NEUROCONTEXT-TTA: SOURCE-ANCHORED CONTEXT CALIBRATION FOR CROSS-SUBJECT EEG EMOTION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-subject EEG emotion recognition must accommodate changes in signal statistics without labels from the new subject. Test-time adaptation can exploit recent observations, but a short, correlated history may not represent every emotion class. Building target class memories from its predictions can therefore entangle subject variation with incomplete class coverage. We propose NeuroContext-TTA, a forward-only adaptation framework that separates source class structure from target context. It translates all source prototypes by a shared context-derived offset, retaining a reference for every class without assigning target pseudo-labels. Retrieval combines similarities in the EEG embedding and encoded log-covariance spaces, and a query-level reliability gate controls residual feature and logit corrections. The calibration modules are trained on source subjects and remain frozen at deployment; only the context statistics and translated references change. Across SEED, SEED-IV, FACED-3 and FACED-9, calibration of mdJPT features improves macro-F1 by 1.35, 0.54, 5.09 and 0.52 percentage points, respectively. Evaluation with five EEG backbones further yields accuracy gains in 19 of 20 backbone–task pairs. These results support source-anchored calibration as a practical way to use unlabeled context without online parameter updates.

1 INTRODUCTION

An EEG emotion recognizer deployed to a new person encounters a familiar label space but different signal statistics. A decision rule learned across source subjects may consequently interpret similar emotional activity differently at deployment. Progress in pretrained EEG representations, including mdJPT, REVE and EMOD, has improved transfer across recordings and tasks (Zhang et al., 2025b; El Ouahidi et al., 2025; Chen et al., 2026). Yet a strong representation does not by itself specify how a frozen model should use the observations collected from its current user. This motivates a deployment question: *how can recent unlabeled EEG inform the interpretation of a query while preserving the class structure learned from source labels?*

Test-time adaptation (TTA) commonly derives an update signal from the model's target predictions. Entropy minimization is a prominent example (Wang et al., 2021); subsequent methods address unreliable samples, forgetting and unstable updates (Niu et al., 2022; 2023; Lee et al., 2024). For EEG emotion recognition, successive windows from a stimulus trial can be strongly correlated. A short history may contain only part of the label space, and repeated predictions need not constitute independent evidence for a class. Treating this history as a target class inventory risks confusing its current emotion composition with a persistent subject effect. The challenge is therefore not only to stabilize an update, but also to decide what information the unlabeled history should be allowed to change.

Memory-based adaptation offers an alternative to repeatedly optimizing the deployed model. Template adjustment, nearest-neighbor methods and retrieval-augmented adapters use reference information to refine predictions (Iwasawa & Matsuo, 2021; Jang et al., 2023; Zhang et al., 2023; Lee et al., 2025). For EEG, task embeddings and covariance descriptors provide two distinct views of a window: the former encode learned discriminative features, whereas the latter summarize statistical relationships in the signal (He & Wu, 2020; Mellot et al., 2024). Combining these views is useful

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.

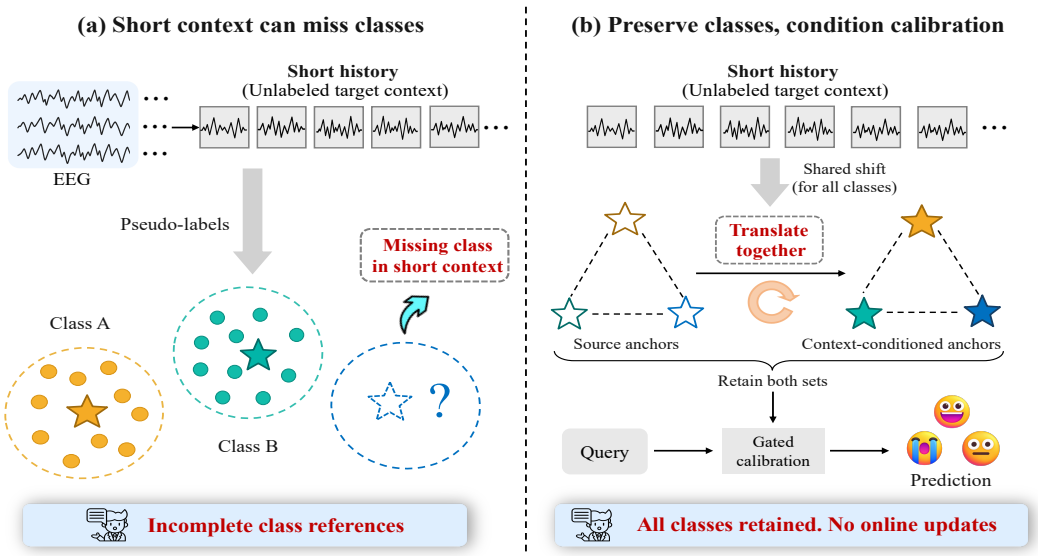


Figure 1: Motivation of NeuroContext-TTA. Source class structure and target context have different roles. (a) A short unlabeled history may omit classes; class-wise memories based on its predictions can be incomplete. (b) NeuroContext-TTA uses aggregate context statistics to translate all source prototypes together, retaining every class anchor. Stars denote class prototypes, and circles denote individual samples. This is a conceptual illustration, not an embedding visualization.

only if the reference construction remains meaningful when context class coverage is incomplete. Retrieving more observations does not, by itself, solve that problem.

We introduce NeuroContext-TTA with a simple division of responsibilities: *source labels define the class anchors; unlabeled context conditions their use*. A shared offset, estimated from the preceding windows, translates every source prototype in both descriptor spaces. This construction does not require target class assignments and preserves all pairwise prototype differences before normalization or nonlinear encoding. We retain the original prototypes alongside their translated versions, so the source reference is not overwritten. Query-to-prototype similarities in the embedding and encoded log-covariance spaces jointly determine a retrieved reference and prototype logits. A learned adapter then proposes feature and logit corrections, whose strength is controlled by the base prediction's certainty and the agreement and margins of the two reference sets.

The adaptation modules are learned with labeled source queries and unlabeled source contexts. At deployment, their parameters and the backbone classifier are fixed. Recent target windows update only the context summary and translated references. This separates learning a calibration rule from applying it to a new stream, avoiding online optimization and target pseudo-label training in the main setting. Figure 1 summarizes the distinction between context as a class inventory and context as a conditioning signal.

