded transformations $E_k(z) = \tanh(W_k z)$ to generate input-dependent candidate corrections. The router R_θ takes the semantic anchor, sample-level context, their interaction, the subject prototype, and the current task prediction as input, and produces a soft mixture over the shared experts:

$$a = \text{softmax}\left(\frac{R_\theta\left([z_{\text{base}}; z_{\text{context}}; z_{\text{base}} \odot z_{\text{context}}; P_{\text{subj}}; p_{\text{base}}]\right)}{\tau}\right), q = \sum_{k=1}^{K} a_k E_k(z_{\text{context}}), \tag{6}$$

where a_k denotes the routing weight of the k-th expert and τ is the routing temperature. Each expert transforms the same context representation into a candidate shift, while the router determines their subject- and task-conditioned combination.

Gated semantic refinement. The expert mixture q provides an subject-conditioned shift direction, but different feature dimensions need not contribute equally to the semantic anchor. We define an effective feature-wise gate $\omega = g \odot u \in (0,1)^D$, where $g = \text{sigmoid}(G_\theta([z_{\text{base}}; z_{\text{context}}; z_{\text{base}} \odot z_{\text{context}}]))$ is a sample-dependent gate. The shared update-rate vector is parameterized as $u = \text{sigmoid}(\eta)$, where $\eta \in \mathbb{R}^D$ is a learnable parameter vector. Rather than replacing the stable semantic anchor, the subject-conditioned shift is incorporated through a gated residual connection (He et al., 2016):

$$\delta = \beta\, \omega \odot q, \qquad z_{\text{cond}} = z_{\text{base}} + \delta, \tag{7}$$

where $0 < \beta < 0.5$ controls the overall magnitude of the residual correction. Together with bounded expert outputs, this gives $\|\delta\|_\infty < 0.5$. It bounds the absolute correction, but does not guarantee a small change relative to the anchor or improved task accuracy. The resulting z_{cond} is used for the final task classification.

A shared classifier maps z_{base} to the task prediction used for expert routing and maps z_{cond} to the final prediction. The entire model is trained end to end using task cross-entropy together with a task-supervised routing contrastive loss (Khosla et al., 2020). The complete objective and implementation details are provided in Appendix C.1.

6 Experiments

6.1 Datasets and evaluation

For a fair comparison, we follow the same training and evaluation protocols as TeCh Yu et al. (2026) across all experiments. We evaluate SAIL on seven datasets against 15 baseline methods,

Table 1: Classification results (%) across seven datasets, averaged over five seeds 41–45. F1 denotes macro-F1, EEG-2 indicates binary classification tasks based on EEG signals. A.Rank averages the 14 accuracy/F1 ranks; Gap is the mean dataset-wise shortfall from the best macro-F1 in percentage points. Bold and underlined values indicate the best and second-best results, respectively. Full metrics and standard deviations are provided in Appendix D.

Model	APAVA (EEG-2)		ADFTD (EEG-3)		TDBRAIN (EEG-2)		PTB (ECG-2)		PTB-XL (ECG-5)		FLAAP (10 classes)		UCI-HAR (6 classes)		Overall (7 datasets)	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	A.Rank	Gap
EEGNet [JNE'18]	73.50	67.52	54.56	47.05	86.35	86.33	76.49	66.57	66.21	47.66	39.09	31.31	85.65	85.61	12.50	19.87
EEG-Conformer [TNSRE'23]	74.17	72.37	45.76	44.11	87.00	86.92	86.67	84.05	70.64	57.32	49.93	47.26	90.35	90.41	9.79	12.67
Transformer [NeurIPS'17]	76.30	73.75	50.47	48.09	87.17	87.10	77.37	68.47	70.59	59.05	76.36	76.05	89.36	89.33	8.57	9.90
Reformer [ICLR'20]	78.70	75.93	50.78	47.94	87.92	87.85	77.96	69.65	71.72	60.69	71.13	70.54	88.70	88.59	8.36	9.99
Informer [AAAI'21]	73.11	69.47	48.45	45.74	89.02	88.98	78.69	70.84	71.43	60.44	72.87	72.59	88.33	88.20	9.36	10.70
Autoformer [NeurIPS'21]	68.64	68.06	45.25	42.59	87.33	87.26	73.35	63.69	61.68	48.85	38.93	33.51	41.86	32.69	14.14	27.78
FEDformer [ICML'22]	74.94	73.51	46.30	43.91	78.13	78.04	76.05	67.14	57.20	47.89	59.51	57.73	76.89	75.03	12.86	18.27
Nonstat. Tr. [NeurIPS'22]	71.89	69.74	49.95	46.96	87.88	87.78	78.66	70.90	70.56	59.10	74.85	74.39	91.04	91.01	8.71	10.18
PatchTST [ICLR'23]	67.03	55.97	44.37	41.97	79.25	79.20	74.74	64.36	73.23	62.61	56.34	55.58	87.67	88.02	12.71	17.63
iTransformer [ICLR'24]	74.55	72.30	52.60	46.79	74.67	74.65	83.89	79.06	69.28	56.20	75.15	74.91	92.41	92.39	8.93	10.69
MTST [AISTATS'24]	71.14	64.01	45.60	44.31	76.96	76.88	76.59	67.38	72.14	61.43	70.57	70.61	90.99	90.83	11.14	13.67
Medformer [NeurIPS'24]	78.74	76.31	53.27	50.65	89.62	89.62	83.50	79.18	72.87	62.02	76.44	76.25	89.62	89.62	5.50	6.78
SoftShape [ICML'25]	81.22	78.71	54.60	48.91	70.40	67.09	83.72	78.42	**73.94**	**62.97**	71.38	70.66	94.54	94.54	6.00	9.98
TeCh [ICLR'26]	85.56	84.73	54.06	47.75	92.40	92.38	86.56	82.37	73.38	62.62	80.50	80.07	93.67	93.80	3.21	3.92
BioFormer [ICML'26]	82.31	79.77	56.73	53.82	90.56	90.54	87.73	84.71	73.37	62.56	76.63	76.55	93.82	93.91	2.86	4.18
SAIL (Ours)	**87.98**	**87.41**	**62.76**	**56.33**	**98.35**	**98.35**	**92.08**	**90.65**	73.19	62.71	**80.93**	**80.73**	**94.62**	**94.69**	**1.36**	**0.04**

using five fixed seeds (41–45) for all experiments. Dataset statistics and evaluation protocols are provided in Appendix B, with complete results and standard deviations reported in Appendix D. During evaluation, SAIL conditions its predictions on unlabeled same-subject context while keeping all model parameters fixed. Additional experiments using batch-local and query-only context are reported in Appendix C.4.

6.2 Classification Performance and Inference Efficiency

Classification performance and inference efficiency. To evaluate predictive performance and inference efficiency, we compare SAIL with 15 baselines. SAIL achieves the highest macro-F1 on six of seven datasets, with an average macro-F1 of 81.55%, the best average rank of 1.36, and the lowest mean gap of 0.04 percentage points (Table 1). On APAVA at batch size 256, SAIL achieves a forward latency of 6.406 ms, providing an $8.94\times$ speedup over TeCh (57.293 ms, Appendix H). These results show that simple, physiologically grounded statistics can provide a stable semantic anchor with low computational cost. SAIL further treats individual variation as a structured modifier of shared semantics rather than nuisance information to suppress. Moreover, SAIL retains comparable performance when subject context is restricted to the query itself, indicating that the gains are not primarily driven by access to additional test-subject windows (Appendix C.4).

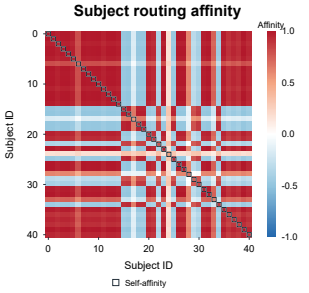

Figure 6: **Expert behavior analysis on PTB across 41 test subjects.** Outlined diagonal cells highlight distinct expert preferences for subject-specific variation modes.

Comparing paradigms for subject information. To evaluate the benefit of subject-conditioned refinement, we compare SAIL with alignment, adversarial, and disentanglement variants (Table 2). Detailed experimental settings for each paradigm are provided in Appendix E.1. SAIL exceeds the strongest alternative by 3.76, 4.37, 2.98, and 2.48 macro-F1 points on APAVA, ADFTD, TDBRAIN, and PTB, respectively. On these tasks, subject-conditioned refinement consistently outperforms other paradigms, showing the benefit of using individual variation to modulate shared task semantics.

6.3 Mechanism Analysis of Subject-Conditioned Refinement

Shared experts learn distinct correction modes. We explore whether different experts in the shared MoE bank learn distinct variation modes and whether unseen subjects combine them differently. On PTB, the output directions of the learned experts are clearly distinct: the cosine similarity is 0.559 between Expert 0 and Expert 2, but -0.881 between Expert 1 and Expert 2, indicating

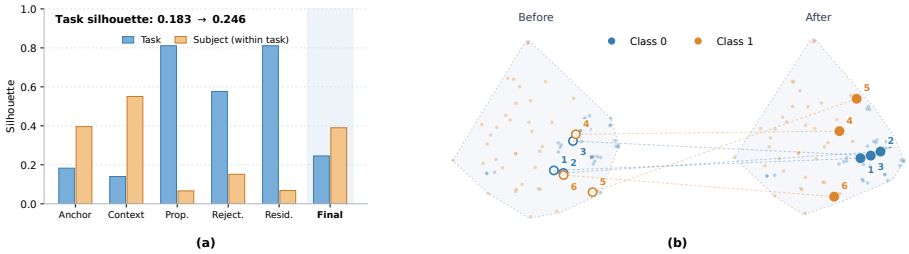

Figure 7: **Semantic geometry under individual-conditioned refinement.** (a) Task and within-task subject silhouettes for Anchor (z_{base}), Context (z_{context}), Prop. (q), Reject. (q_{rej}), Resid. (δ), and Final (z_{cond}). (b) Matched t-SNE views before and after refinement, showing boundary samples redistributed into more class-consistent regions as individual variation reshapes shared semantics.

that different experts capture different transformation directions (Appendix F.2). Moreover, routing affinities vary across 41 unseen subjects showing that different individuals favor different combinations of the shared experts, as shown in Figure 6. Together, these results suggest that SAIL learns a set of reusable variation modes and adapts their combination according to subject context.

Individual-conditioned semantic refinement. We examine whether individual variation can refine the semantic anchor, as shown in Figure 7. On PTB, the subject-rich context has a task silhouette of 0.140, whereas the conditionally selected residual reaches 0.811, showing that individual variation can be transformed into strongly task-relevant semantic directions. Incorporating this residual further increases the task silhouette from 0.183 to 0.246. The matched t-SNE views further show that samples near the original class boundary become more dispersed along class-consistent directions after refinement. Rather than removing individual variation, SAIL therefore uses it to conditionally redefine how shared task semantics are expressed. Moreover, we verify that the MoE learns reusable modes that are shared across unseen subjects, as shown in Appendix F.3.

Table 2: Subject-information, anchor–context, and fusion variants. Macro-F1 is reported as mean ± SD (%) over five runs. Full results: Appendix E.

Variant	APAVA	ADFTD	TDBRAIN	PTB
Subject-information strategies				
V1 Alignment	83.65±3.38	50.43±1.96	91.30±3.50	88.17±1.29
V2 Adversarial	80.54±2.57	47.59±2.71	95.37±3.96	86.88±2.22
V3 Disentangle.	81.25±0.79	51.96±2.45	87.63±5.84	87.72±1.75
Anchor–context variants				
V4 Residual only	75.42±2.62	52.58±3.19	99.73±0.22	88.63±1.80
V5 Dual routing	70.28±2.57	53.12±3.89	99.04±0.56	87.65±1.90
V6 Swap bands	62.09±5.50	52.32±3.04	98.83±0.34	86.44±4.95
V7 Anchor only	86.30±3.27	50.72±2.98	96.68±1.39	88.89±1.99
Fusion strategies				
V8 Concat-MLP	66.56±5.16	45.56±1.94	98.17±0.43	88.09±2.42
V9 Add-MLP	73.27±3.16	45.03±2.55	98.73±0.50	86.25±1.37
SAIL (MoE)	87.41±1.82	56.33±1.05	98.35±1.00	90.65±0.82

Anchor–context complementarity. To examine why SAIL preserves an explicit anchor and transforms context into a correction, we ablate the prediction pathway $C(z_{\text{base}}+\delta)$, where $\delta = \beta(g \odot u) \odot q$ is the gated expert correction (Table 2). Anchor only (V7) predicts from $C(z_{\text{base}})$, whereas Residual only (V4) predicts from $C(\delta)$ without adding the anchor. On APAVA, these variants achieve 86.30% and 75.42% macro-F1, respectively, compared with 87.41% for SAIL. The anchor thus provides the primary task representation, while the subject-conditioned residual provides complementary information to the semantic anchor. We next examine whether the anchor and context can be treated symmetrically or directly fused. Routing both branches (V5), which applies separate expert banks to the anchor and context, and swapping their spectral roles (V6), which exchanges the frequency regions used to construct the two representations, both degrade performance. Replacing conditional refinement with simple feature fusion (V8 and V9) leads to a similar decline. These results highlight a key design choice in SAIL, in which the anchor remains a stable semantic reference, while individual context is transformed into a conditional update rather than directly mixed with the anchor.

7 Conclusion

In this work, we rethink cross-subject classification by modeling individual variation as reusable conditional shifts rather than nuisance information to be removed. We propose *SAIL*, which combines a physiology-inspired semantic anchor with individual-conditioned MoE adaptation. Across five medical datasets, SAIL improves the average macro-F1 over the lightweight TeCh baseline by 5.12% , while achieving approximately 9× faster inference.

AI USE STATEMENT

Generative AI was used solely for language polishing and code debugging. All research ideas, experimental results, analyses, and conclusions were produced and verified by the authors.

REFERENCES

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge L. Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *Proceedings of the 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pp. 437–442, 2013.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems*, volume 29, pp. 343–351, 2016.

R. Bousseljot, D. Kreiseler, and A. Schnabel. Nutzung der EKG-signaldatenbank CARDIODAT der PTB über das Internet. *Biomedizinische Technik*, 40(s1):317–318, 1995.

Manh-Ha Bui, Toan Tran, Anh Tran, and Dinh Phung. Exploiting domain-specific features to enhance domain generalization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 21189–21201, 2021.

György Buzsáki and Andreas Draguhn. Neuronal oscillations in cortical networks. *Science*, 304 (5679):1926–1929, 2004.

Matteo Demuru and Matteo Fraschini. EEG fingerprinting: Subject-specific signature based on the aperiodic component of power spectrum. *Computers in Biology and Medicine*, 120:103748, 2020.

Thomas Donoghue, Matar Haller, Erik J. Peterson, Paroma Varma, Priyadarshini Sebastian, Richard Gao, Torben Noto, Antonio H. Lara, Joni D. Wallis, Robert T. Knight, Avgusta Shestyuk, and Bradley Voytek. Parameterizing neural power spectra into periodic and aperiodic components. *Nature Neuroscience*, 23:1655–1665, 2020.

Guikang Du, Haoran Li, Xinyu Liu, Zhibo Zhang, Xiaoli Gong, and Jin Zhang. Bioformer: Rethinking cross-subject generalization via spectral structural alignment in biomedical time-series. In *Forty-third International Conference on Machine Learning*, 2026.