Our contributions are threefold. First, we formulate source-anchored context calibration that keeps every class represented without constructing pseudo-labeled target prototypes. Second, we combine dual-space prototype retrieval with query-gated feature and logit correction in a fully forward deployment procedure. Third, we evaluate the method on four emotion-recognition tasks, five EEG backbones, component ablations, context-size and context-source interventions, feature-space perturbations, and paired margin changes. The resulting evidence establishes where calibration helps and distinguishes that benefit from stronger claims about subject-specific context recovery.

2 Related Work

Cross-subject EEG representation learning. EEGNet and DGCNN model EEG with convolutional and graph-based architectures (Lawhern et al., 2018; Song et al., 2020). RGNN and EEG

Conformer extend spatial modeling and convolution–attention integration (Zhong et al., 2022; Song et al., 2023). Self-supervised encoders such as BENDR, BIOT, LaBraM and EEGPT learn reusable representations from heterogeneous recordings (Kostas et al., 2021; Yang et al., 2023; Jiang et al., 2024; Wang et al., 2024). More recent mdJPT, REVE and EMOD address broad EEG transfer or emotion-oriented pretraining (Zhang et al., 2025b; El Ouahidi et al., 2025; Chen et al., 2026). Our method operates on frozen representations and is trained separately for each backbone; it complements representation learning rather than replacing it.

Emotion transfer and distribution alignment. Domain-adversarial learning and source-free adaptation provide established approaches to distribution shift (Ganin et al., 2016; Liang et al., 2020). In EEG, MS-DCDA and SDC-Net use domain adaptation, FG-MSTGNN models frequency-dependent spatial–temporal structure, and ShareLink learns subject-common representations (Xiao et al., 2025; Tang et al., 2025; Zhang et al., 2025a; Huang et al., 2025). MMASE-DG and EmotionCLIP further study multi-view generalization and EEG–text supervision (Cheng et al., 2026; Yan et al., 2026). These methods differ in representation training and target access. We therefore distinguish complete-system comparisons from the paired question of whether calibration improves a particular frozen backbone.

Test-time optimization and reference-based adaptation. Test-time training and TTT++ learn or exploit self-supervised adaptation objectives (Sun et al., 2020; Liu et al., 2021). TENT, EATA, SAR and DeYO refine entropy-based adaptation (Wang et al., 2021; Niu et al., 2022; 2023; Lee et al., 2024); CoTTA, NOTE and RoTTA address continual or correlated streams (Wang et al., 2022; Gong et al., 2022; Yuan et al., 2023). T3A, TAST, AdaNPC and LAME instead adjust templates, use neighbors or refine predictions (Iwasawa & Matsuo, 2021; Jang et al., 2023; Zhang et al., 2023; Boudiaf et al., 2022). TDA, BoostAdapter and RA-TTA incorporate visual memory and retrieval (Karmanov et al., 2024; Zhang et al., 2024; Lee et al., 2025). Our retriever uses a fixed number of class prototypes, not a growing bank of target examples. Its target reference is obtained by a common translation rather than class-wise pseudo-label accumulation.

Covariance and forward-only EEG adaptation. Euclidean alignment and GOPSA motivate using EEG covariance information for transfer (He & Wu, 2020; Mellot et al., 2024). Our covariance branch applies a learned encoder to log-covariance vectors and uses cosine similarity; it is not geodesic nearest-neighbor search. Recent NeuroAdapt-Bench and BFT study EEG TTA and backpropagation-free deployment, while predictive-coding adaptation and ACCUP address related source-free or prototype-based settings (Lee et al., 2026; Li et al., 2026; Imtiaz & Khan, 2026; Gong et al., 2025). Forward-only computation and reliability weighting are thus not claimed as new in isolation. Our focus is how source class anchors, a shared context translation and two-view retrieval can be combined under bounded-history access.

3 Source-anchored context calibration

3.1 Setting and overview

Let $\mathcal{D}_{\mathrm{tr}}$ contain labeled source-subject windows and $\mathcal{D}_{\mathrm{val}}$ a separate source validation subject. The held-out target subject provides an ordered stream $x_1, \ldots, x_N$ with C emotion classes and no labels available to the adaptation procedure. A frozen EEG backbone produces $h_i = F_\theta(x_i) \in \mathbb{R}^d$. Each window also has a precomputed log-covariance descriptor $v_i \in \mathbb{R}^p$. A source-trained classifier g_ω maps h_i to base logits ℓ_i^0. Calibration uses these exported descriptors and does not modify the backbone.

At the start of query block t, we use the preceding K windows, $\mathcal{C}_t = \{(h_i, v_i) : t - K \le i < t\}$. All queries in that block share the context; the next block refreshes it from the preceding history. The first K windows initialize context and are excluded from scoring for both NoAdapt and the calibrated model. No query label, later window or another query from the current block enters its context summary. This information boundary applies to the calibration module operating on the ordered descriptor stream.

Figure 2 distinguishes block-level context and reference construction from query-specific retrieval and calibration. The original and translated references remain available to each query.

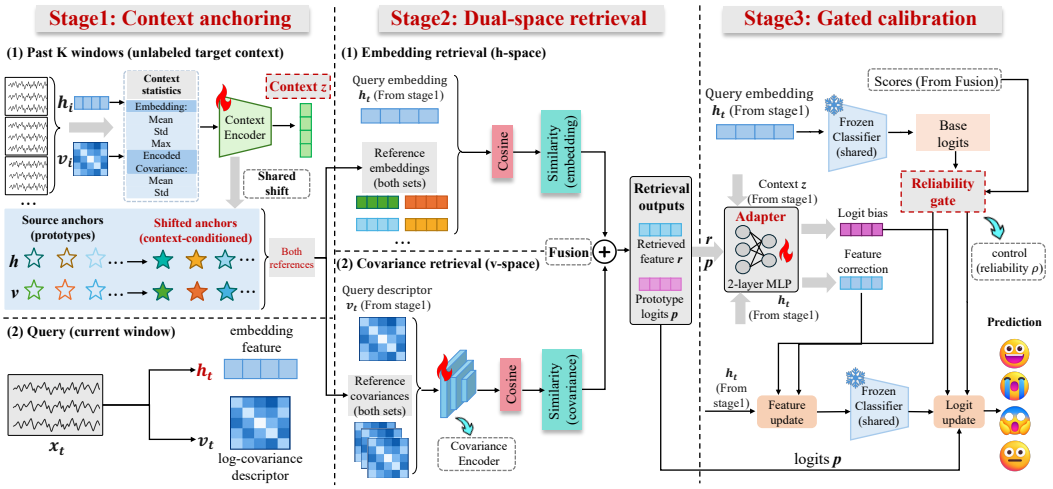


Figure 2: Forward-only context calibration. Left: context statistics define a state z and translate complete source prototype sets. Here, h denotes the backbone embedding and v denotes the log-covariance descriptor. Center: query matching combines embedding and encoded-covariance similarities, returning a retrieved feature r and prototype logits p (denoted ℓ^{proto} in the text). Right: the adapter and reliability gate operate in parallel, with the gate scaling feature and logit corrections. Flames denote modules learned during source training, while snowflakes denote frozen modules. All learned parameters remain fixed during deployment.

3.2 Complete class references from an unlabeled context

For descriptor $u \in \{h, v\}$, let p_c^u be the mean of source training descriptors with label c. Define the equally weighted center of class prototypes as $\bar{p}^u = C^{-1} \sum_c p_c^u$, and the context mean as $\bar{u}_t = K^{-1} \sum_{i=t-K}^{t-1} u_i$. We construct

$$\widetilde{p}_{t,c}^u = p_c^u + \lambda(\bar{u}_t - \bar{p}^u), \qquad u \in \{h, v\}, \quad \lambda = 0.35. \tag{1}$$

The center weights source classes equally. Every class retains an original and a translated prototype, even if absent from the context. Since the offset is shared,

$$\widetilde{p}_{t,a}^u - \widetilde{p}_{t,b}^u = p_a^u - p_b^u. \tag{2}$$

This preserves pairwise Euclidean distances in raw descriptor space, but not necessarily cosine similarities or distances after nonlinear encoding.

Context state. A learned covariance encoder $S_\eta : \mathbb{R}^p \to \mathbb{R}^{128}$ gives $e_i = S_\eta(v_i)$. We form a complementary context vector

$$z_t = E_\phi([\text{mean}(h), \text{std}(h), \max(h), \text{mean}(e), \text{std}(e)]_{\mathcal{C}_t}) \,. \tag{3}$$

The elementwise statistics use population standard deviations, with $z_t \in \mathbb{R}^{128}$. Translation uses raw descriptor means; the encoder also captures embedding variability and learned covariance statistics. Pooling is permutation invariant.

3.3 Dual-space retrieval and reliability-guided correction

For each class, source and context-conditioned retrieval scores are

$$\begin{aligned} a_c^{\text{src}} &= \alpha \, \cos(h_q, p_c^h) + (1 - \alpha) \, \cos(S_\eta(v_q), S_\eta(p_c^v)), \\ a_c^{\text{ctx}} &= \alpha \, \cos(h_q, \widetilde{p}_{t,c}^h) + (1 - \alpha) \, \cos(S_\eta(v_q), S_\eta(\widetilde{p}_{t,c}^v)), \end{aligned} \tag{4}$$

where $\alpha = \text{sigmoid}(a)$ is learned during source training. Covariance prototypes are averaged and translated *before* the nonlinear encoder. With $w^j = \text{softmax}(a^j/\tau)$ for $j \in \{\text{src}, \text{ctx}\}$, retrieval returns

$$r_q = \tfrac{1}{2} \sum_c \left(w_c^{\text{src}} p_c^h + w_c^{\text{ctx}} \widetilde{p}_{t,c}^h \right), \qquad \ell_q^{\text{proto}} = (a^{\text{src}} + a^{\text{ctx}})/(2\tau), \tag{5}$$

using $\tau = 0.1$ by default. The covariance view informs retrieval weights, while the retrieved values stay in the backbone feature space.

Reliability is a query diagnostic. The gate ρ_q combines base-prediction certainty, the confidence and top-two margins of prototype scores, and agreement between source and translated reference predictions. Both reference predictions already use both descriptor spaces. Clipping to $[0.02, 0.8]$ limits its range. The gate scales query corrections; it does not filter or relabel context windows. Appendix A specifies the calculation.

Residual feature and logit correction. An MLP adapter takes $[h_q, z_t, r_q, \ell_q^{\text{proto}}]$ and produces a bounded feature residual $\Delta h_q = \tanh(A_h(\cdot))$ and an unconstrained class-bias vector $b_q = A_b(\cdot)$. The output is

$$\widetilde{h}_q = h_q + \beta \rho_q \Delta h_q, \qquad \ell_q = g_\omega(\widetilde{h}_q) + \beta \rho_q b_q + \gamma \rho_q \ell_q^{\text{proto}}. \tag{6}$$

The feature correction obeys $\|\widetilde{h}_q - h_q\|_\infty \le |\beta| \rho_{\max}$, but the learned bias is not bounded by this expression. Both adapter output heads are initialized to zero; exact equality with NoAdapt additionally requires $\gamma = 0$, since the prototype-logit branch otherwise remains active.

3.4 Source training and deployment

We first train g_ω on frozen source embeddings. We then freeze its weights and train S_η, E_ϕ, the retrieval mixture, and the adapter. Each source subject supplies its first K windows as unlabeled context and its remaining windows as labeled training queries. Source prototypes are shared across the training subjects, including the subject supplying the current training queries. The loss is

$$\mathcal{L} = \text{CE}(\ell_q, y_q) + \lambda_p \,\text{CE}(\ell_q^{\text{proto}}, y_q) + \lambda_r \left(\text{mean}(\Delta h_q^2) + 0.1\,\text{mean}(b_q^2)\right), \tag{7}$$

with $\lambda_p = 0.2$ and $\lambda_r = 0.005$. Classification and prototype losses train the correction and retrieval pathways; residual regularization discourages large changes. Checkpoint and calibration-strength selection use only source validation. At deployment, all weights remain frozen; only context statistics and translated prototypes change. The reference bank has two prototypes per class in each space, independent of training-set size. Retrieval adds forward computation but requires no backward pass or source raw windows. Appendix A gives the architecture, gate, optimization and deployment procedure.

4 Experiments

4.1 Datasets, protocol and comparisons

We evaluate three-class SEED and four-class SEED-IV, each with 15 subjects (Zheng & Lu, 2015; Zheng et al., 2019), and two label granularities of FACED, which contains 123 participants (Chen et al., 2023). FACED-3 groups the negative and positive emotions into two classes while retaining neutral; FACED-9 uses the original nine categories. A separate classifier, prototype set and calibration model are trained for each task. The four tasks cover both coarse valence discrimination and finer emotion recognition.