J. Escudero, D. Abasolo, R. Hornero, P. Espino, and M. Lopez. Analysis of electroencephalograms in Alzheimer's disease patients with multiscale entropy. *Physiological Measurement*, 27(11): 1091–1106, 2006.

Yaroslav Ganin, E. Ustinova, Hana Ajakan, Pascal Germain, H. Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. In *Journal of machine learning research*, 2016.

Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia, and Andrew Y. Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25:65–69, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv: Learning*, 2016.

He Huang, Xiaopeng Si, Yumeng Han, and Dong Ming. A novel conditional adversarial domain adaptation network for EEG cross-subject emotion recognition. *IEEE Transactions on Affective Computing*, 16(4):2905–2917, 2025.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456. PMLR, 2015.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Eunjin Jeon, Wonjun Ko, Jee Seok Yoon, and Heung-Il Suk. Mutual information-driven subject-invariant and class-relevant deep representation learning in BCI. *IEEE Transactions on Neural Networks and Learning Systems*, 34(2):739–749, 2023.

Ian T. Jolliffe and Jorge Cadima. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020.

Prabhat Kumar and S. Suresh. FLAAP: An open human activity recognition (HAR) dataset for learning and finding the associated activity patterns. *Procedia Computer Science*, 212:64–73, 2022.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024.

Zhen Liu, Yicheng Luo, Boyuan Li, Emadeldeen Eldele, Min Wu, and Qianli Ma. Learning soft sparse shapes for efficient time-series classification. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39032–39059. PMLR, 2025.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 97–105. PMLR, 2015.

Musa Mahmood, Deogratias Mzurikwao, Yun-Soung Kim, Yongkuk Lee, Saswat Mishra, Robert Herbert, Audrey Duarte, Chee Siang Ang, and Woon-Hong Yeo. Fully portable and wireless universal brain–machine interfaces enabled by flexible scalp electronics and deep-learning algorithm. *Nature Machine Intelligence*, 1:412–422, 2019.

Andreas Miltiadous, Katerina D. Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Nikolaos Grigoriadis, Dimitrios G. Tsalikakis, Pantelis Angelidis, Markos G. Tsipouras, Euripidis Glavas, Nikolaos Giannakeas, and Alexandros T. Tzallas. A dataset of scalp EEG recordings of Alzheimer's disease, frontotemporal dementia and healthy subjects from routine EEG. *Data*, 8 (6):95, 2023.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Anders Nørskov, Alexander Neergaard Zahid, and Morten Mørup. CSLP-AE: A contrastive split-latent permutation autoencoder framework for zero-shot electroencephalography signal conversion. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Soheil Rayatdoost, Yufeng Yin, David Rudrauf, and Mohammad Soleymani. Subject-invariant EEG representation learning for emotion recognition. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3955–3959, 2021.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Xinke Shen, Xianggen Liu, Xin Hu, Dan Zhang, and Sen Song. Contrastive learning of subject-invariant EEG representations for cross-subject emotion recognition. *IEEE Transactions on Affective Computing*, 14(3):2496–2511, 2023.

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

Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision Workshops*, pp. 443–450, 2016.

Weilong Tan, Hongyi Zhang, Yingbei Wang, Weimin Wen, Liang Chen, Han Li, Xingen Gao, and Nianyin Zeng. SEDA-EEG: A semi-supervised emotion recognition network with domain adaptation for cross-subject EEG analysis. *Neurocomputing*, 622:129315, 2025.

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

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7:154, 2020.

Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A multi-granularity patching transformer for medical time-series classification. In *Advances in Neural Information Processing Systems*, volume 37, pp. 36314–36341, 2024a.

Yiming Wang, Bin Zhang, and Yujiao Tang. DMMR: Cross-subject domain generalization for EEG-based emotion recognition via denoising mixed mutual reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 628–636, 2024b.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

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

Yikun Yang, Lifei Duan, Kechen Hou, Zhongfeng Kang, Xiaowei Zhang, and Bin Hu. FDDGNet: An information bottleneck-inspired feature disentanglement network for cross-subject EEG-based emotion recognition. *Neurocomputing*, 668:132368, 2026.

Guoqi Yu, Juncheng Wang, Chen Yang, Jing Qin, Angelica I. Aviles-Rivero, and Shujun Wang. Decentralized attention fails centralized signals: Rethinking transformers for medical time series. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yitian Zhang, Liheng Ma, Soumyasundar Pal, Yingxue Zhang, and Mark Coates. Multi-resolution time-series transformer for long-term forecasting. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4222–4230, 2024.

Han Zhao, Rémi Tachet des Combes, Kun Zhang, and Geoffrey J. Gordon. On learning invariant representations for domain adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7523–7532, 2019.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

A MOTIVATION

Figure 8 extends the analysis of reusable individual variation from APAVA to PTB and ADFTD. Across both datasets, subject residual spectra exhibit higher similarity than frequency-shuffled controls, while their training-set variance concentrates in a small number of principal components. Beyond this within-training structure, a PCA basis learned from training subjects achieves lower reconstruction error on held-out subjects than the mean and random baselines. Its advantage over shuffled PCA varies with the dataset and retained subspace dimension.

Together, these findings suggest that individual variation contains shared, low-dimensional structure that extends to unseen subjects. This motivates SAIL's use of a shared bank of context transformations: rather than modeling each subject independently, SAIL combines reusable transformations in a subject-conditioned manner to refine the shared semantic anchor.

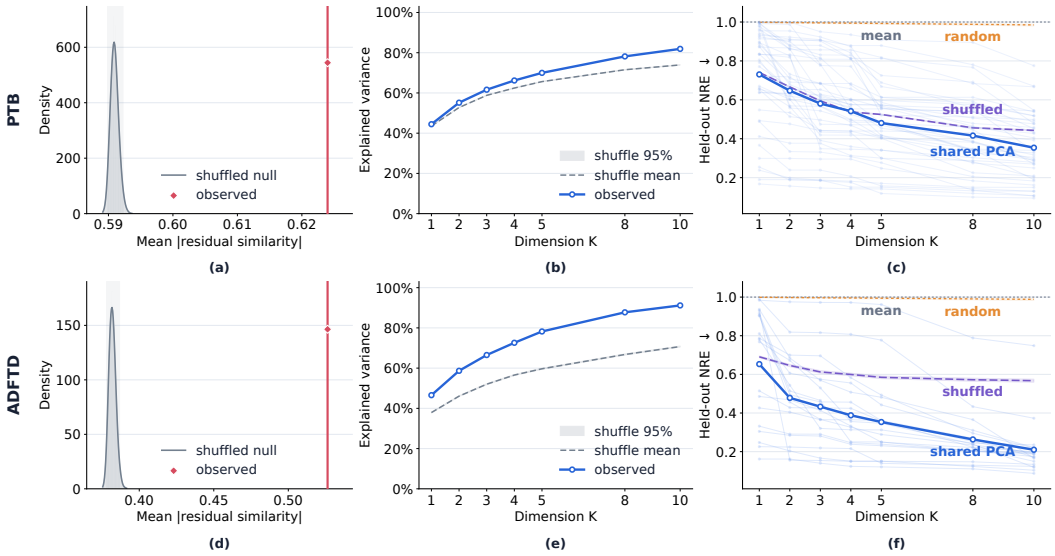

Figure 8: Reusable structure in subject-specific spectral variation on PTB (top) and ADFTD (bottom). Left: subject residual similarity compared with frequency-shuffled controls. Middle: explained variance of training residuals. Right: normalized reconstruction error on held-out subjects, comparing a training-fitted PCA basis with baseline representations.

B DATASETS AND EVALUATION PROTOCOLS

We evaluate SAIL on five clinical datasets and two supplementary activity-recognition datasets, covering EEG, ECG, and inertial signals. We use the preprocessed datasets associated with Medformer (Wang et al., 2024a) and refer to the original paper and its accompanying implementation for preprocessing details. Table 3 summarizes the dataset characteristics and subject splits.

Table 3: Dataset characteristics and subject splits. Subject counts are reported as training/validation/test. "Disjoint" indicates subject separation across all three splits; "partial" indicates disjoint test subjects only.

Dataset	Signal	Classes	Subjects	Disjoint	Labels/subject
APAVA	EEG	2	15/4/4	yes	fixed
ADFTD	EEG	3	51/18/19	yes	fixed
TDBRAIN	EEG	2	34/8/8	yes	fixed
PTB	ECG	2	118/39/41	yes	fixed
PTB-XL	ECG	5	10,556/3,519/3,521	yes	fixed
FLAAP	Inertial	10	8/8/8	no	varying
UCI-HAR	Inertial	6	21/21/9	partial	varying

Clinical datasets. The five clinical datasets cover neurological and cardiovascular classification tasks. APAVA (Escudero et al., 2006; Wang et al., 2024a) contains EEG recordings from 23 subjects for binary classification of Alzheimer's disease and healthy controls. ADFTD (Miltiadous et al., 2023) includes 88 subjects across three diagnostic categories: Alzheimer's disease, frontotemporal dementia, and healthy controls. TDBRAIN (van Dijk et al., 2022) comprises 50 subjects and 6,240 EEG windows for distinguishing Parkinson's disease from healthy controls. PTB (Bousseljot et al., 1995) includes 198 subjects for binary classification of myocardial infarction and healthy controls. PTB-XL (Wagner et al., 2020) considers five diagnostic classes: normal ECG, myocardial infarction, ST/T changes, conduction disturbances, and hypertrophy. Following the adopted preprocessing, subjects with inconsistent diagnostic labels across recordings are excluded from PTB-XL.

Activity-recognition datasets. FLAAP and UCI-HAR extend the evaluation to activity recognition from inertial signals. FLAAP (Kumar & Suresh, 2022) contains smartphone accelerometer and gyroscope recordings from eight subjects across ten activity classes. UCI-HAR (Anguita et al., 2013) includes six activity classes from 30 subjects. These datasets complement the clinical benchmarks with tasks in which activity labels vary within each subject.

Evaluation protocols. All five clinical datasets use mutually disjoint training, validation, and test subjects. APAVA and TDBRAIN use fixed subject partitions, while ADFTD, PTB, and PTB-XL use class-stratified subject-level splits with an approximate ratio of 60/20/20. For FLAAP, samples from all eight subjects are randomly partitioned into training, validation, and test sets with a ratio of 60/20/20. For UCI-HAR, we retain the official division of 21 development subjects and nine test subjects, and partition the development samples into training and validation sets with an 80/20 ratio.

C Implementation and reproducibility details

C.1 Model operations and training objective

SAIL is trained end to end with two complementary objectives: cross-entropy on the final prediction and task-supervised contrastive regularization of the routing weights (Khosla et al., 2020):

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}}(C(z_{\mathrm{cond}}), y) + \lambda \mathcal{L}_{\mathrm{route}}(a, y). \tag{8}$$

We set $\lambda = 0.05$ for all datasets except PTB-XL, where $\lambda = 0.10$.

Spectral anchor–context pooling. For an input window x_n, let $h^{(n)}$ denote the encoder output and $H^{(n)} = \mathcal{F}(h^{(n)})$ its latent spectrum. The spectral power at frequency bin f is $P_f^{(n)} = |H_f^{(n)}|^2 \in \mathbb{R}^D$. A cutoff f_{cut} partitions the latent spectrum into a base-frequency region Ω_b and a context-frequency region Ω_c. For each region $i \in \{\Omega_b, \Omega_c\}$, we compute

$$\mu_i^{(n)} = \frac{1}{|i|} \sum_{f \in i} P_f^{(n)}, \tag{9}$$

$$\sigma_i^{(n)} = \sqrt{\frac{1}{|i|} \sum_{f \in i} (P_f^{(n)} - \mu_i^{(n)})^2 + 10^{-8}}, \tag{10}$$

$$z_i^{(n)} = \phi_i \left([\log(1 + \mu_i^{(n)}); \log(1 + \sigma_i^{(n)})] \right). \tag{11}$$

Here, $\mu_i^{(n)}, \sigma_i^{(n)} \in \mathbb{R}^D$ summarize the spectral strength and dispersion across frequency bins within each latent channel. The square, square root, and logarithms are applied elementwise, with the numerical-stability constant added to each channel. Each ϕ_i consists of layer normalization (Ba et al., 2016), a $2D$-to-D linear projection, GELU (Hendrycks & Gimpel, 2016), and a final layer normalization. The resulting $z_{\Omega_b}^{(n)} = z_{\mathrm{base}}^{(n)}$ provides the semantic anchor, while $z_{\Omega_c}^{(n)} = z_{\mathrm{context}}^{(n)}$ retains complementary information for subject-conditioned refinement. The cutoff is dataset-dependent and is parameterized by the normalized cutoff ρ, measured relative to the maximum represented FFT frequency. The two regions are assigned complementary modeling roles rather than being constrained to contain exclusively task- or subject-specific information.

Spectral anchor–context pooling. For input window x_i, let $h^{(i)}$ denote the encoder output and $H^{(i)} = \mathcal{F}(h^{(i)})$ its latent spectrum. The spectral power at frequency bin f is $P_f^{(i)} = |H_f^{(i)}|^2 \in \mathbb{R}^D$.

For each band $v \in \{b, c\}$, the binary frequency mask m_{vf} is normalized as $w_{vf} = m_{vf} / \sum_{f'} m_{vf'}$. The band representation is constructed from its power statistics:

$$\mu_v^{(i)} = \sum_f w_{vf} P_f^{(i)}, \tag{12}$$

$$\sigma_v^{(i)} = \sqrt{\sum_f w_{vf} (P_f^{(i)} - \mu_v^{(i)})^2 + 10^{-8}}, \tag{13}$$

$$z_v^{(i)} = \phi_v \left([\log(1 + \mu_v^{(i)}); \log(1 + \sigma_v^{(i)})] \right). \tag{14}$$

Here, $\mu_v^{(i)}, \sigma_v^{(i)} \in \mathbb{R}^D$ collect the power statistics across latent channels; the square, square root, and logarithms are applied elementwise, with the numerical-stability constant added to each channel. The outputs for $v = b$ and $v = c$ are the semantic anchor $z_{\text{base}}^{(i)}$ and context representation $z_{\text{context}}^{(i)}$, respectively. Each ϕ_v consists of layer normalization (Ba et al., 2016), a $2D$-to-D linear projection, GELU (Hendrycks & Gimpel, 2016), and a final layer normalization. The band masks are separated by the normalized cutoff ρ, measured relative to the maximum represented FFT frequency. This pooling operation summarizes the distribution of spectral power within each band.

Subject-conditioned refinement. For a mini-batch $\mathcal{B}$, the subject prototype is

$$P_{\text{subj}}^{(s_i)} = \text{LN}\left(\frac{1}{|\mathcal{B}_{s_i}|} \sum_{j \in \mathcal{B}_{s_i}} z_{\text{context}}^{(j)} \right), \qquad \mathcal{B}_{s_i} = \{j \in \mathcal{B} : s_j = s_i\}. \tag{15}$$

The aggregation includes the query window. The router receives the semantic anchor, local context, their element-wise interaction, the subject prototype, and the anchor's task prediction:

$$\left[z_{\text{base}}^{(i)}; z_{\text{context}}^{(i)}; z_{\text{base}}^{(i)} \odot z_{\text{context}}^{(i)}; P_{\text{subj}}^{(s_i)}; p_{\text{base}}^{(i)} \right],$$

$$p_{\text{base}}^{(i)} = \text{softmax}(W_y z_{\text{base}}^{(i)} + b_y).$$

The classifier parameters W_y and b_y are shared between the anchor prediction and final classification. A temperature-scaled softmax produces the routing weights $a^{(i)}$. Each expert applies the bounded transformation $E_k(z_{\text{context}}) = \tanh(W_k z_{\text{context}})$, giving

$$q^{(i)} = \sum_{k=1}^{K} a_k^{(i)} E_k(z_{\text{context}}^{(i)}).$$

The resulting proposal is a convex mixture of bounded, input-dependent corrections.