We use leave-one-subject-out evaluation. For each target subject, one remaining source subject is reserved for validation, and the other subjects provide training data. The implementation selects the last subject in the remaining source-index list for validation. The classifier is trained on frozen embeddings before the calibration modules are fitted. Source validation selects the checkpoint; target labels are used only to calculate evaluation metrics. The primary context sizes are $K = 256$ for SEED and both FACED tasks and $K = 512$ for SEED-IV, with blocks of 128 queries. NoAdapt and NeuroContext-TTA use the same scored suffix in each paired comparison. All main results use pure forward adaptation, with no online updates, source-prior alignment or temporal smoothing.

The main comparison reports accuracy, macro-precision, macro-recall, macro-F1, one-vs-rest macro-AUROC and macro average precision (AUPRC). The backbone table reports accuracy and macro-F1 together, while the SEED ablation table includes all six metrics; the appendix gives the complete cross-task results and fold dispersions. All are expressed as percentages. Reported variation is the standard deviation across held-out-subject folds, not a confidence interval; the implementation uses population standard deviation. Changes in the text and plots are percentage-point

differences between the displayed means. Paired changes are examined separately in the fold visualizations; aggregate means and SDs alone do not establish significance. This distinction is relevant for EEG evaluation and TTA protocol design (Liu et al., 2024; Zhao et al., 2023).

The mdJPT pair isolates calibration; standalone methods assess complete-system performance, and the backbone study tests reuse with other representations. Appendix J separately reports fixed-feature TTA comparisons under a different configuration.

4.2 Main performance and standalone methods

Table 1 shows improvements on all four tasks. SEED macro-F1 rises from 72.37 to 73.72, and SEED-IV from 51.54 to 52.08. FACED-3 shows the largest gain, from 30.99 to 36.08, together with a 2.01-point accuracy improvement. FACED-9 improves from 15.71 to 16.23 macro-F1 and from 18.70 to 19.05 accuracy. The gains also extend to precision, recall, AUROC and AUPRC in every task (Table 1). Thus the paired improvements are not confined to accuracy or a single threshold-dependent summary.

FACED-3 gains more in macro-F1 than accuracy, indicating a benefit beyond the overall hit rate. The smaller FACED-9 gain reflects the remaining difficulty of fine-grained recognition.

Table 1 compares the evaluated pretrained, convolutional, graph and adaptation methods. Among the listed methods, calibrated mdJPT has the highest means on all six metrics for SEED, SEED-IV and FACED-9. On FACED-3, it leads on accuracy, precision, recall, macro-F1 and AUPRC, while EEGNet has the highest AUROC (60.22 versus 59.59). The adjacent mdJPT rows isolate the calibration effect; comparisons with other complete systems also involve differences in representation training and target access.

4.3 Generality across EEG backbones

Table 2 reports all 20 backbone–task pairs for mdJPT, REVE, DGCNN, EEGNet and EMOD. Calibration modules are trained separately for each representation. Accuracy improves in 19 pairs and macro-F1 in 15. On SEED, macro-F1 gains are positive for all five backbones: 1.35 for mdJPT, 7.20 for REVE, 5.06 for DGCNN, 19.93 for EEGNet and 7.55 for EMOD. The largest gain occurs for the comparatively weak EEGNet baseline; Table 2 reports absolute values, so this difference should not be read as the best final model.

On FACED, some backbones gain accuracy but lose macro-F1, indicating uneven class-wise benefits. All five gain macro-F1 on SEED. These results support reuse of the calibration design, not universal improvement or zero-shot transfer of one adapter across backbones.

4.4 Component contributions

Table 3 reports all six metrics for SEED. The full model leads on five metrics and ties for the highest AUROC. Removing the gate gives the largest component-wise macro-F1 reduction (73.72 to 72.85); removing either retrieval view or the adapter also lowers macro-F1. These results support the combined design on SEED; Appendix F reports all four tasks.

These are implemented-pathway interventions, not independent abstract factors: context encoding, translation and retrieval share inputs. Appendix F specifies exactly which quantities each ablation removes and analyzes the cross-task differences.

4.5 Context sensitivity and feature perturbations

We vary K over $\{64, 128, 256, 512\}$, or $\{128, 256, 512, 1024\}$ for SEED-IV (Figure 3). Performance is non-monotonic; longer context is not consistently better. Because the first K windows are excluded, this study also changes the scored suffix, rather than isolating context length on fixed queries.

Source-derived context often retains most gains, while shuffling context leaves results unchanged (Appendix I), consistent with permutation-invariant pooling. These interventions support statistical conditioning but do not establish that subject-matched context is necessary.

Table 1: Main comparison across four tasks and six metrics (%). NC denotes NeuroContext-TTA on frozen mdJPT features. All entries are evaluated fold means; precision, recall, F1, AUROC and AUPRC use macro averaging. Bold and underline mark the best and second-best distinct means within each task and metric; ties are shared. Methods differ in representation training and adaptation; the mdJPT pair isolates calibration. Fold standard deviations and implementation details are in Appendices C and D.

Method	Acc.	Precision	Recall	F1	AUROC	AUPRC
SEED						
REVE	39.58	41.67	39.26	36.69	58.44	41.07
DGCNN	49.30	53.24	49.39	47.02	71.24	55.03
EEGNet	37.01	44.16	36.77	26.71	62.73	46.19
EMOD	53.37	59.37	53.31	50.09	76.54	62.87
MS-DCDA	54.99	61.37	55.17	51.81	77.66	64.31
SDC-Net	56.30	60.06	56.44	55.43	77.38	62.60
ShareLink	53.20	59.29	53.08	49.95	75.35	60.81
mdJPT (NoAdapt)	<u>72.97</u>	<u>74.31</u>	<u>72.99</u>	<u>72.37</u>	<u>88.45</u>	<u>80.43</u>
mdJPT + NC	**74.05**	**75.25**	**74.06**	**73.72**	**88.76**	**80.68**
SEED-IV						
REVE	26.50	14.06	24.92	15.48	49.93	25.05
DGCNN	28.55	15.92	26.05	15.71	51.60	26.20
EEGNet	27.16	13.05	25.19	13.61	50.45	25.34
EMOD	26.63	7.84	25.00	11.19	50.54	25.61
MS-DCDA	27.03	10.20	25.53	11.66	52.30	27.15
SDC-Net	26.85	6.71	25.00	10.58	49.97	25.00
ShareLink	28.43	15.55	26.92	16.54	53.14	27.53
mdJPT (NoAdapt)	<u>52.18</u>	<u>52.96</u>	<u>52.50</u>	<u>51.54</u>	<u>76.52</u>	<u>54.79</u>
mdJPT + NC	**52.56**	**53.12**	**52.86**	**52.08**	**76.61**	**55.02**
FACED-3						
REVE	45.33	31.64	33.80	30.50	51.19	34.55
DGCNN	45.29	30.16	33.46	31.14	50.99	34.55
EEGNet	<u>49.47</u>	34.60	<u>36.35</u>	<u>31.54</u>	**60.22**	<u>41.16</u>
EMOD	46.61	31.23	34.39	30.39	54.85	36.92
MS-DCDA	48.65	<u>36.29</u>	36.16	30.82	59.01	40.82
SDC-Net	46.81	31.27	34.32	27.58	55.34	37.50
ShareLink	48.41	33.42	35.77	31.50	57.43	39.28
mdJPT (NoAdapt)	48.60	33.08	36.00	30.99	58.68	40.28
mdJPT + NC	**50.61**	**38.91**	**37.71**	**36.08**	<u>59.59</u>	**41.28**
FACED-9						
REVE	12.78	10.33	11.65	8.01	51.48	12.47
DGCNN	14.26	9.39	11.18	6.69	51.25	12.44
EEGNet	17.71	15.44	14.83	9.46	59.43	16.61
EMOD	17.29	11.61	13.95	8.35	59.24	16.53
MS-DCDA	15.94	14.20	14.48	9.93	58.99	16.70
SDC-Net	15.98	4.02	11.32	3.74	53.50	13.72
ShareLink	17.11	13.38	14.06	8.93	58.52	16.30
mdJPT (NoAdapt)	<u>18.70</u>	<u>17.21</u>	<u>17.12</u>	<u>15.71</u>	<u>60.06</u>	<u>17.26</u>
mdJPT + NC	**19.05**	**17.69**	**17.45**	**16.23**	**60.16**	**17.38**

The stress test perturbs query embedding and log-covariance dimensions with scaled Gaussian noise, keeping context clean. At rate 0.3, macro-F1 drops by 2.25, 1.58, 0.63 and 1.13 points across the four tasks (Figure 3; full results in Appendix H). This measures feature-noise sensitivity, not comparative robustness against NoAdapt.

Table 2: Generality across five frozen backbones. Each task reports accuracy / macro-F1 (%). NC is trained separately for each backbone. Bold and underline compare the two means within each backbone–task–metric pair. All six metrics and fold SDs appear in Appendix E.

Backbone	Setting	SEED	SEED-IV	FACED-3	FACED-9
mdJPT	NoAdapt	72.97 / 72.37	52.18 / 51.54	48.60 / 30.99	18.70 / 15.71
	+ NC	**74.05 / 73.72**	**52.56 / 52.08**	**50.61 / 36.08**	**19.05 / 16.23**
REVE	NoAdapt	39.58 / 36.69	**26.50** / 15.48	45.33 / 30.50	12.78 / 8.01
	+ NC	**44.08 / 43.89**	26.47 / **15.54**	**45.56 / 31.81**	**13.20 / 10.77**
DGCNN	NoAdapt	49.30 / 47.02	28.55 / 15.71	45.29 / **31.14**	14.26 / **6.69**
	+ NC	**52.55 / 52.08**	**28.63 / 15.79**	**45.66** / 30.54	**15.18** / 5.86
EEGNet	NoAdapt	37.01 / 26.71	27.16 / 13.61	49.47 / 31.54	17.71 / **9.46**
	+ NC	**48.26 / 46.64**	**27.28 / 13.76**	**50.41 / 33.52**	**18.48** / 9.34
EMOD	NoAdapt	53.37 / 50.09	26.63 / 11.19	46.61 / **30.39**	17.29 / **8.35**
	+ NC	**58.21 / 57.64**	**26.70 / 11.29**	**47.05** / 29.99	**17.87** / 8.11

Table 3: Component ablations on SEED across six metrics (%). Bold indicates the best mean per metric; ties are shared. The full model is listed last. Appendix F reports fold SDs and the ablations on all four tasks.

Variant	Acc.	Precision	Recall	F1	AUROC	AUPRC
NoAdapt	72.97	74.31	72.99	72.37	88.45	80.43
w/o context encoder	73.79	75.00	73.79	73.49	88.56	80.52
w/o translated memory	73.64	74.74	73.64	73.34	88.63	80.67
w/o SPD retrieval	73.62	74.82	73.63	73.38	88.54	80.51
w/o embedding retrieval	73.59	74.76	73.61	73.26	88.76	80.64
w/o reliability gate	73.28	74.24	73.31	72.85	88.41	80.13
w/o residual adapter	73.67	74.90	73.67	73.32	88.63	80.62
NeuroContext-TTA	**74.05**	**75.25**	**74.06**	**73.72**	**88.76**	**80.68**

4.6 Sample-level prediction changes

For logits ℓ and true class y, define $m = \ell_y - \max_{c \neq y} \ell_c$. A positive margin implies a correct top-1 prediction; a negative margin implies an error, excluding ties. Target labels are used only for this post-hoc analysis. Figure 4 shows two positive examples: SEED Fold 09 has mean gain 0.541 and 138 fixed versus 18 broken predictions; SEED-IV Fold 12 has gain 0.047 and 61 versus 9. Across all supplied folds, fixed counts exceed broken counts in 10/15 folds on each dataset. The totals are 1,074/307 for SEED and 1,176/620 for SEED-IV. These results show net error correction in the visualized query sets, while the complete fold analysis records heterogeneous effects.

5 Discussion and conclusion

NeuroContext-TTA uses recent unlabeled EEG to condition a source reference without reconstructing the target label space from predictions. Shared translation keeps every class represented, dual-space retrieval supplies a query-specific reference, and reliability scales a learned residual correction. The source-trained parameters remain fixed during deployment. The main mdJPT results improve all six metrics across four tasks, and the backbone study establishes broad, though not universal, benefits.