The effective feature-wise gate is $\omega^{(i)} = g^{(i)} \odot u$, where

$$g^{(i)} = \text{sigmoid}\left(G_\theta([z_{\text{base}}^{(i)}; z_{\text{context}}^{(i)}; z_{\text{base}}^{(i)} \odot z_{\text{context}}^{(i)}]) \right), \tag{16}$$

$$u = \text{sigmoid}(\eta). \tag{17}$$

The sample-dependent gate $g^{(i)}$ controls how the proposal is incorporated, while u is the shared feature-wise update rate. The residual scale is parameterized as $\beta = 0.5\,\text{sigmoid}(\gamma)$, yielding

$$\delta^{(i)} = \beta\, \omega^{(i)} \odot q^{(i)}, \qquad z_{\text{cond}}^{(i)} = z_{\text{base}}^{(i)} + \delta^{(i)}.$$

We initialize β to 0.2 and each coordinate of u to 0.5. For representation analysis, the rejected component is defined as $q_{\text{rej}}^{(i)} = (\mathbf{1} - g^{(i)}) \odot u \odot q^{(i)}$.

Task-supervised routing contrast. We normalize each routing vector as $\widetilde{a}^{(i)} = a^{(i)} / \|a^{(i)}\|_2$. Within a mini-batch, let $\mathcal{P}(i) = \{j \neq i : y_j = y_i\}$ denote the positive set and $\mathcal{I} = \{i : |\mathcal{P}(i)| > 0\}$ the valid anchors. The routing objective is

$$\mathcal{L}_{\text{route}} = -\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \frac{1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp((\widetilde{a}^{(i)})^\top \widetilde{a}^{(p)} / \tau_c)}{\sum_{j \neq i} \exp((\widetilde{a}^{(i)})^\top \widetilde{a}^{(j)} / \tau_c)}. \tag{18}$$

Positive pairs are defined by task labels, allowing same-class samples from different subjects to contribute to shared routing structure. Differently labeled samples serve as negatives. When a mini-batch contains no valid anchors, the routing loss is set to zero.

C.2 MODEL CONFIGURATIONS

Table 4 summarizes the dataset-specific configurations of SAIL.

Table 4: SAIL configurations. All configurations use a hidden width of 128, a learning rate of 10^{-4}, at most 100 training epochs, early-stopping patience of 10, and routing temperature $\tau = 1$. Each configuration is evaluated over five runs. λ denotes the routing-loss weight.

Dataset	Batch	Encoder	Kernels	Cutoff	Experts	λ	Augmentation
APAVA	64	spectral	–	0.40	2	0.05	frequency0.1, flip0.1
ADFTD	64	spectral	–	0.10	7	0.05	drop0.1, jitter0.1, mask0.4
TDBRAIN	32	dual	5,3	0.40	3	0.05	frequency0.40
PTB	512	temporal	9,7	0.10	3	0.05	drop0.25
PTB-XL	256	temporal	15,9	0.10	16	0.10	scale0.40
FLAAP	128	temporal	11,7	0.40	7	0.05	flip0.40
UCI-HAR	32	spectral	–	0.40	3	0.05	channel0.25

Encoder architecture. The temporal encoder consists of two bias-free convolutions with channel widths $C_{\text{in}} \to D/2 \to D$, where C_{in} is the number of input channels. Each convolution is followed by batch normalization (Ioffe & Szegedy, 2015) and GELU (Hendrycks & Gimpel, 2016). The spectral encoder applies learned complex-valued channel-mixing matrices to the retained frequency modes, enforces real-valued DC and Nyquist coefficients where applicable, and returns to the time domain through an inverse FFT. The dual encoder combines the two branches as

$$h = \sigma(\alpha_{\text{stem}})h_{\text{time}} + [1 - \sigma(\alpha_{\text{stem}})]h_{\text{spectral}}.$$

A learned positive scale is constrained to $[0.5, 2]$. The resulting encoder features are passed through a separate real FFT for spectral anchor–context pooling.

Router, experts, and sample-dependent gate. With hidden dimension D and L task classes, the router has input dimension $4D + L$. It applies layer normalization followed by two linear layers mapping $4D + L \to D \to K$, with GELU between them. Each expert is initialized to the identity plus independent Gaussian perturbations with scale 0.01. All experts are evaluated, and their outputs are combined through soft routing. The sample-dependent gate $g^{(i)}$ is produced by a $3D \to D \to D$ network with GELU and an output sigmoid.

C.3 TRAINING AND INFERENCE

We optimize SAIL using Adam (Kingma & Ba, 2015) with gradient-norm clipping at 4. Early stopping is based on validation macro-F1, with a patience of 10 epochs and a tolerance of 10^{-5}. The routing contrastive loss uses a default temperature of $\tau_c = 0.2$ across all experiments.

During training, subject prototypes are computed from same-subject windows within each mini-batch, with gradients propagated through the aggregation operation. During validation and testing, windows are first encoded in mini-batches. Their band features are then concatenated, and subject prototypes are constructed from all available windows of each subject in the corresponding split. Classification is performed using these subject-level prototypes.

Subject IDs serve only as grouping keys, without introducing identity embeddings or subject-specific parameters. Neither the classifier nor the router receives target labels, and no parameter updates are performed during evaluation. Appendix C.4 evaluates the same trained models with batch-local and query-only subject prototypes.

C.4 BATCH-LOCAL SUBJECT PROTOTYPES AT TEST TIME

We examine how the amount of available subject context affects prediction by varying test-time prototype construction while keeping the five trained models per dataset fixed. The full-test reference aggregates all available unlabeled windows of the same subject. Batch-local inference instead aggregates only same-subject context within the current batch, including the query, without cross-batch caching. At $B = 1$, the prototype is constructed solely from the query. The final incomplete batch is retained, and labels are used only for scoring.

We evaluate APAVA (1,431 windows from four subjects), PTB (7,148 windows from 41 subjects), and TDBRAIN (960 windows from eight subjects) at $B \in \{1, 32, 64, 128, 512\}$. For $B > 1$,

we consider the original sample order and three random permutations with seeds 0–2. The same permutations are shared across model seeds and batch sizes, and batches are not reorganized by label or subject. Random augmentation is disabled, and the task-probability and routing temperatures are both set to 1. This yields 90 evaluations per dataset. For random-order results, we first average the three permutations within each model seed and then aggregate across the five model seeds with equal weight.

Table 5: Effect of test-time prototype scope with fixed model parameters. Macro-F1 is reported as mean ± sample SD over five model seeds (%); Δ denotes the mean paired change from the full-test reference in percentage points. Random-order permutations are averaged within each model seed.

		APAVA		PTB		TDBRAIN	
Scope	Order	F1	Δ	F1	Δ	F1	Δ
Full test	–	87.41±2.03	–	90.65±0.91	–	98.35±1.12	–
$B = 1$	Original	87.40±1.98	-0.012	89.93±1.05	-0.723	98.06±1.25	-0.292
$B = 32$	Original	87.40±2.03	-0.015	90.27±0.93	-0.383	98.31±1.15	-0.042
$B = 32$	Random	87.41±2.02	-0.005	90.25±0.98	-0.396	98.30±1.16	-0.049
$B = 64$	Original	87.40±2.03	-0.015	90.39±0.94	-0.260	98.35±1.12	0.000
$B = 64$	Random	87.40±2.03	-0.010	90.43±0.94	-0.224	98.33±1.14	-0.021
$B = 128$	Original	87.40±2.03	-0.015	90.55±0.94	-0.103	98.35±1.12	0.000
$B = 128$	Random	87.40±2.03	-0.010	90.52±0.94	-0.133	98.34±1.13	-0.014
$B = 512$	Original	87.41±2.03	0.000	90.65±0.91	+0.002	98.35±1.12	0.000
$B = 512$	Random	87.41±2.03	-0.005	90.63±0.94	-0.023	98.35±1.12	0.000

Table 5 shows that SAIL maintains strong predictive performance even without access to the full set of test windows for a subject. With query-only prototypes ($B = 1$), mean macro-F1 decreases by 0.012, 0.723, and 0.292 percentage points on APAVA, PTB, and TDBRAIN, respectively. Larger batches generally approach the full-test reference. At $B = 128$ with random sample order, the corresponding differences are -0.010, -0.133, and -0.014 percentage points. These results demonstrate the flexibility of subject-conditioned refinement under limited contextual information, while showing that additional same-subject context can further improve prediction.

C.5 Comparison methods

We compare SAIL with 15 baseline models: Transformer (Vaswani et al., 2017), Reformer (Kitaev et al., 2020), Informer (Zhou et al., 2021), Autoformer (Wu et al., 2021), FEDformer (Zhou et al., 2022), Non-stationary Transformer (Liu et al., 2022), PatchTST (Nie et al., 2023), iTransformer (Liu et al., 2024), MTST (Zhang et al., 2024), Medformer (Wang et al., 2024a), SoftShape (Liu et al., 2025), BioFormer (Du et al., 2026), TeCh (Yu et al., 2026), EEGNet (Lawhern et al., 2018), and EEG-Conformer (Song et al., 2023). All baselines are retrained and evaluated on the same dataset splits.

D Complete classification results

Table 6 reports the complete classification results of all 16 models across the five clinical and two activity-recognition datasets. For each dataset, we report accuracy, macro-precision, macro-recall, macro-F1, AUROC, and AUPRC.

The Rank and Gap columns are computed from macro-F1 within each dataset. Specifically, for model m on dataset d,

$$R_{md} = \mathrm{rank}_d(m), \qquad G_{md} = \max_j F_{jd} - F_{md}, \tag{19}$$

where F_{md} denotes the macro-F1 score of model m on dataset d. Ranks are computed among the same 16 models, with ties assigned their average rank, while Gap measures the difference from the best macro-F1 on that dataset in percentage points.

The main table (Table 1) summarizes performance across the seven datasets. A.Rank averages the 14 accuracy and macro-F1 ranks, while Gap is the equal-dataset mean of the per-dataset macro-F1

gaps:

$$\mathrm{Gap}_m = \frac{1}{|\mathcal{D}|}\sum_{d\in\mathcal{D}} G_{md} = \frac{1}{|\mathcal{D}|}\sum_{d\in\mathcal{D}}\left(\max_j F_{jd} - F_{md}\right), \qquad |\mathcal{D}| = 7. \tag{20}$$

Both A.Rank and Gap are lower-is-better, and Gap is measured in percentage points.

Table 6: Complete classification results for 16 models. Dataset blocks report percentages (mean ± SD). Rank and Gap use macro-F1 only, with lower values better. Gap is in percentage points.

Model	Acc.	Precision	Recall	F1	AUROC	AUPRC	Rank	Gap
APAVA								
SAIL	87.98±1.72	88.03±1.83	87.01±1.83	87.41±1.82	95.19±1.09	94.92±1.34	1.00	0.00
TeCh	85.56±1.43	86.01±1.42	84.22±2.03	84.73±1.68	92.73±1.91	92.48±2.02	2.00	2.68
BioFormer	82.31±1.55	87.87±0.67	78.50±1.94	79.77±2.09	88.52±2.25	88.16±2.40	3.00	7.64
SoftShape	81.22±0.62	85.65±0.22	77.55±0.87	78.71±0.91	82.24±2.43	83.93±2.19	4.00	8.70
Medformer	78.74±0.64	81.11±0.84	75.40±0.66	76.31±0.71	83.20±0.91	83.66±0.92	5.00	11.10
Reformer	78.70±2.00	82.50±3.95	75.00±1.61	75.93±1.82	73.94±1.40	76.04±1.14	6.00	11.48
Transformer	76.30±4.72	77.64±5.95	73.09±5.01	73.75±5.38	72.50±6.60	73.23±7.60	7.00	13.66
FEDformer	74.94±2.15	74.59±1.50	73.56±3.55	73.51±3.39	83.72±1.97	82.94±2.37	8.00	13.90
iTransformer	74.55±1.66	74.77±2.10	71.76±1.72	72.30±1.79	85.59±1.55	84.39±1.57	10.00	15.11
Nonstat. Tr.	71.89±3.81	71.80±4.58	69.44±3.56	69.74±3.84	70.55±2.96	70.78±4.08	11.00	17.67
Informer	73.11±4.40	75.17±6.06	69.17±4.56	69.47±5.06	70.46±4.91	70.75±5.27	12.00	17.94
Autoformer	68.64±1.82	68.48±2.10	68.77±2.27	68.06±1.94	75.94±3.61	74.38±4.05	13.00	19.35
EEGNet	73.50±8.43	77.04±12.54	68.55±9.96	67.52±12.80	78.39±14.00	79.20±14.30	14.00	19.89
MTST	71.14±1.59	79.30±0.97	65.27±2.28	64.01±3.16	68.87±2.34	71.06±1.60	15.00	23.40
EEG-Conformer	74.17±4.58	75.07±5.79	72.00±3.19	72.37±3.77	77.54±4.80	76.88±5.40	9.00	15.04
PatchTST	67.03±1.65	78.76±1.28	59.91±2.02	55.97±3.10	65.65±0.28	67.99±0.76	16.00	31.44
ADFTD								
SAIL	62.76±1.47	57.92±1.64	57.88±1.06	56.33±1.05	74.23±1.02	56.84±0.97	1.00	0.00
BioFormer	56.73±2.51	54.66±2.59	54.60±3.02	53.82±3.03	74.74±2.08	58.07±2.72	2.00	2.51
Medformer	53.27±1.54	51.02±1.57	50.71±1.55	50.65±1.51	70.93±1.19	51.21±1.32	3.00	5.68
SoftShape	54.60±0.82	50.33±1.24	49.17±0.89	48.91±1.13	70.08±0.98	50.90±1.12	4.00	7.42
Transformer	50.47±2.14	49.13±1.83	48.01±1.53	48.09±1.59	67.93±1.59	48.93±2.02	5.00	8.24
Reformer	50.78±1.17	49.64±1.49	49.89±1.67	47.94±0.69	69.17±1.58	51.73±1.94	6.00	8.39
TeCh	54.06±0.55	51.83±0.83	48.32±1.12	47.75±1.52	67.88±1.01	49.62±1.00	7.00	8.58
EEG-Conformer	45.76±1.35	44.29±1.30	44.50±1.68	44.11±1.50	63.16±1.05	45.62±0.84	13.00	12.22
EEGNet	54.56±0.95	50.73±2.22	48.12±1.04	47.05±1.50	68.46±1.64	50.79±1.97	8.00	9.28
Nonstat. Tr.	49.95±1.05	47.71±0.97	47.46±1.50	46.96±1.35	66.23±1.37	47.33±1.78	9.00	9.37
iTransformer	52.60±1.59	46.79±1.27	47.28±1.29	46.79±1.13	67.26±1.16	49.53±1.21	10.00	9.54
Informer	48.45±1.96	46.54±1.68	46.06±1.84	45.74±1.38	65.87±1.27	47.60±1.30	11.00	10.59
MTST	45.60±2.03	44.70±1.33	45.05±1.30	44.31±1.74	62.50±0.81	45.16±0.85	12.00	12.02
FEDformer	46.30±0.59	46.05±0.76	44.22±1.38	43.91±1.37	62.62±1.75	46.11±1.44	14.00	12.42
Autoformer	45.25±1.48	43.67±1.94	42.96±2.03	42.59±1.85	61.02±1.82	43.10±2.30	15.00	13.74
PatchTST	44.37±0.95	42.40±1.13	42.06±1.48	41.97±1.37	60.08±1.50	42.49±1.79	16.00	14.36
TDBRAIN								
SAIL	98.35±1.00	98.40±0.95	98.35±1.00	98.35±1.00	99.96±0.03	99.96±0.03	1.00	0.00
TeCh	92.40±0.67	92.65±0.52	92.40±0.67	92.38±0.68	98.48±0.08	98.53±0.08	2.00	5.97
BioFormer	90.56±1.72	90.92±1.70	90.56±1.72	90.54±1.73	97.34±0.85	97.40±0.77	3.00	7.81
EEG-Conformer	87.00±2.66	87.83±2.56	87.00±2.66	86.92±2.69	95.93±1.65	96.11±1.59	10.00	11.43
Medformer	89.62±0.81	89.68±0.78	89.62±0.81	89.62±0.81	96.41±0.35	96.51±0.33	4.00	8.73
Informer	89.02±2.50	89.43±2.14	89.02±2.50	88.98±2.54	96.64±0.68	96.75±0.63	5.00	9.37
Reformer	87.92±2.01	88.64±1.40	87.92±2.01	87.85±2.08	96.30±0.54	96.40±0.45	6.00	10.50
Nonstat. Tr.	87.88±2.48	88.86±1.84	87.88±2.48	87.78±2.56	97.05±0.68	96.99±0.68	7.00	10.57
Autoformer	87.33±3.79	88.06±3.56	87.33±3.79	87.26±3.84	93.81±2.26	93.32±2.42	8.00	11.09
Transformer	87.17±1.67	87.99±1.68	87.17±1.67	87.10±1.68	96.28±0.92	96.34±0.81	9.00	11.25
EEGNet	86.35±6.11	86.55±6.09	86.35±6.11	86.33±6.13	94.40±4.49	94.62±4.22	11.00	12.02
PatchTST	79.25±3.79	79.60±4.09	79.25±3.79	79.20±3.77	87.95±4.96	86.36±6.67	12.00	19.15
FEDformer	78.13±1.98	78.52±1.91	78.13±1.98	78.04±2.01	86.56±1.86	86.48±1.99	13.00	20.31
MTST	76.96±3.76	77.24±3.59	76.96±3.76	76.88±3.83	85.27±4.46	82.81±5.64	14.00	21.47
iTransformer	74.67±1.06	74.71±1.06	74.67±1.06	74.65±1.06	83.37±1.14	83.73±1.27	15.00	23.70
SoftShape	70.40±10.30	76.77±6.22	70.40±10.30	67.09±13.97	87.01±7.40	86.81±7.59	16.00	31.26
PTB								
SAIL	92.08±0.68	91.78±0.80	89.74±0.90	90.65±0.82	95.74±0.41	94.44±0.67	1.00	0.00
BioFormer	87.73±1.73	88.71±1.28	82.63±2.80	84.71±2.51	94.71±0.94	93.53±0.96	2.00	5.94
TeCh	86.56±2.17	90.29±2.09	79.50±3.32	82.37±3.23	95.83±1.34	95.37±1.56	4.00	8.28
EEG-Conformer	86.67±1.89	85.58±1.90	83.08±2.94	84.05±2.56	94.27±0.48	92.71±0.86	3.00	6.60
Medformer	83.50±2.01	85.19±0.94	77.11±3.39	79.18±3.31	92.81±1.48	90.32±1.54	5.00	11.47

Continued on next page

Table 6 (continued)

Model	Acc.	Precision	Recall	F1	AUROC	AUPRC	Rank	Gap
iTransformer	83.89±0.71	88.25±1.18	76.39±1.01	79.06±1.06	91.18±1.16	90.93±0.98	6.00	11.59
SoftShape	83.72±0.83	86.70±0.57	75.74±1.28	78.42±1.30	94.38±0.76	93.00±0.88	7.00	12.23
Nonstat. Tr.	78.66±0.49	82.77±0.86	69.12±0.87	70.90±1.00	89.37±2.51	86.67±2.38	8.00	19.75
Informer	78.69±1.68	82.87±1.02	69.19±2.90	70.84±3.47	92.09±0.53	90.02±0.60	9.00	19.81
Reformer	77.96±2.13	81.72±1.61	68.20±3.35	69.65±3.88	91.13±0.74	88.42±1.30	10.00	21.00
Transformer	77.37±1.02	81.84±0.66	67.14±1.80	68.47±2.19	90.08±1.76	87.22±1.68	11.00	22.18
MTST	76.59±1.90	79.88±1.90	66.31±2.95	67.38±3.71	86.86±2.75	83.75±2.84	12.00	23.27
FEDformer	76.05±2.54	77.58±3.61	66.10±3.55	67.14±4.37	85.93±4.31	82.59±5.42	13.00	23.51
EEGNet	76.49±1.23	77.88±1.51	65.36±1.94	66.57±2.39	92.42±1.40	88.18±1.84	14.00	24.08
PatchTST	74.74±1.62	76.94±1.51	63.89±2.71	64.36±3.38	88.79±0.91	83.39±0.96	15.00	26.29
Autoformer	73.35±2.10	72.11±2.89	63.24±3.17	63.69±3.84	78.54±3.48	74.25±3.53	16.00	26.96
PTB-XL								
EEG-Conformer	70.64±0.29	62.95±0.97	55.90±0.61	57.32±0.81	87.64±0.39	62.69±0.74	12.00	5.65
SoftShape	73.94±0.20	65.67±0.20	61.97±0.52	62.97±0.52	90.59±0.09	68.46±0.24	1.00	0.00
SAIL	73.19±0.35	64.93±0.82	61.45±0.47	62.71±0.45	90.14±0.16	66.89±0.32	2.00	0.26
TeCh	73.38±0.34	65.79±0.91	60.87±0.50	62.62±0.46	89.94±0.13	67.21±0.45	3.00	0.35
PatchTST	73.23±0.25	65.70±0.64	60.82±0.76	62.61±0.34	89.74±0.19	67.32±0.22	4.00	0.36
BioFormer	73.37±0.39	65.62±0.52	60.81±0.36	62.56±0.30	90.37±0.12	67.61±0.35	5.00	0.41
Medformer	72.87±0.23	64.14±0.42	60.60±0.46	62.02±0.37	89.66±0.13	66.39±0.22	6.00	0.95
MTST	72.14±0.27	63.84±0.72	60.01±0.81	61.43±0.38	88.97±0.33	65.83±0.51	7.00	1.54
Reformer	71.72±0.43	63.12±1.02	59.20±0.75	60.69±0.18	88.80±0.24	64.72±0.47	8.00	2.28
Informer	71.43±0.32	62.64±0.60	59.12±0.47	60.44±0.43	88.65±0.09	64.76±0.17	9.00	2.53
Nonstat. Tr.	70.56±0.55	61.57±0.66	57.75±0.72	59.10±0.66	88.32±0.36	63.40±0.79	10.00	3.87
Transformer	70.59±0.44	61.57±0.65	57.62±0.35	59.05±0.25	88.21±0.16	63.36±0.29	11.00	3.92
iTransformer	69.28±0.22	59.59±0.45	54.62±0.18	56.20±0.19	86.71±0.10	60.27±0.21	13.00	6.77
Autoformer	61.68±2.72	51.60±1.64	49.10±1.52	48.85±2.27	82.04±1.44	51.93±1.71	14.00	14.12
FEDformer	57.20±9.47	52.38±6.09	49.04±7.26	47.89±8.44	82.13±4.17	52.31±7.03	15.00	15.08
EEGNet	66.21±0.22	49.79±0.72	47.45±0.23	47.66±0.27	83.48±0.46	53.88±0.97	16.00	15.31
FLAAP								
SAIL	80.93±0.60	80.85±0.56	80.99±0.34	80.73±0.49	97.99±0.15	87.01±0.68	1.00	0.00
TeCh	80.50±0.61	80.22±0.55	80.20±0.81	80.07±0.69	97.68±0.06	86.08±0.55	2.00	0.66
BioFormer	76.63±0.39	77.31±0.43	76.55±0.58	76.55±0.45	97.07±0.09	81.76±0.58	3.00	4.18
Medformer	76.44±0.64	76.61±1.13	76.63±1.36	76.25±0.65	95.44±0.26	81.12±1.60	4.00	4.48
Transformer	76.36±1.21	76.53±1.25	76.23±0.98	76.05±1.16	96.65±0.11	80.70±0.63	5.00	4.68
iTransformer	75.15±0.48	75.09±0.53	75.14±0.47	74.91±0.51	96.64±0.14	80.81±0.60	6.00	5.82
Nonstat. Tr.	74.85±1.76	75.19±1.37	74.51±1.85	74.39±1.80	96.43±0.27	79.29±1.90	7.00	6.34
Informer	72.87±0.89	73.20±0.97	72.76±0.92	72.59±0.96	95.91±0.24	77.57±1.21	8.00	8.14
SoftShape	71.38±0.52	71.96±0.80	70.48±0.57	70.66±0.58	95.40±0.19	75.34±0.70	9.00	10.07
EEG-Conformer	49.93±0.97	55.03±2.11	48.04±1.08	47.26±1.39	87.45±0.27	48.67±0.52	14.00	33.47
MTST	70.57±0.54	71.09±0.73	70.97±0.73	70.61±0.57	94.56±0.18	73.28±0.99	10.00	10.12
Reformer	71.13±1.64	71.20±1.81	70.57±1.66	70.54±1.79	95.16±0.42	73.80±2.09	11.00	10.19
FEDformer	59.51±9.03	59.84±8.10	58.57±8.97	57.73±9.99	89.75±5.37	60.88±9.63	12.00	23.00
PatchTST	56.34±0.31	56.36±0.63	55.29±0.32	55.58±0.45	89.24±0.11	58.92±0.36	13.00	25.15
Autoformer	38.93±1.01	38.22±1.31	37.40±1.17	33.51±1.14	74.12±0.35	35.77±0.91	15.00	47.22
EEGNet	39.09±1.97	36.92±6.90	34.34±2.34	31.31±2.88	77.83±1.16	33.14±1.42	16.00	49.42
UCI-HAR								
SAIL	94.62±0.11	94.84±0.11	94.72±0.11	94.69±0.12	99.59±0.03	98.25±0.11	1.00	0.00
SoftShape	94.54±0.59	94.85±0.56	94.59±0.57	94.54±0.58	99.68±0.07	98.70±0.31	2.00	0.15
BioFormer	93.82±0.76	94.02±0.77	93.92±0.72	93.91±0.72	99.39±0.14	97.17±0.81	3.00	0.78
TeCh	93.67±0.42	93.81±0.41	93.90±0.38	93.80±0.42	99.28±0.04	96.41±0.22	4.00	0.89
EEG-Conformer	90.35±0.39	90.60±0.40	90.56±0.39	90.41±0.37	98.95±0.08	95.37±0.21	8.00	4.28
iTransformer	92.41±0.63	92.24±0.63	92.33±0.67	92.39±0.64	99.07±0.07	96.01±0.39	5.00	2.30
Nonstat. Tr.	91.04±0.58	90.98±0.60	91.14±0.56	91.01±0.60	99.02±0.09	96.07±0.37	6.00	3.68
MTST	90.99±0.84	90.96±0.79	90.92±0.85	90.83±0.88	98.21±0.11	96.14±0.59	7.00	3.86
Medformer	89.62±0.81	89.70±0.18	89.80±0.14	89.62±0.81	98.11±0.06	94.80±0.72	9.00	5.07
Transformer	89.36±1.74	89.33±1.70	89.49±1.69	89.33±1.75	98.87±0.23	95.58±0.68	10.00	5.36
Reformer	88.70±1.14	88.82±1.03	88.82±1.13	88.59±1.19	98.68±0.26	94.60±1.07	11.00	6.10
Informer	88.33±1.26	88.28±1.20	88.47±1.20	88.20±1.29	98.36±0.14	94.20±0.33	12.00	6.49
PatchTST	87.67±0.39	88.37±0.43	87.97±0.37	88.02±0.38	98.50±0.09	93.86±0.40	13.00	6.67
EEGNet	85.65±1.19	86.04±1.31	85.62±1.22	85.61±1.20	98.20±0.19	92.77±0.83	14.00	9.08
FEDformer	76.89±9.59	75.66±9.46	77.56±9.79	75.03±9.77	95.16±4.66	83.28±8.14	15.00	19.66
Autoformer	41.86±2.46	49.62±11.48	44.30±2.55	32.69±2.60	83.72±2.53	58.56±4.67	16.00	62.00

E COMPLETE ABLATIONS AND SENSITIVITY STUDIES

We investigate the contributions of SAIL's representation design, expert routing, fusion mechanism, training objectives, and input encoders through a series of ablation studies. Table 8 reports six classification metrics and test classification loss across seven datasets. Results are presented as mean $\pm$ population SD over five independent training runs. Parameter counts are expressed in millions (M), with dashes indicating unreported entries.

Table 7: Variant definitions used in the ablation studies. V1–V9 follow the order in Table 2; V10–V24 are additional controls reported in this appendix. SAIL denotes the full-model reference.

ID	Variant
Reference configuration	
SAIL	Full model: $C(z_{\text{base}} + \delta)$
Subject-information strategies	
V1	Alignment: subject covariance alignment
V2	Adversarial: gradient-reversed subject classification
V3	Disentanglement: task/private representation branches
Anchor–context variants	
V4	Residual only: $C(\delta)$
V5	Dual routing: separate expert banks for both branches
V6	Swap bands: exchange anchor and context frequency regions
V7	Anchor only: $C(z_{\text{base}})$
Fusion strategies	
V8	Concat-MLP: $C(\text{MLP}([z_{\text{base}}; z_{\text{context}}]))$
V9	Add-MLP: $C(\text{MLP}(z_{\text{base}} + z_{\text{context}}))$
Additional context control	
V10	No context proposal
Expert routing structure	
V11	One expert
V12	Hard routing
Frequency-band design	
V13	Full-band routing
V14	Full-band anchor
Router inputs	
V15	Router: anchor, context
V16	Router: + prototype
V17	Router: prototype only
V18	Router: no task probabilities
V19	Router: no interaction
Signal encoder configurations	
V20	Spectral encoder
V21	Temporal encoder
V22	Dual encoder
Training objectives	
V23	Extra negative-route separation
V24	Self/subject routing objective

E.1 CONTROL DEFINITIONS

Representation and routing controls. We ablate the prediction pathway $C(z_{\text{base}} + \delta)$, where $\delta = \beta(g \odot u) \odot q$. Residual only (V4) predicts from $C(\delta)$ without adding the anchor, while Anchor only (V7) predicts from $C(z_{\text{base}})$. Dual routing (V5) applies separate expert banks to the anchor and context, and Swap bands (V6) exchanges the frequency regions used to construct the two representations. The additional no-context-proposal control is denoted V10. Structural variants use a single expert (V11) or replace soft routing with hard argmax selection (V12). For the frequency partition, a cutoff of zero removes the dedicated lower-frequency band (V13), whereas a cutoff of one pools the full spectrum into the anchor and disables the context pathway (V14). Router-input variants (V15–V19) modify the conditioning components used for expert selection, as listed in Table 7. Task probabilities are predicted from the semantic anchor.