Preserving class structure is distinct from identifying subject shift: an unlabeled mean can change with class prevalence. Source anchors and gated corrections moderate reliance on that estimate, without resolving the ambiguity. Context interventions and mixed backbone results support this

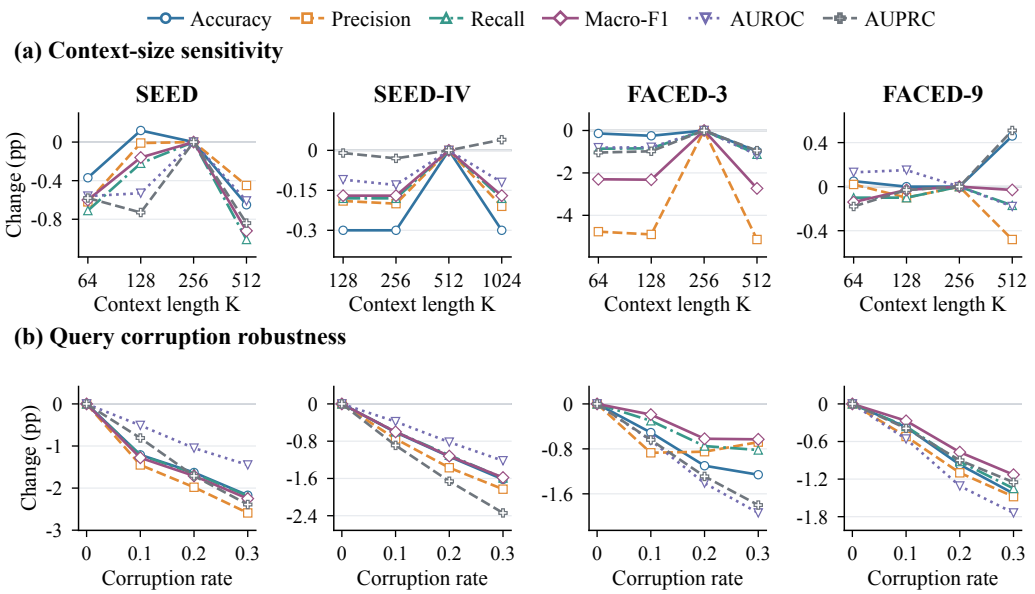


Figure 3: Six-metric sensitivity profiles. Colors and markers identify metrics. Each curve shows the change in fold mean (percentage points): (a) relative to the primary context length, $K = 512$ for SEED-IV and 256 otherwise; (b) relative to zero query corruption, with clean context. Axes vary by task; lines connect evaluated settings without smoothing. Absolute means and fold SDs are in Appendices G and H.

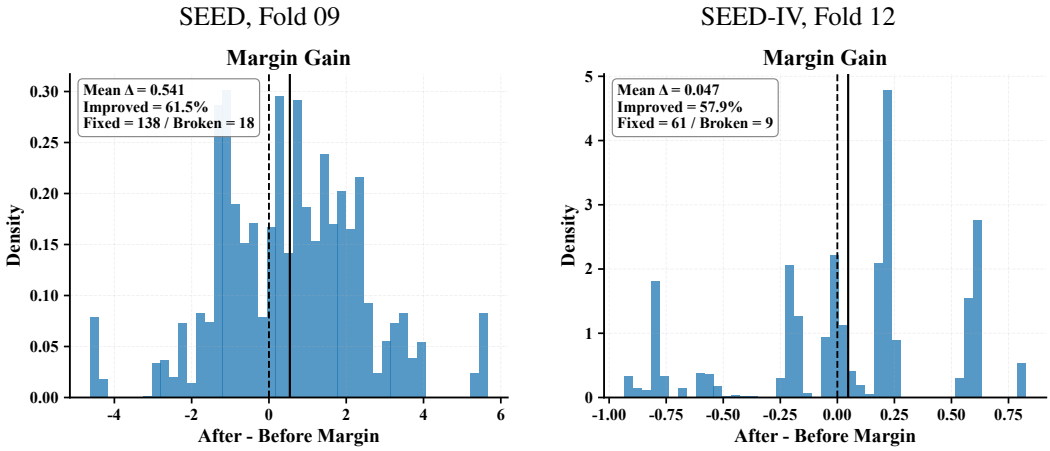


Figure 4: Positive fold examples of true-class margin change. Histograms show $m_{\text{after}} - m_{\text{before}}$; the dashed line marks zero. These folds were selected for positive mean gain and more fixed than broken predictions, not as an unbiased cross-fold summary. All 30 folds and their full three-panel views appear in Appendix K.

bounded interpretation. Future work should test class-controlled contexts and paired subject-level effects. The fold analysis adds a sample-level perspective to the aggregate gains: calibration corrects more predictions than it breaks in ten of fifteen folds on each dataset, although mean margin changes and error corrections need not agree. This makes decision changes a useful complement to score-distribution plots when evaluating adaptation. For deployment, the method retains a fixed class reference and bounded context while avoiding online optimization. Its practical benefit is therefore a learned correction rule that can use recent observations without maintaining a growing pseudo-labeled target memory.

AI USE STATEMENT

Generative AI tools assisted with language editing, manuscript organization, LaTeX preparation, reference checking, and limited code review and debugging. Experimental results were not generated by these tools. The authors reviewed and verified all AI-assisted content and are responsible for the final manuscript, citations, code, and figures.

ETHICS STATEMENT

This work concerns secondary analysis of EEG emotion-recognition benchmarks. EEG data can reveal sensitive information, and stimulus-associated emotion labels are not direct measurements of an individual's private mental state. The reported results do not establish suitability for diagnosis, consequential decision-making or covert emotion monitoring. Use of the datasets and any derived representations must respect their access conditions and participant privacy.

REPRODUCIBILITY STATEMENT

Section 3 defines the calibration operations and training objective. Appendix A gives architecture, optimization and deployment details, while the experimental appendices report the six available metrics for all evaluated configurations. The schematics distinguish conceptual operations from quantitative evidence, and quantitative plots use the reported tables or the supplied per-fold prediction analyses. The code is provided in the supplementary material.

REFERENCES

Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-Free Online Test-Time Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8344–8353, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Boudiaf_Parameter-Free_Online_Test-Time_Adaptation_CVPR_2022_paper.html.

Jingjing Chen, Xiaobin Wang, Chen Huang, Xin Hu, Xinke Shen, and Dan Zhang. A Large Finer-grained Affective Computing EEG Dataset. *Scientific Data*, 10(1):740, 2023. doi: 10.1038/s41597-023-02650-w. URL https://doi.org/10.1038/s41597-023-02650-w.

Yuning Chen, Sha Zhao, Shijian Li, and Gang Pan. EMOD: A unified eeg emotion representation framework leveraging v-a guided contrastive learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(21):17427–17435, 2026. doi: 10.1609/aaai.v40i21.38796. URL https://doi.org/10.1609/aaai.v40i21.38796.

Cheng Cheng, Wenzhe Liu, Ziyu Jia, and Weiqi He. Multi-view multi-scale adapter squeeze excitation network with domain generalization for cross-subject emotion recognition. *Expert Systems with Applications*, 297:129222, 2026. doi: 10.1016/j.eswa.2025.129222. URL https://doi.org/10.1016/j.eswa.2025.129222.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A Foundation Model for EEG – Adapting to Any Setup with Large-Scale Pretraining on 25,000 Subjects. In *Advances in Neural Information Processing Systems*, volume 38, pp. 25583–25619, 2025. doi: 10.52202/085713-0760. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/20a917f77773ac0fa8bea2bdd6606b66-Abstract-Conference.html.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. URL https://jmlr.org/papers/v17/15-239.html.