MLP fusion controls. Concat-MLP (V8) and Add-MLP (V9) retain the encoder architecture, frequency partition, and separate Spectral Anchor–Context Pooling modules. They predict from $C(\text{MLP}([z_{\text{base}}; z_{\text{context}}]))$ and $C(\text{MLP}(z_{\text{base}} + z_{\text{context}}))$, respectively. Both MLPs use a single

hidden layer with GELU and are trained with task cross-entropy. Direct feature fusion replaces subject prototypes, expert routing, the effective feature-wise gate ω, and residual refinement.

Shared-encoder comparisons. We compare three approaches to handling subject information: adversarial learning, distribution alignment, and representation disentanglement. These controls retain the convolutional encoder and replace the post-encoder pathway with temporal mean and standard-deviation pooling, a $2D$-to-D projection, and a task-prediction branch. Comparisons are reported on APAVA, ADFTD, TDBRAIN, PTB, FLAAP, and UCI-HAR.

Adversarial learning (V2). A subject classifier is connected to the task representation through a gradient-reversal layer (Ganin et al., 2016). The subject classifier learns to predict subject identity, while the reversed gradients encourage the encoder to reduce subject-discriminative information in the task features. The adversarial objective has a weight of 0.1.

Distribution alignment (V1). We compute subject-specific feature covariance matrices and minimize their average pairwise discrepancy, normalized by $4D^2$ (Sun & Saenko, 2016). This objective encourages subjects to share similar second-order feature statistics and is combined with task cross-entropy using a weight of 0.1.

Representation disentanglement (V3). A private subject branch complements the task branch (Bousmalis et al., 2016). An uncentered, normalized cross-second-moment penalty discourages overlap between the task and private representations. A reconstruction objective further encourages information retention by reconstructing the pooled encoder feature, with the reconstruction target detached from gradient computation. The combined auxiliary objective is weighted by 0.1.

E.2 Training objectives and input encoders

Routing objectives. The reference objective combines final-prediction cross-entropy with task-supervised routing contrast. We consider two alternatives to examine how routing supervision affects the learned expert combinations.

The additional negative-separation objective penalizes differently labeled pairs whose routing cosine similarity exceeds 0.2, using the mean squared hinge loss. The self/subject variant instead treats each routing vector as its own positive and uses samples from other subjects in the mini-batch as negatives, excluding same-subject peers. Its loss weight is 0.05, and its contrastive temperature is 0.2. Because normalized self-similarity is constant, this formulation primarily encourages separation between subjects in routing space.

Input encoders. We compare spectral, temporal, and dual temporal–spectral encoders to examine how the input representation influences subject-conditioned refinement.

E.3 Complete retraining results

Table 8: Ablation results across seven datasets. Classification metrics are reported as percentages and parameter counts in millions (M). Results are mean ± population SD over five independent training runs. SAIL denotes the recorded full-model reference, not a variant. Variants with no recorded results are omitted; dashes indicate missing individual entries. Variant IDs are defined in Table 7.

ID	Acc.	Precision	Recall	F1	AUROC	AUPRC	Params
APAVA							
SAIL	87.98±1.72	88.03±1.83	87.01±1.83	87.41±1.82	95.19±1.09	94.92±1.34	2.61
V1	84.58±3.11	85.00±3.16	83.08±3.43	83.65±3.38	92.80±2.22	92.44±2.58	–
V2	81.96±2.07	83.02±2.03	79.87±2.75	80.54±2.57	91.49±1.55	90.89±1.86	–
V3	82.54±0.66	83.53±1.40	80.51±1.01	81.25±0.79	91.76±1.02	91.17±1.26	–
V4	78.25±2.67	82.14±4.62	74.57±2.36	75.42±2.62	77.70±5.73	79.20±4.86	–
V5	74.59±1.63	79.19±2.60	70.05±2.09	70.28±2.57	84.85±1.87	84.88±1.95	–
V6	69.15±3.30	73.39±2.88	63.57±4.29	62.09±5.50	68.95±2.92	68.71±2.94	–
V7	87.13±2.76	87.83±2.03	85.76±3.54	86.30±3.27	95.29±1.13	94.98±1.40	–
V8	72.59±2.76	79.49±2.10	67.31±3.86	66.56±5.16	84.86±1.39	84.81±1.47	2.62
V9	76.66±2.53	80.49±3.64	72.66±2.80	73.27±3.16	82.94±3.44	83.77±3.34	2.62
V10	86.15±4.53	87.01±3.36	84.52±5.62	85.07±5.49	94.19±2.69	94.00±2.65	–

Continued on next page

Table 8 (continued)

ID	Acc.	Precision	Recall	F1	AUROC	AUPRC	Params
V11	83.66±2.54	84.95±2.72	81.80±3.08	82.45±2.89	93.07±2.34	92.46±3.01	–
V12	87.38±2.14	88.01±1.79	86.01±2.70	86.61±2.46	95.19±1.09	94.92±1.33	–
V13	87.11±1.77	87.63±1.96	85.73±1.89	86.37±1.87	95.11±1.11	94.87±1.24	–
V14	86.74±1.45	87.27±1.92	85.38±1.43	85.98±1.48	94.87±1.07	94.54±1.24	–
V15	85.49±2.47	86.24±3.36	83.99±1.89	84.65±2.33	93.72±2.57	93.05±3.16	–
V16	86.01±3.85	86.67±3.93	84.62±4.11	85.17±4.14	94.23±2.88	93.70±3.51	–
V17	82.61±2.12	84.20±2.45	80.49±2.62	81.19±2.54	92.79±2.34	92.11±2.93	–
V18	83.17±3.36	84.22±3.22	81.23±3.92	81.89±3.86	92.40±2.38	91.73±2.81	–
V19	81.90±1.29	84.59±2.01	79.29±2.33	80.08±1.97	92.99±2.19	92.40±2.86	–
V20	87.98±1.72	88.03±1.83	87.01±1.83	87.41±1.82	95.19±1.09	94.92±1.34	–
V21	75.74±6.83	78.01±8.77	72.37±6.90	72.96±7.37	82.90±6.54	82.91±6.81	–
V22	78.77±5.81	79.57±7.12	76.76±5.88	77.28±6.10	87.80±5.66	87.38±6.62	–
V23	88.08±1.67	88.12±1.76	87.13±1.79	87.52±1.77	95.19±1.10	94.92±1.35	–
V24	86.19±4.39	87.14±3.24	84.54±5.47	85.11±5.33	94.30±2.40	94.08±2.38	2.61
ADFTD							
SAIL	62.76±1.47	57.92±1.64	57.88±1.06	56.33±1.05	74.23±1.02	56.84±0.97	–
V1	53.19±2.09	52.36±2.14	51.15±1.59	50.43±1.96	71.10±2.08	52.34±1.90	–
V2	51.31±2.22	49.19±2.87	48.05±2.75	47.59±2.71	67.93±2.19	49.16±2.56	–
V3	54.38±3.32	53.66±3.50	52.57±2.43	51.96±2.45	72.21±2.90	54.12±3.36	–
V4	55.37±2.45	56.59±2.59	54.56±2.99	52.58±3.19	73.21±2.52	56.46±2.40	–
V5	56.84±2.97	59.46±5.43	54.26±3.01	53.12±3.89	76.06±2.18	59.75±5.14	–
V6	55.24±3.56	55.19±2.99	53.48±2.80	52.32±3.04	74.13±2.77	55.84±3.03	–
V7	53.09±3.08	52.07±3.81	51.14±2.82	50.72±2.98	69.97±2.56	51.49±3.57	–
V8	49.84±1.81	47.18±1.71	46.72±1.51	45.56±1.94	66.43±1.72	47.85±1.65	0.36
V9	50.47±1.83	48.68±2.95	46.40±2.20	45.03±2.55	67.28±3.50	48.68±3.77	0.36
V10	52.71±0.56	51.60±0.59	50.73±0.98	49.91±1.01	70.32±0.15	52.00±0.50	–
V11	52.40±2.21	51.84±2.83	50.48±2.88	49.62±3.02	70.01±1.75	51.24±2.70	–
V12	53.89±2.35	53.82±1.89	51.47±2.19	50.41±2.85	71.58±2.06	53.68±2.38	–
V13	57.90±3.84	61.79±7.17	53.08±3.86	50.67±5.02	73.94±3.32	58.64±5.35	–
V14	51.99±1.55	50.95±2.17	48.43±2.24	47.33±2.95	68.58±2.40	50.01±2.33	–
V15	53.44±1.03	52.17±1.40	51.11±1.68	50.42±2.12	70.59±0.93	53.12±1.06	–
V16	55.73±1.93	56.47±2.51	53.26±1.30	52.14±1.42	71.95±1.55	55.57±2.30	–
V17	55.32±1.88	55.63±2.80	52.85±1.54	51.77±2.10	73.40±1.14	57.61±1.70	–
V18	52.36±1.88	51.42±1.97	50.33±1.77	49.19±1.98	69.84±0.98	52.20±1.74	–
V19	52.55±4.23	52.08±5.05	49.79±4.45	48.39±4.20	69.82±2.97	52.30±4.05	–
V20	62.76±1.47	57.92±1.64	57.88±1.06	56.33±1.05	74.23±1.02	56.84±0.97	–
V21	54.93±3.43	56.67±4.59	53.04±3.20	52.48±3.36	71.59±2.94	55.67±4.98	–
V22	58.93±1.64	57.96±2.26	55.48±1.28	54.07±1.19	74.06±0.93	57.53±1.55	–
V23	54.85±2.18	56.49±2.99	52.98±2.84	51.39±3.09	72.55±1.87	55.87±2.84	–
V24	56.05±2.77	56.78±3.16	53.21±2.49	51.72±2.52	73.38±3.05	56.14±4.17	0.35
TDBRAIN							
SAIL	98.35±1.00	98.40±0.95	98.35±1.00	98.35±1.00	99.96±0.03	99.96±0.03	2.94
V1	91.33±3.48	91.93±3.26	91.33±3.48	91.30±3.50	98.14±1.56	98.19±1.49	–
V2	95.38±3.95	95.54±3.81	95.38±3.95	95.37±3.96	98.98±1.45	99.02±1.38	–
V3	87.85±5.56	89.62±3.98	87.85±5.56	87.63±5.84	97.56±1.64	97.64±1.58	–
V4	99.73±0.22	99.73±0.22	99.73±0.22	99.73±0.22	100.00±0.00	100.00±0.00	–
V5	99.04±0.56	99.05±0.55	99.04±0.56	99.04±0.56	99.94±0.06	99.94±0.06	–
V6	98.83±0.34	98.84±0.34	98.83±0.34	98.83±0.34	99.96±0.04	99.96±0.04	–
V7	96.69±1.39	96.82±1.31	96.69±1.39	96.68±1.39	99.83±0.21	99.84±0.21	–
V8	98.17±0.42	98.20±0.41	98.17±0.42	98.17±0.43	99.84±0.13	99.84±0.14	2.95
V9	98.73±0.50	98.74±0.50	98.73±0.50	98.73±0.50	99.91±0.07	99.92±0.07	2.95
V10	95.69±2.47	95.96±2.18	95.69±2.47	95.68±2.48	99.59±0.48	99.60±0.48	–
V11	96.33±1.94	96.50±1.78	96.33±1.94	96.33±1.95	99.68±0.34	99.68±0.33	–
V12	96.31±2.03	96.49±1.85	96.31±2.03	96.31±2.04	99.70±0.33	99.70±0.33	–
V13	98.56±1.23	98.58±1.20	98.56±1.23	98.56±1.23	99.91±0.14	99.92±0.13	–
V14	95.58±2.58	95.86±2.33	95.58±2.58	95.57±2.59	99.58±0.41	99.59±0.40	–
V15	96.21±2.31	96.41±2.06	96.21±2.31	96.20±2.33	99.63±0.41	99.63±0.41	–
V16	97.79±1.19	97.87±1.13	97.79±1.19	97.79±1.19	99.94±0.04	99.94±0.04	–
V17	97.19±1.38	97.29±1.32	97.19±1.38	97.19±1.38	99.89±0.11	99.89±0.11	–
V18	97.52±1.51	97.62±1.43	97.52±1.51	97.52±1.52	99.89±0.14	99.89±0.13	–
V19	97.98±1.49	98.04±1.44	97.98±1.49	97.98±1.50	99.90±0.14	99.90±0.14	–
V20	93.33±3.25	94.20±2.60	93.33±3.25	93.29±3.30	99.81±0.12	99.82±0.12	–
V21	95.71±4.85	96.28±3.98	95.71±4.85	95.67±4.93	99.25±1.46	99.27±1.43	–
V22	98.35±1.00	98.40±0.95	98.35±1.00	98.35±1.00	99.96±0.03	99.96±0.03	–
V23	97.71±1.72	97.81±1.63	97.71±1.72	97.71±1.72	99.90±0.14	99.90±0.14	–
V24	97.38±1.31	97.46±1.25	97.38±1.31	97.37±1.31	99.88±0.16	99.88±0.15	2.94
PTB							

Continued on next page

Table 8 (continued)

ID	Acc.	Precision	Recall	F1	AUROC	AUPRC	Params
SAIL	92.08±0.68	91.78±0.80	89.74±0.90	90.65±0.82	95.74±0.41	94.44±0.67	0.32
V1	90.25±0.96	90.76±0.77	86.48±1.54	88.17±1.29	95.39±0.40	93.16±0.71	–
V2	89.20±1.65	89.44±1.63	85.31±2.57	86.88±2.22	94.84±0.62	94.07±0.93	–
V3	89.94±1.28	90.58±0.98	85.93±2.02	87.72±1.75	95.24±0.84	93.47±1.95	–
V4	90.24±1.34	89.09±1.10	88.45±2.70	88.63±1.80	94.70±1.23	93.45±0.48	–
V5	89.92±1.39	90.73±0.90	85.80±2.34	87.65±1.90	95.60±0.40	93.58±1.28	–
V6	88.58±4.12	87.57±4.52	85.67±5.33	86.44±4.95	94.34±2.80	92.68±3.67	–
V7	90.71±1.55	90.56±1.40	87.70±2.35	88.89±1.99	95.24±0.52	93.80±0.98	–
V8	90.11±1.90	90.10±1.88	86.75±2.79	88.09±2.42	95.81±1.09	94.27±1.49	0.33
V9	88.86±0.95	89.85±0.61	84.20±1.62	86.25±1.37	95.64±0.64	93.96±0.92	0.33
V10	90.21±1.20	89.79±1.17	87.31±1.88	88.35±1.53	95.31±0.63	94.01±1.07	–
V11	90.22±1.33	89.99±1.08	87.12±2.19	88.31±1.76	95.29±0.72	93.87±1.24	–
V12	89.78±1.60	89.90±1.33	86.19±2.62	87.64±2.14	95.39±0.63	94.13±1.08	–
V13	90.68±1.69	91.33±1.12	86.99±2.78	88.66±2.29	95.43±0.74	93.92±1.23	–
V14	90.10±1.25	90.44±0.87	86.44±2.09	88.01±1.73	94.99±0.64	93.60±0.82	–
V15	89.79±1.46	89.42±1.40	86.69±2.30	87.81±1.88	95.15±0.64	93.87±1.31	–
V16	90.75±2.24	91.04±1.58	87.41±3.70	88.79±3.05	95.53±0.50	94.05±0.86	–
V17	91.07±1.59	90.91±1.87	88.45±2.40	89.40±1.99	95.56±0.39	94.08±1.20	–
V18	90.71±1.39	90.74±0.95	87.56±2.39	88.84±1.88	95.41±0.60	94.02±0.81	–
V19	90.35±1.45	90.67±1.14	86.80±2.40	88.32±1.98	95.54±0.62	94.06±0.93	–
V20	86.14±2.68	87.10±1.68	80.42±4.54	82.43±4.11	92.83±1.57	91.17±1.77	–
V21	92.08±0.68	91.78±0.80	89.74±0.90	90.65±0.82	95.74±0.41	94.44±0.67	–
V22	91.29±1.26	90.70±1.48	89.00±1.66	89.75±1.50	95.04±0.34	93.27±0.61	–
V23	90.40±1.66	90.51±1.10	87.07±2.83	88.42±2.31	95.18±0.56	93.55±1.10	–
V24	90.17±1.09	90.35±1.23	86.64±1.42	88.15±1.35	95.20±0.61	93.53±0.85	0.32
PTB-XL							
SAIL	73.19±0.35	64.93±0.82	61.45±0.47	62.71±0.45	90.14±0.16	66.89±0.32	0.55
V4	72.73±0.27	64.20±0.95	59.90±0.37	61.24±0.50	88.93±0.21	65.30±0.41	–
V5	73.19±0.49	64.99±1.03	60.88±0.61	62.25±0.30	90.11±0.34	66.91±0.89	–
V6	72.81±0.26	64.35±0.33	60.14±0.41	61.50±0.53	89.43±0.19	65.88±0.50	–
V7	73.07±0.39	64.88±0.87	61.18±0.24	62.36±0.45	90.15±0.20	67.03±0.48	–
V8	72.50±0.45	63.99±1.09	60.43±0.50	61.65±0.43	89.63±0.24	65.67±0.69	0.56
V9	72.53±0.39	63.92±0.76	59.69±0.75	60.99±0.25	89.71±0.19	65.75±0.37	0.56
V10	72.92±0.40	64.55±0.95	61.21±0.31	62.32±0.51	90.18±0.18	67.02±0.44	–
V11	72.96±0.45	64.87±1.07	61.13±0.39	62.38±0.50	90.14±0.19	66.97±0.49	–
V12	73.00±0.50	64.73±1.02	60.96±0.53	62.14±0.30	90.06±0.25	66.83±0.66	–
V13	73.34±0.31	65.24±1.08	60.72±0.88	62.29±0.54	90.05±0.24	66.91±0.62	–
V14	73.31±0.21	65.27±0.80	60.53±0.77	62.19±0.57	90.09±0.27	67.03±0.66	–
V15	72.94±0.44	64.67±0.89	61.08±0.65	62.25±0.57	90.08±0.28	66.81±0.58	–
V16	73.05±0.48	64.92±1.03	61.30±0.32	62.53±0.54	90.16±0.19	66.99±0.36	–
V17	73.18±0.40	65.03±0.87	61.28±0.42	62.51±0.43	90.16±0.25	67.12±0.43	–
V18	72.99±0.47	64.55±1.07	60.98±0.69	62.12±0.55	90.12±0.26	66.90±0.78	–
V19	73.14±0.52	65.07±1.10	61.30±0.59	62.52±0.60	90.20±0.13	67.04±0.39	–
V20	70.17±0.79	61.51±1.25	57.89±0.77	59.02±0.36	88.64±0.53	63.45±0.90	–
V21	73.14±0.36	64.86±0.83	61.30±0.40	62.60±0.35	90.12±0.16	66.82±0.32	–
V22	71.65±0.60	63.50±1.30	59.11±0.87	60.15±0.89	89.42±0.45	65.45±0.89	–
V23	73.10±0.47	65.03±1.07	61.03±0.56	62.30±0.12	90.12±0.11	66.83±0.32	–
V24	73.16±0.53	64.92±1.09	61.36±0.49	62.50±0.49	90.20±0.20	67.20±0.57	0.55
FLAAP							
SAIL	80.93±0.60	80.85±0.56	80.99±0.34	80.73±0.49	97.99±0.15	87.01±0.68	0.38
V1	77.94±0.81	78.10±0.88	77.90±0.84	77.68±0.90	97.26±0.11	83.02±0.79	–
V2	76.80±0.52	76.90±0.46	76.76±0.86	76.48±0.62	97.12±0.17	82.35±0.85	–
V3	77.76±0.62	78.02±0.57	77.91±0.68	77.56±0.71	97.33±0.06	83.53±0.31	–
V4	77.30±0.56	76.67±0.70	76.79±0.71	76.57±0.70	96.50±0.09	81.69±0.44	–
V5	80.04±0.72	80.01±0.74	79.89±1.08	79.81±0.90	97.76±0.07	85.68±0.42	–
V6	77.77±0.91	77.52±0.89	77.43±0.82	77.35±0.87	97.11±0.09	83.18±0.64	–
V7	77.39±1.12	77.58±1.11	77.41±1.35	77.11±1.26	97.28±0.21	82.98±1.46	–
V8	77.49±1.05	77.61±0.98	77.56±1.20	77.37±1.14	97.07±0.21	82.74±0.59	0.39
V9	77.07±1.14	77.12±1.14	77.03±1.50	76.82±1.34	97.06±0.07	82.49±0.81	0.39
V10	77.91±1.24	78.00±1.28	77.82±1.43	77.54±1.35	97.34±0.24	83.33±1.35	–
V11	77.43±2.07	77.49±1.83	77.31±2.42	77.00±2.19	97.24±0.29	82.77±1.76	–
V12	78.45±1.18	78.56±1.25	78.33±1.45	78.14±1.33	97.41±0.20	83.75±1.32	–
V13	78.55±1.37	78.60±1.39	78.17±1.56	78.11±1.47	97.61±0.37	84.99±1.70	–
V14	77.67±0.81	77.79±0.58	77.53±0.99	77.32±0.76	97.30±0.12	83.25±0.53	–
V15	78.48±0.84	78.79±0.92	78.56±1.18	78.41±1.08	97.51±0.08	84.23±0.52	–
V16	79.92±1.98	80.00±1.90	79.78±2.21	79.68±2.15	97.84±0.41	86.18±2.32	–
V17	78.85±0.75	79.10±0.71	78.93±0.83	78.72±0.78	97.60±0.16	85.04±0.78	–
V18	79.59±1.42	79.50±1.63	79.50±1.91	79.29±1.76	97.67±0.25	85.16±1.52	–

Continued on next page

Table 8 (continued)

ID	Acc.	Precision	Recall	F1	AUROC	AUPRC	Params
V19	81.21±0.47	81.34±0.57	81.44±0.61	81.18±0.58	98.03±0.10	87.15±0.53	–
V20	77.03±0.85	77.09±0.61	76.58±0.84	76.56±0.85	97.03±0.16	81.92±1.13	–
V21	80.93±0.60	80.85±0.56	80.99±0.34	80.73±0.49	97.99±0.15	87.01±0.68	–
V22	79.00±0.83	78.81±0.52	78.54±0.91	78.47±0.81	97.54±0.08	84.86±0.60	–
V23	80.98±0.60	80.99±0.62	81.06±0.47	80.77±0.52	97.88±0.13	86.31±0.71	–
V24	81.97±0.20	82.27±0.14	81.90±0.26	81.83±0.19	98.27±0.07	88.60±0.37	0.38
UCI-HAR							
SAIL	94.62±0.11	94.84±0.11	94.72±0.11	94.69±0.12	99.59±0.03	98.25±0.11	1.39
V1	96.91±0.33	96.98±0.33	96.94±0.33	96.91±0.35	99.91±0.01	99.62±0.05	–
V2	96.78±0.60	96.86±0.60	96.82±0.61	96.79±0.64	99.90±0.02	99.59±0.07	–
V3	96.65±0.44	96.72±0.46	96.69±0.44	96.65±0.46	99.91±0.01	99.61±0.04	–
V4	92.31±0.58	92.69±0.57	92.41±0.57	92.32±0.60	99.36±0.03	97.39±0.14	–
V5	92.91±0.32	93.17±0.33	93.01±0.32	92.91±0.32	99.55±0.03	98.02±0.12	–
V6	92.70±0.31	92.94±0.33	92.83±0.31	92.71±0.32	99.36±0.05	97.43±0.24	–
V7	93.51±0.84	93.66±0.86	93.67±0.80	93.61±0.83	99.46±0.10	97.59±0.45	–
V8	93.10±0.63	93.33±0.56	93.21±0.66	93.14±0.67	99.50±0.05	97.85±0.21	1.40
V9	92.98±0.32	93.15±0.32	93.11±0.32	93.02±0.32	99.48±0.04	97.75±0.18	1.40
V10	94.20±0.30	94.40±0.36	94.30±0.29	94.27±0.31	99.58±0.03	98.20±0.11	–
V11	93.84±0.17	94.06±0.11	93.95±0.18	93.90±0.18	99.55±0.08	98.11±0.22	–
V12	94.41±0.37	94.64±0.38	94.51±0.36	94.47±0.38	99.58±0.02	98.23±0.10	–
V13	94.08±1.15	94.26±1.21	94.21±1.08	94.17±1.12	99.50±0.14	97.81±0.66	–
V14	94.03±1.14	94.27±1.20	94.16±1.08	94.12±1.11	99.49±0.14	97.73±0.63	–
V15	94.13±0.43	94.37±0.41	94.21±0.43	94.16±0.44	99.58±0.04	98.21±0.12	–
V16	93.90±0.49	94.15±0.42	93.98±0.50	93.91±0.52	99.59±0.03	98.28±0.06	–
V17	93.98±0.50	94.20±0.53	94.07±0.50	94.01±0.53	99.58±0.05	98.23±0.15	–
V18	93.96±0.21	94.25±0.36	94.04±0.20	93.98±0.23	99.59±0.03	98.21±0.14	–
V19	94.50±0.39	94.70±0.39	94.60±0.40	94.56±0.41	99.58±0.04	98.17±0.16	–
V20	94.62±0.11	94.84±0.11	94.72±0.11	94.69±0.12	99.59±0.03	98.25±0.11	–
V21	90.89±0.45	90.98±0.48	91.11±0.46	90.93±0.49	99.27±0.04	96.96±0.11	–
V22	92.68±0.61	92.81±0.64	92.85±0.58	92.74±0.60	99.38±0.07	97.46±0.28	–
V23	94.46±0.45	94.69±0.47	94.56±0.45	94.53±0.48	99.58±0.04	98.22±0.14	–
V24	94.41±0.62	94.61±0.65	94.50±0.61	94.46±0.63	99.58±0.05	98.24±0.18	1.39

E.4 APAVA SINGLE-FACTOR SENSITIVITY

We examine the sensitivity of SAIL to three hyperparameters on APAVA: the spectral cutoff ρ, the number of experts K, and the routing-loss weight λ. Each factor is varied independently, with the other two held at their default values of $\rho = 0.40$, $K = 2$, and $\lambda = 0.05$. Each setting is evaluated over five independent training runs.

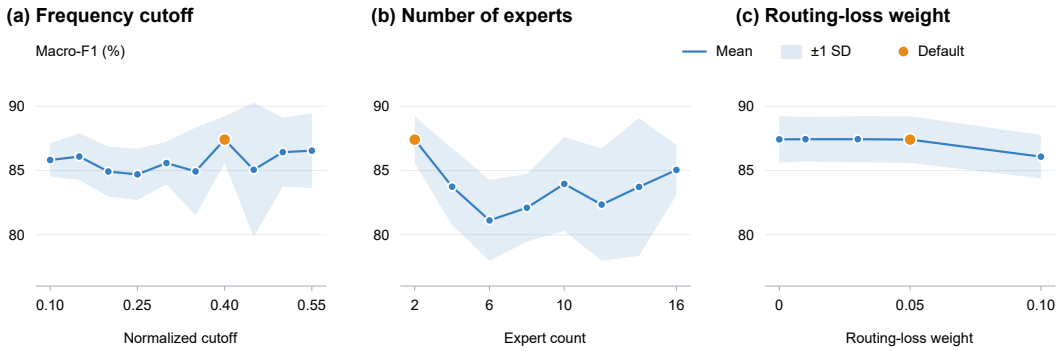

Figure 9: Sensitivity to the spectral cutoff, expert count, and routing-loss weight on APAVA. Points denote five-run means, and shaded regions indicate one population standard deviation. Orange markers identify the default settings.

Spectral cutoff. Across the evaluated cutoffs, macro-F1 ranges from 84.70% to 87.41%, with the highest value at $\rho = 0.40$. The results highlight the role of the frequency partition in balancing the semantic anchor and contextual information.

Number of experts. The two-expert configuration achieves the highest macro-F1 of 87.41%, while larger expert banks yield 81.12–85.05%. This suggests that a compact set of shared transformations is sufficient for effective subject-conditioned refinement on APAVA.

Routing-loss weight. Macro-F1 remains within 87.41–87.44% for the evaluated weights between 0 and 0.05, compared with 86.08% at $\lambda = 0.10$. Overall, the APAVA results provide stronger support for the anchor–context architecture than for the necessity of the auxiliary routing objective.

Table 9: APAVA: spectral cutoff (five-seed mean ± population SD).

Value	Acc.	Precision	Recall	F1	AUROC	AUPRC	Loss
0.1	86.58 ± 1.17	87.10 ± 1.33	85.26 ± 1.42	85.82 ± 1.29	94.92 ± 0.43	94.62 ± 0.57	0.373 ± 0.027
0.15	86.62 ± 1.75	86.41 ± 1.94	85.95 ± 1.82	86.09 ± 1.82	94.31 ± 1.22	93.82 ± 1.58	0.352 ± 0.011
0.2	85.69 ± 1.70	86.01 ± 1.67	84.53 ± 2.17	84.92 ± 1.95	93.89 ± 1.28	93.45 ± 1.60	0.429 ± 0.140
0.25	85.49 ± 1.81	85.68 ± 1.77	84.21 ± 2.10	84.70 ± 2.00	93.53 ± 1.44	93.00 ± 1.82	0.424 ± 0.116
0.3	86.18 ± 1.59	85.96 ± 1.73	85.32 ± 1.65	85.58 ± 1.66	93.68 ± 1.47	93.13 ± 1.89	0.366 ± 0.011
0.35	85.59 ± 3.28	85.44 ± 3.55	84.58 ± 3.33	84.93 ± 3.41	93.15 ± 2.66	92.73 ± 2.74	0.452 ± 0.202
0.4	87.98 ± 1.72	88.03 ± 1.83	87.01 ± 1.83	87.41 ± 1.82	95.19 ± 1.09	94.92 ± 1.34	0.338 ± 0.025
0.45	86.12 ± 4.30	86.95 ± 3.10	84.53 ± 5.39	85.06 ± 5.23	94.16 ± 2.46	93.96 ± 2.42	0.425 ± 0.179
0.5	87.27 ± 2.26	88.14 ± 1.72	85.80 ± 2.98	86.43 ± 2.69	95.40 ± 1.08	95.14 ± 1.32	0.398 ± 0.127
0.55	87.39 ± 2.44	88.33 ± 1.76	85.92 ± 3.22	86.55 ± 2.92	95.58 ± 1.13	95.36 ± 1.38	0.383 ± 0.104

Table 10: APAVA: number of experts (five-seed mean ± population SD).