Peiliang Gong, Mohamed Ragab, Min Wu, Zhenghua Chen, Yongyi Su, Xiaoli Li, and Daoqiang Zhang. Augmented Contrastive Clustering with Uncertainty-Aware Prototyping for Time Series

Test Time Adaptation. *arXiv preprint arXiv:2501.01472*, 2025. URL https://arxiv.org/abs/2501.01472.

Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. NOTE: Robust Continual Test-time Adaptation Against Temporal Correlation. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1976. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/ae6c7dbd9429b3a75c41b5fb47e57c9e-Abstract-Conference.html.

He He and Dongrui Wu. Transfer Learning for Brain–Computer Interfaces: A Euclidean Space Data Alignment Approach. *IEEE Transactions on Biomedical Engineering*, 67(2):399–410, 2020. doi: 10.1109/TBME.2019.2913914. URL https://doi.org/10.1109/TBME.2019.2913914.

Zixuan Huang, Lingyao Kong, Licheng Ao, Shiyi Yao, An Xiang, and Fen Miao. ShareLink: Neuro-Inspired EEG-Based Cross-Subject Emotion Recognition via Shared Bi-hemisphere. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pp. 605–615. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-05162-2_58. URL https://papers.miccai.org/miccai-2025/0829-Paper5190.html.

Md Niaz Imtiaz and Naimul Khan. An Enhanced Source-Free Unsupervised Domain Adaptation Framework for Cross-Dataset EEG Emotion Recognition via Predictive Coding and Test-Time Training. *arXiv preprint arXiv:2606.28202*, 2026. doi: 10.48550/arXiv.2606.28202. URL https://arxiv.org/abs/2606.28202.

Yusuke Iwasawa and Yutaka Matsuo. Test-Time Classifier Adjustment Module for Model-Agnostic Domain Generalization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2427–2440, 2021. URL https://papers.nips.cc/paper/2021/hash/1415fe9fea0fa1e45dddcff5682239a0-Abstract.html.

Minguk Jang, Sae-Young Chung, and Hye Won Chung. Test-Time Adaptation via Self-Training with Nearest Neighbor Information. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2207.10792.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QzTpTRVtrP.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient Test-Time Adaptation of Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14162–14171, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Karmanov_Efficient_Test-Time_Adaptation_of_Vision-Language_Models_CVPR_2024_paper.html.

Demetres Kostas, Stéphane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using Transformers and a Contrastive Self-Supervised Learning Task to Learn From Massive Amounts of EEG Data. *Frontiers in Human Neuroscience*, 15:653659, 2021. doi: 10.3389/fnhum.2021.653659. URL https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2021.653659/full.

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. doi: 10.1088/1741-2552/aace8c. URL https://doi.org/10.1088/1741-2552/aace8c.

Gabriel Jason Lee, Jathurshan Pradeepkumar, and Jimeng Sun. Test-Time Adaptation for EEG Foundation Models: A Systematic Study under Real-World Distribution Shifts. In *Proceedings of the 11th Machine Learning for Healthcare Conference*, volume 340 of *Proceedings of Machine Learning Research*, pp. 948–984. PMLR, 2026. URL https://proceedings.mlr.press/v340/lee26a.html.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not Enough for Test-Time Adaptation: From the Perspective of Disentangled Factors. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/cd0665986c1e9c15f6569ac944bcf88a-Abstract-Conference.html.

Youngjun Lee, Doyoung Kim, Junhyeok Kang, Jihwan Bang, Hwanjun Song, and Jae-Gil Lee. RA-TTA: Retrieval-Augmented Test-Time Adaptation for Vision-Language Models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/fa1790d7c3036c691d0b2fb3b9a0ce64-Abstract-Conference.html.

Siyang Li, Jiayi Ouyang, Zhenyao Cui, Ziwei Wang, Tianwang Jia, Feng Wan, and Dongrui Wu. Backpropagation-Free Test-Time Adaptation for Lightweight EEG-Based Brain-Computer Interfaces. *IEEE Journal of Biomedical and Health Informatics*, pp. 1–13, 2026. doi: 10.1109/JBHI.2026.3725326.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 6028–6039, 2020. URL https://proceedings.mlr.press/v119/liang20a.html.

Huan Liu, Shusen Yang, Yuzhe Zhang, Mengze Wang, Fanyu Gong, Chengxi Xie, Guanjian Liu, Zejun Liu, Yong-Jin Liu, Bao-Liang Lu, and Dalin Zhang. LibEER: A Comprehensive Benchmark and Algorithm Library for EEG-based Emotion Recognition. *arXiv preprint arXiv:2410.09767*, 2024. doi: 10.48550/arXiv.2410.09767. URL https://arxiv.org/abs/2410.09767.

Yuejiang Liu, Parth Kothari, Bastien van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. TTT++: When Does Self-Supervised Test-Time Training Fail or Thrive? In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/b618c3210e934362ac261db280128c22-Abstract.html.

Apolline Mellot, Antoine Collas, Sylvain Chevallier, Alexandre Gramfort, and Denis A. Engemann. Geodesic Optimization for Predictive Shift Adaptation on EEG data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 32828–32855, 2024. doi: 10.52202/079017-1033. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/399888844ffc8fa95c2db196f3f1b60b-Abstract-Conference.html.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient Test-Time Model Adaptation without Forgetting. In *International Conference on Machine Learning*, volume 162, pp. 16888–16905, 2022. URL https://proceedings.mlr.press/v162/niu22a.html.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards Stable Test-Time Adaptation in Dynamic Wild World. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=g2YraF75Tj.

Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. EEG Emotion Recognition Using Dynamical Graph Convolutional Neural Networks. *IEEE Transactions on Affective Computing*, 11 (3):532–541, 2020. doi: 10.1109/taffc.2018.2817622. URL https://doi.org/10.1109/TAFFC.2018.2817622.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG Conformer: Convolutional Transformer for EEG Decoding and Visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023. doi: 10.1109/tnsre.2022.3230250. URL https://doi.org/10.1109/tnsre.2022.3230250.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 9229–9248, 2020. URL https://proceedings.mlr.press/v119/sun20b.html.