Value	Acc.	Precision	Recall	F1	AUROC	AUPRC	Loss
2	87.98 ± 1.72	88.03 ± 1.83	87.01 ± 1.83	87.41 ± 1.82	95.19 ± 1.09	94.92 ± 1.34	0.338 ± 0.025
4	85.10 ± 2.42	87.19 ± 1.82	82.74 ± 3.07	83.74 ± 3.00	94.17 ± 2.49	94.07 ± 2.36	0.538 ± 0.243
6	82.70 ± 2.45	84.91 ± 2.71	80.39 ± 3.17	81.12 ± 3.14	92.65 ± 2.18	92.07 ± 2.86	0.588 ± 0.259
8	83.42 ± 2.24	84.90 ± 2.43	81.35 ± 2.72	82.10 ± 2.64	92.89 ± 2.32	92.31 ± 2.85	0.499 ± 0.166
10	85.23 ± 3.01	86.89 ± 2.22	83.10 ± 3.83	83.96 ± 3.65	94.68 ± 1.42	94.37 ± 1.72	0.434 ± 0.124
12	83.89 ± 3.11	85.88 ± 1.25	81.68 ± 4.41	82.35 ± 4.38	93.44 ± 0.99	93.10 ± 1.08	0.558 ± 0.322
14	84.98 ± 4.21	86.05 ± 2.46	83.25 ± 5.51	83.72 ± 5.37	93.43 ± 2.42	93.13 ± 2.54	0.525 ± 0.319
16	85.93 ± 1.72	86.72 ± 2.13	84.44 ± 2.10	85.05 ± 1.95	94.43 ± 1.30	94.00 ± 1.62	0.399 ± 0.112

Table 11: APAVA: routing-loss weight (five-seed mean ± population SD).

Value	Acc.	Precision	Recall	F1	AUROC	AUPRC	Loss
0	87.99 ± 1.72	88.03 ± 1.85	87.04 ± 1.81	87.43 ± 1.81	95.17 ± 1.08	94.91 ± 1.33	0.339 ± 0.026
0.01	88.01 ± 1.66	88.05 ± 1.79	87.04 ± 1.73	87.44 ± 1.74	95.17 ± 1.08	94.91 ± 1.33	0.339 ± 0.026
0.03	88.01 ± 1.70	88.06 ± 1.83	87.05 ± 1.79	87.44 ± 1.79	95.18 ± 1.08	94.92 ± 1.34	0.339 ± 0.025
0.05	87.98 ± 1.72	88.03 ± 1.83	87.01 ± 1.83	87.41 ± 1.82	95.19 ± 1.09	94.92 ± 1.34	0.338 ± 0.025
0.1	86.74 ± 1.57	86.78 ± 1.60	85.66 ± 1.78	86.08 ± 1.69	94.37 ± 1.28	93.94 ± 1.61	0.353 ± 0.012

F Routing structure and expert responses

We examine how SAIL combines shared experts through subject-relative routing patterns, expert-response geometry, and sample-level task conditioning.

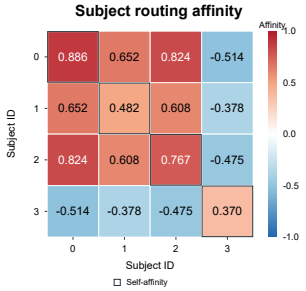

Figure 10: APAVA subject-relative routing affinity with the full router, averaged over five seeds. Positive and negative values indicate aligned and opposing preferences relative to uniform routing.

Subject-relative routing affinity. The APAVA analysis uses 1,431 test windows from four held-out subjects, two experts, and five independent training runs. In Table 12, $P_{\text{subj}}^{(s)}$ denotes the subject prototype and p_{base} denotes the anchor's predicted class-probability vector.

Table 12: APAVA router-input ablations over five runs. Accuracy and macro-F1 are reported as mean ± population SD; expert-response statistics use sample SD. Expert silhouette and proposal cosine characterize the differentiation of expert responses.

Router input	Acc.	F1	Expert sil.	Proposal cos.
$P^{(s)}_{\text{subj}}$	82.61±2.12	81.19±2.54	0.111 ± 0.139	0.895 ± 0.136
$[z_{\text{base}}; z_{\text{context}}; P^{(s)}_{\text{subj}}]$	86.01±3.85	85.17±4.14	0.193 ± 0.204	0.830 ± 0.188
$[z_{\text{base}}; z_{\text{context}}; z_{\text{base}} \odot z_{\text{context}}; P^{(s)}_{\text{subj}}]$	83.17±3.36	81.89±3.86	0.130 ± 0.162	0.885 ± 0.135
$[z_{\text{base}}; z_{\text{context}}; z_{\text{base}} \odot z_{\text{context}}; P^{(s)}_{\text{subj}}; p_{\text{base}}]$	87.98±1.72	87.41±1.82	0.148 ± 0.065	0.900 ± 0.050

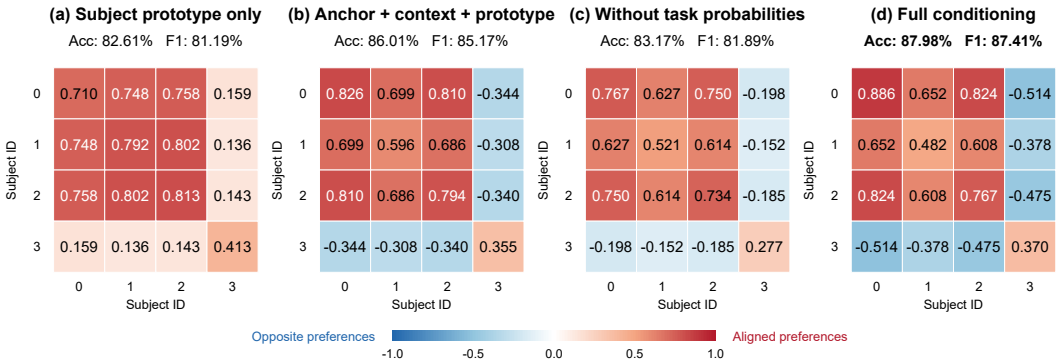

Figure 11: APAVA subject-relative routing affinities for four router-input configurations, averaged over five seeds. Panels (a)–(d) correspond to V17, V16, V18, and SAIL, respectively. All panels share a zero-centered scale from -1 to 1: blue denotes opposing preferences, white denotes zero affinity, and red denotes aligned preferences. Affinities are computed using Equation 21, with the same definition applied to the PTB analysis in Figure 6.

Let $\bar{a}_s = N_s^{-1} \sum_{i:s_i=s} a^{(i)}$ be the mean routing vector for subject s, and let $u_K = (1/K, \ldots, 1/K)$ denote uniform routing. We define the affinity between subjects s and t as

$$A_{st} = \frac{K}{K-1}(\bar{a}_s - u_K)^\top(\bar{a}_t - u_K). \tag{21}$$

For $K = 2$, this reduces to $(\bar{a}_{s0} - \bar{a}_{s1})(\bar{a}_{t0} - \bar{a}_{t1})$. The statistic is invariant to a common permutation of expert indices, allowing affinity matrices to be averaged across independently trained models.

Positive affinity indicates aligned routing preferences, whereas negative affinity indicates opposing preferences relative to uniform routing. Diagonal entries quantify the strength of each subject's expert preference. Off-diagonal affinities near zero can reflect weak preferences or variation in relative preferences across runs. The resulting matrix characterizes how different subjects draw on the same bank of transformations.

Expert-response geometry. Proposal cosine compares $\tanh(W_a z^{(i)}_{\text{context}})$ and $\tanh(W_b z^{(i)}_{\text{context}})$ for the same input and averages the resulting similarities over test windows. Expert-identity silhouette (Rousseeuw, 1987) is computed from all $1{,}431 \times 2 = 2{,}862$ candidate proposals after joint standardization, using cosine distance and expert identity as the grouping label. Together, these statistics describe the geometric differentiation of the shared transformations.

Table 12 shows that the full router achieves the highest accuracy and macro-F1 among the input variants, while the anchor/context/prototype router produces more separated expert responses. This comparison highlights the importance of selecting task-relevant corrections, rather than maximizing response diversity alone. By combining semantic, contextual, and subject-level cues, SAIL directs the shared experts toward effective refinement of the task representation.

F.1 TASK-PROBABILITY CONDITIONING AT THE SAMPLE LEVEL

The router conditions expert selection on the complete anchor prediction vector p_{base}. We characterize the confidence of this semantic hypothesis by

$$c_i = \max_{\ell} p^{(i)}_{\text{base},\ell}.$$

Figure 12 groups the 1,431 APAVA test windows into five confidence bins for each of five trained models with seeds 41–45. The analysis uses full-test unlabeled subject prototypes, with augmentation disabled and the task-prediction and routing temperatures both set to 1. Within each bin, statistics are first averaged over windows within each seed and then equally across seeds with nonempty bins. Error bars indicate sample standard deviations across seeds.

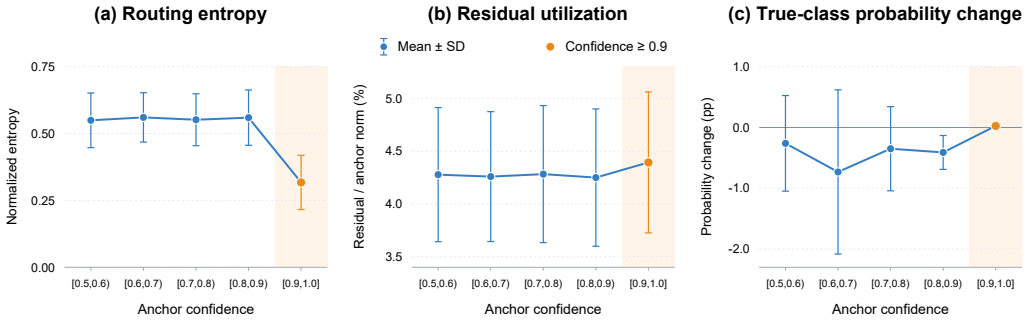

Figure 12: APAVA routing and refinement behavior by anchor confidence $\max(p_{\text{base}})$. Points and error bars denote five-seed bin means and sample standard deviations: (a) normalized routing entropy; (b) residual-to-anchor norm ratio; (c) change in true-class probability after correction. Task labels are used only to evaluate the probability change in panel (c).

For $c_i \geq 0.9$, mean normalized routing entropy is 0.317, compared with 0.550–0.560 in the lower-confidence bins. Meanwhile, the residual-to-anchor norm ratio remains between 4.25% and 4.39%. Greater semantic confidence is therefore associated with more concentrated expert selection while maintaining a comparable correction scale. The mean change in true-class probability ranges from -0.732 to -0.261 percentage points in the lower-confidence bins and reaches $+0.026$ percentage points in the highest-confidence bin.

By conditioning the routing weights, p_{base} guides the composition of the expert proposal and thereby influences the subsequent gated residual update. The effective feature-wise gate ω then regulates how this task-conditioned proposal is incorporated into the semantic anchor. This coupling connects the current semantic hypothesis to both expert selection and gated refinement, enabling task-aware, sample-specific adaptation.

F.2 PTB CORRECTION DIRECTIONS

We analyze three experts from a trained PTB model using 5,852 test windows from 41 subjects (Figure 13). An eight-dimensional PCA basis (Jolliffe & Cadima, 2016) is fitted to within-class, between-subject context variation among the training subjects and then applied to the test corrections. Each expert's directional signature is obtained by averaging its per-window unit-normalized projected corrections.

Pairwise cosines compare expert responses to the same input. For subject-macro statistics, similarities are first averaged within each subject and then across subjects with equal weight. For a given input, the expert-specific corrections use the same sample-dependent gate $g^{(i)}$, shared feature-wise update rate u, and residual scale β, allowing their transformation directions to be compared under a common refinement mechanism.

The experts exhibit distinct correction directions. E0 and E2 produce aligned corrections, with a subject-macro cosine of 0.5588, whereas E1 and E2 show strongly opposing directions, with a cosine of -0.8807. These relationships persist in the training-fitted PCA subspace, where the corresponding cosines are 0.5917 and -0.8565. E1 and E2 also exhibit high directional consistency,

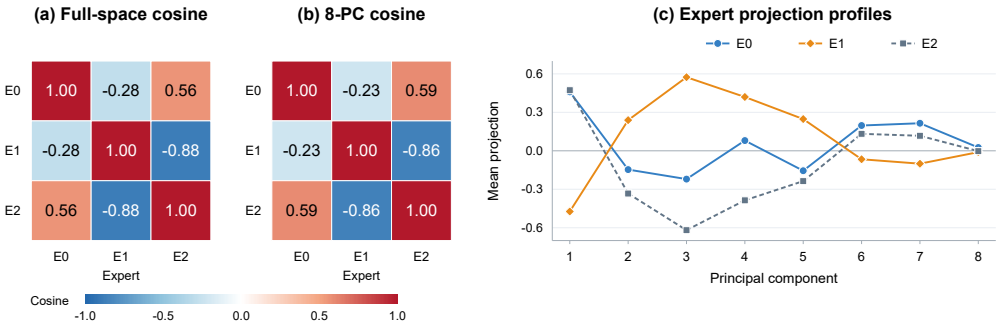

Figure 13: PTB expert correction directions: full-space cosine similarities, cosine similarities in a training-fitted eight-component PCA subspace, and mean unit-projection profiles. Cosines compare experts on the same input and weight subjects equally.

Table 13: PTB same-input expert directions. Mean cosine averages over windows, while subject-macro cosine gives equal weight to each subject. Angles are mean pairwise angles in degrees.

Space	Pair	Mean cosine	Subject-macro	Angle (°)
Proposal (full)	E0–E1	-0.2750	-0.2789	106.26
Correction (full)	E0–E1	-0.2796	-0.2835	106.55
Correction (PCA)	E0–E1	-0.2185	-0.2266	103.42
Proposal (full)	E0–E2	0.5581	0.5587	55.80
Correction (full)	E0–E2	0.5580	0.5588	55.80
Correction (PCA)	E0–E2	0.5900	0.5917	51.72
Proposal (full)	E1–E2	-0.8735	-0.8776	154.16
Correction (full)	E1–E2	-0.8768	-0.8807	154.77
Correction (PCA)	E1–E2	-0.8528	-0.8565	152.55

Table 14: PTB expert correction statistics: mean ℓ_2 norm, fraction of energy retained in the eight-dimensional PCA subspace, and direction consistency. Direction consistency is the norm of the mean unit-projection signature.

Expert	Mean norm	Energy retained	Direction consistency
E0	0.7875	0.2754	0.6305
E1	1.0539	0.2463	0.9289
E2	1.1157	0.2535	0.9766

with mean unit-projection signature norms of 0.9289 and 0.9766, respectively. Together, these patterns show that the expert bank provides reusable yet distinct correction directions that subject- and task-conditioned routing can combine to refine the semantic anchor.

F.3 Shared expert reuse and conditional routing

We examine how the shared expert pool is reused across unseen individuals. Figure 14 first visualizes the subject-wise allocation of routing weights on PTB. Each expert receives routing mass from multiple held-out subjects, while different experts exhibit distinct subject-wise allocation patterns. This indicates that the learned experts are shared across individuals rather than being tied to a one-to-one subject–expert assignment. We then test whether these shared experts can be recombined arbitrarily by replacing the native routing coefficients with same-class donor routes while keeping the query representation, expert outputs, gating, and classifier fixed. Routing substitutions close to the native combination have almost no effect, whereas increasingly mismatched combinations lead to substantial degradation: the farthest 10% substitutions reduce macro-F1 by 4.44 points, and the farthest substitutions by 6.22 points. These results suggest that SAIL reuses a common set of experts across unseen subjects, while their effective combination remains conditioned on the current individual context.

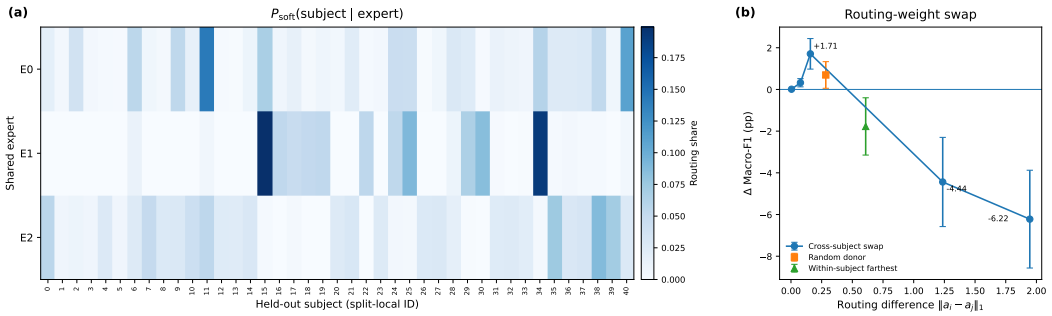

Figure 14: (a) Subject–expert routing patterns. Subject-wise allocation of routing weights across the shared expert pool on held-out PTB subjects. Each expert receives routing mass from multiple unseen subjects, while different subjects exhibit distinct expert-selection patterns. (b) Routing perturbation. Performance change after replacing the native routing coefficients with same-class donor routes at different routing distances. Small perturbations have little effect, whereas sufficiently mismatched routing combinations substantially reduce macro-F1.

G REPRESENTATION GEOMETRY

We examine how subject-conditioned refinement transforms contextual variation into task-relevant corrections and improves the geometry of the semantic anchor. Figure 15 complements the PTB analysis in Figure 7, while Table 15 summarizes task and subject structure across the representations involved in refinement.

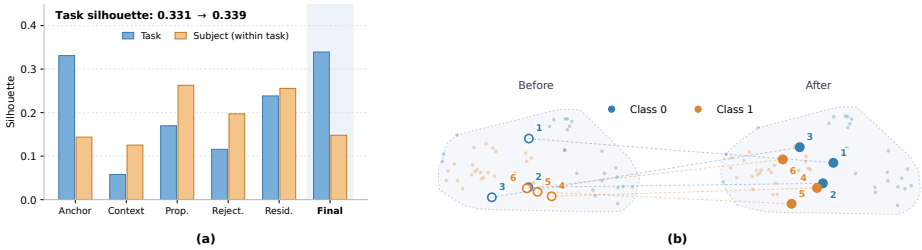

Figure 15: APAVA representation geometry, complementing Figure 7. Left: task and within-task subject silhouettes for Anchor (z_{base}), Context (z_{context}), Prop. (q), Reject. (q_{rej}), Resid. (δ), and Final (z_{cond}). Right: t-SNE views of z_{base} (Before) and z_{cond} (After), showing all 1,431 test observations colored by task label. Rings and dashed links identify the same three selected sample groups across views.

Table 15: Representation geometry on APAVA and PTB. Task silhouette (TS), within-task subject silhouette (SS), task distance gap (TG), and within-task subject distance gap (SG) are computed using standardized features and cosine distance. All representations are evaluated on the same observations: 1,431 for APAVA and 2,000 for PTB.

Dataset	Representation	TS	SS	TG	SG
APAVA	z_{base}	0.3310	0.1437	0.4204	0.0991
APAVA	z_{context}	0.0579	0.1254	0.0637	0.1261
APAVA	q	0.1698	0.2628	0.2141	0.3017
APAVA	q_{rej}	0.1157	0.1973	0.1335	0.2037
APAVA	δ	0.2383	0.2556	0.3102	0.2892
APAVA	z_{cond}	0.3393	0.1479	0.4339	0.1030
PTB	z_{base}	0.1829	0.3958	0.1931	0.5465
PTB	z_{context}	0.1400	0.5506	0.1337	0.6988
PTB	q	0.8111	0.0663	1.4950	0.1710
PTB	q_{rej}	0.5764	0.1515	0.8961	0.4128
PTB	δ	0.8109	0.0679	1.4945	0.1711
PTB	z_{cond}	0.2455	0.3900	0.2739	0.5321

Representations and evaluation metrics. We analyze six complementary representations: the semantic anchor z_{base}, context representation z_{context}, expert mixture q, rejected component q_{rej}, accepted residual δ, and refined representation z_{cond}. Following Equation equation 7, the accepted residual and refined representation are

$$\delta = \beta\, \omega \odot q, \qquad \omega = g \odot u, \qquad z_{\text{cond}} = z_{\text{base}} + \delta.$$

For the geometric analysis, we additionally define the rejected component as

$$q_{\text{rej}} = (\mathbf{1} - g) \odot u \odot q, \tag{22}$$

where $\mathbf{1} \in \mathbb{R}^D$ is the all-ones vector, g is the sample-dependent gate, and u is the shared feature-wise update rate. Their elementwise product ω is the effective feature-wise gate. Together, these representations characterize how contextual information is transformed, selected, and incorporated into the semantic anchor.

All geometry metrics use per-coordinate standardized features and cosine distances. The analysis includes all 1,431 APAVA test observations and a fixed stratified subset of 2,000 PTB test observations, with identical samples used across representations. Task silhouette measures separation by task label (Rousseeuw, 1987). Within-task subject silhouette measures subject separation within each task class, with valid class-level values weighted by sample count. Task distance gap is the mean different-task distance minus the mean same-task distance. Within-task subject distance gap similarly compares different-subject and same-subject distances within a task.

Task-relevant contextual corrections. On APAVA, the accepted residual δ exhibits a higher task silhouette than the context representation z_{context} (0.2383 versus 0.0579), alongside a higher within-task subject silhouette (0.2556 versus 0.1254). The correction therefore carries stronger task structure while retaining individual variation. On PTB, the accepted residual δ has a task silhouette of 0.8109, compared with 0.1400 for z_{context}, while its within-task subject silhouette is lower (0.0679 versus 0.5506), revealing a pronounced shift toward task-discriminative structure.

The accepted residual δ also exhibits stronger task separation than the rejected component q_{rej} on both datasets: task silhouette is 0.2383 versus 0.1157 on APAVA and 0.8109 versus 0.5764 on PTB. These patterns are consistent with selective incorporation of task-relevant contextual information.

Refinement of the semantic anchor. Adding δ to z_{base} improves task separation in the refined representation z_{cond}. Task silhouette increases from 0.3310 to 0.3393 on APAVA and from 0.1829 to 0.2455 on PTB. The task distance gap also increases, from 0.4204 to 0.4339 and from 0.1931 to 0.2739, respectively. Together, these results support the central role of subject-conditioned refinement: transforming individual variation into useful contextual updates that strengthen the shared task representation.

Matched local t-SNE groups. We visualize z_{base} and z_{cond} using t-SNE (van der Maaten & Hinton, 2008) with perplexity 30 and 1,000 iterations, displaying 1,431 APAVA windows and 1,500 PTB windows. Figures 7 and 15 highlight three local sample groups per dataset, with group membership held fixed across the two representation views. The colored rings and dashed links identify the same samples, providing a qualitative view of their local task neighborhoods before and after refinement. These matched groups complement the global geometry metrics by illustrating how contextual updates reshape local task organization within the learned representation.

H Efficiency and training dynamics

We evaluate the computational efficiency of SAIL on APAVA and examine how its gated residual mechanism evolves during training.

H.1 Inference efficiency

We compare model size, batch-forward latency, profiled operation counts, and peak CUDA memory alongside classification performance. The hardware platform is equipped with eight NVIDIA V100 GPUs with 32 GB memory each and an Intel Xeon Gold 6130 CPU at 2.10 GHz.

The APAVA benchmark uses a batch size of 256, with 20 warm-up passes followed by five timing trials. Each trial consists of 20 forward passes, and latency is reported as the mean and sample

standard deviation of the trial-average latencies. Timing covers model forward computation with batch-local subject grouping, excluding data loading and signal acquisition.

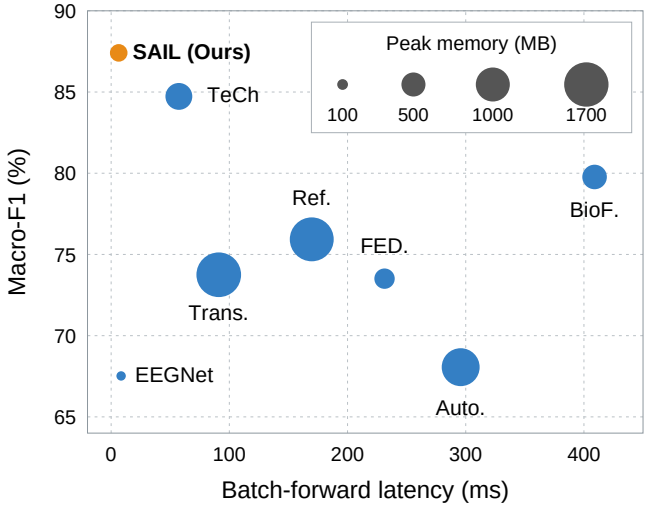

Figure 16: Classification performance and inference efficiency on APAVA for eight selected models. The horizontal axis shows batch-forward latency at a batch size of 256, the vertical axis shows macro-F1, and bubble area represents peak CUDA memory. Table 16 provides results for all 16 models.

Table 16: APAVA inference efficiency and classification performance for 16 models at batch size 256. Latency is measured in milliseconds per batch. Parameter counts and profiled operations are reported in millions (M) and billions (G), respectively. Memory denotes peak CUDA memory, and F1 denotes macro-F1 (%).

Model	Params (M)	Latency (ms)	Ops (G)	Memory	F1
SAIL	2.61	6.406 ± 0.056	0.1108	267.16	87.41 ± 1.82
TeCh	6.00	57.293 ± 0.105	206.7458	623.07	84.73 ± 1.68
BioFormer	1.36	409.004 ± 1.650	53.5577	521.01	79.77 ± 2.09
SoftShape	0.50	74.580 ± 0.775	10.0403	127.87	78.71 ± 0.91
Medformer	7.41	1005.120 ± 7.305	144.8019	764.18	76.31 ± 0.71
Reformer	0.77	169.722 ± 0.269	45.9528	1666.04	75.93 ± 1.82
Transformer	0.87	91.045 ± 0.127	52.3952	1723.19	73.75 ± 5.38
FEDformer	1.46	231.308 ± 1.338	52.0933	358.35	73.51 ± 3.39
iTransformer	0.83	56.899 ± 1.224	3.3838	39.90	72.30 ± 1.79
Nonstat. Tr.	0.94	112.294 ± 0.123	52.4204	1728.34	69.74 ± 3.84
Informer	1.07	91.304 ± 1.189	52.3952	445.27	69.47 ± 5.06
Autoformer	3.61	295.720 ± 7.878	232.4819	1242.63	68.06 ± 1.94
EEGNet	< 0.01	8.434 ± 0.508	0.5844	76.42	67.52 ± 12.80
MTST	2.59	2329.603 ± 19.835	146.6696	795.44	64.01 ± 3.16
EEG-Conformer	0.27	18.031 ± 0.572	2.8456	171.17	72.37 ± 3.77
PatchTST	0.93	777.688 ± 1.121	104.2536	735.92	55.97 ± 3.10

As shown in Table 16, SAIL achieves the highest macro-F1 of 87.41% and the lowest measured batch-forward latency of 6.406 ± 0.056 ms among the compared models. The latency ratios of EEGNet, TeCh, and BioFormer relative to SAIL are $1.32\times$, $8.94\times$, and $63.8\times$, respectively. SAIL also improves macro-F1 over TeCh and BioFormer by 2.68 and 7.64 percentage points while requiring less peak CUDA memory than either model. These results demonstrate that subject-conditioned semantic refinement combines strong classification performance with efficient batched inference.

H.2 Training dynamics

We examine the evolution of residual utilization, the sample-dependent gate, the shared feature-wise update rate, and the residual scale on APAVA, PTB, and TDBRAIN.

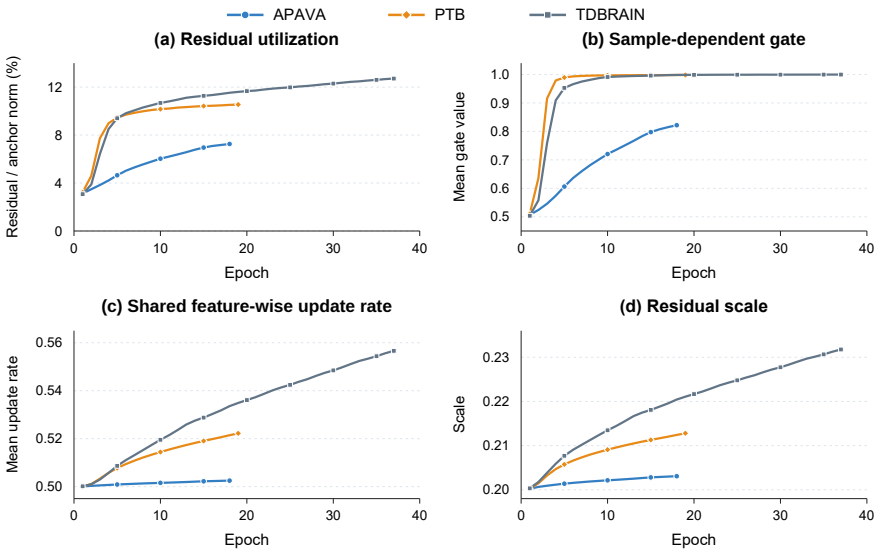

Figure 17: Training dynamics on APAVA, PTB, and TDBRAIN, with one training run shown for each dataset. Curves track epoch-mean residual utilization, the sample-dependent gate g, the shared feature-wise update rate u, and the residual scale β.

For sample i, residual utilization measures the correction magnitude relative to the semantic anchor:

$$U_i = \frac{\|\delta^{(i)}\|_2}{\|z_{\text{base}}^{(i)}\|_2 + 10^{-6}}. \tag{23}$$

The associated directional adjustment is characterized by $1 - \cos(z_{\text{base}}^{(i)}, z_{\text{cond}}^{(i)})$, where $z_{\text{cond}}^{(i)} = z_{\text{base}}^{(i)} + \delta^{(i)}$. Together, these quantities describe the relative magnitude and orientation of contextual refinement.

Figure 17 tracks residual utilization alongside the mean sample-dependent gate $g^{(i)}$, the mean shared feature-wise update rate u, and the learned residual scale β. These components regulate complementary aspects of the update: $g^{(i)}$ selects how strongly each feature accepts the contextual proposal, u learns feature-wise update rates across samples, and β controls the overall correction scale.

The sample-dependent gate and shared feature-wise update rate together form the effective feature-wise gate $\omega^{(i)}$. Through the factorized update $\delta^{(i)} = \beta(g^{(i)} \odot u) \odot q^{(i)}$, SAIL learns both which contextual corrections to apply and how strongly to incorporate them. This adaptive regulation preserves the semantic anchor as the basis of prediction while allowing subject-dependent information to refine it throughout optimization.