Jiahao Tang, Youjun Li, Xiangting Fan, Yangxuan Zheng, Siyuan Lu, Xueping Li, Peng Fang, Chenxi Li, and Zi-Gang Huang. SDC-Net: A Domain Adaptation Framework with Semantic-Dynamic Consistency for Cross-Subject EEG Emotion Recognition. *arXiv preprint arXiv:2507.17524*, 2025. URL https://arxiv.org/abs/2507.17524.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully Test-time Adaptation by Entropy Minimization. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2006.10726.

Guagnyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *Advances in Neural Information Processing Systems*, volume 37, pp. 39249–39280. Curran Associates, Inc., 2024. doi: 10.52202/079017-1239. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4540d267eeec4e5dbd9dae9448f0b739-Paper-Conference.pdf.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual Test-Time Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Continual_Test-Time_Domain_Adaptation_CVPR_2022_paper.html.

Yun Xiao, Yimeng Zhang, Xiaopeng Peng, Shuzheng Han, Xia Zheng, Dingyi Fang, and Xiaojiang Chen. Multi-source EEG emotion recognition via dynamic contrastive domain adaptation. *Biomedical Signal Processing and Control*, 102:107337, 2025. doi: 10.1016/j.bspc.2024.107337. URL https://doi.org/10.1016/j.bspc.2024.107337.

Rui Yan, Yibo Li, Han Ding, and Fei Wang. Cross-Domain EEG-Based Emotion Recognition with Contrastive Learning. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7887–7891, 2026. doi: 10.1109/ICASSP55912.2026.11460520.

Chaoqi Yang, M Westover, and Jimeng Sun. BIOT: Biosignal Transformer for Cross-data Learning in the Wild. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78240–78260. Curran Associates, Inc., 2023. doi: 10.52202/075280-3420. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f6b30f3e2dd9cb53bbf2024402d02295-Paper-Conference.pdf.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust Test-Time Adaptation in Dynamic Scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15922–15932, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Yuan_Robust_Test-Time_Adaptation_in_Dynamic_Scenarios_CVPR_2023_paper.html.

Chenchen Zhang, Yanrong Hao, Xin Wen, Mengni Zhou, Fei Yuan, and Rui Cao. FG-MSTGNN: Cross-subject EEG Emotion Recognition via Frequency-guided Multi-period Spatial-temporal Graph Neural Network. In *Proceedings of the 17th Asian Conference on Machine Learning*, volume 304, pp. 1102–1117, 2025a. URL https://proceedings.mlr.press/v304/zhang26c.html.

Qingzhu Zhang, Jiani Zhong, Zongsheng Li, Xinke Shen, and Quanying Liu. Multi-dataset Joint Pre-training of Emotional EEG Enables Generalizable Affective Computing. In *Advances in Neural Information Processing Systems*, volume 38, pp. 183384–183411, 2025b. doi: 10.52202/085713-5515. URL https://papers.neurips.cc/paper_files/paper/2025/hash/f1b8d443042f376aa3654b6c68de6297-Abstract-Conference.html.

Taolin Zhang, Jinpeng Wang, Hang Guo, Tao Dai, Bin Chen, and Shu-Tao Xia. BoostAdapter: Improving Vision-Language Test-Time Adaptation via Regional Bootstrapping. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2165. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7d60bfd8458b67acbbaf18b892338d00-Abstract-Conference.html.

Yifan Zhang, Xue Wang, Kexin Jin, Kun Yuan, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. AdaNPC: Exploring Non-Parametric Classifier for Test-Time Adaptation. In *International Conference on Machine Learning*, volume 202, pp. 41647–41676, 2023. URL https://proceedings.mlr.press/v202/zhang23am.html.

Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On Pitfalls of Test-Time Adaptation. In *International Conference on Machine Learning*, volume 202, pp. 42058–42080, 2023. URL https://proceedings.mlr.press/v202/zhao23d.html.

Wei-Long Zheng and Bao-Liang Lu. Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015. doi: 10.1109/TAMD.2015.2431497. URL https://doi.org/10.1109/TAMD.2015.2431497.

Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. EmotionMeter: A Multimodal Framework for Recognizing Human Emotions. *IEEE Transactions on Cybernetics*, 49(3):1110–1122, 2019. doi: 10.1109/TCYB.2018.2797176. URL https://pubmed.ncbi.nlm.nih.gov/29994384/.

Peixiang Zhong, Di Wang, and Chunyan Miao. EEG-Based Emotion Recognition Using Regularized Graph Neural Networks. *IEEE Transactions on Affective Computing*, 13:1290–1301, 2022. doi: 10.1109/taffc.2020.2994159. URL https://doi.org/10.1109/taffc.2020.2994159.

A IMPLEMENTATION AND DEPLOYMENT

Descriptor interface and data split. The calibration implementation consumes precomputed backbone embeddings and log-covariance vectors for each subject and window. It does not fit an EEG backbone or construct covariance matrices from raw signals. The log-covariance vector is encoded by a learned MLP; similarity is cosine similarity in its output space, not a Riemannian distance between covariance matrices. The source class means are computed in the input descriptor spaces before this encoder. In each leave-one-subject-out fold, the last remaining source index is used for validation and all other source subjects for training. Target labels are never inputs to the calibrator. Exclusion of the held-out subject from downstream training is distinct from the provenance of any externally pretrained backbone.

Architecture. The frozen downstream classifier is an MLP with widths $d \to 256 \to 256 \to C$, LayerNorm, GELU and dropout. The covariance encoder has widths $p \to 512 \to 256 \to 128$. The context encoder maps the concatenated $3d + 256$ statistics through a 256-dimensional hidden layer to a 128-dimensional output. The adapter takes $[h_q, z_t, r_q, \ell_q^{\text{proto}}]$, uses two 256-dimensional hidden layers, and has separate feature and class-bias output heads. The feature residual uses $\tanh$; the bias head is unconstrained. Both heads are initialized to zero. The prototype-logit branch can still change predictions at initialization when its coefficient is nonzero.

Training and selection. The default implementation uses 12 epochs for the classifier and 12 for calibration, batches of 128, AdamW with learning rate 10^{-3} and weight decay 10^{-3}, and gradient clipping at norm 5. The default random seed is 2026. The classifier checkpoint is selected by source-validation accuracy; the calibration checkpoint uses macro-F1 plus 0.2 times AUPRC. Source context consists of the first K windows of each training subject; the remaining labeled queries are shuffled for training. Source prototypes include all source training subjects. This is supervised training of a context-conditioned rule, not leave-one-source-subject-out meta-training.

Reliability calculation. Let $p_q^0 = \text{softmax}(\ell_q^0)$, $H_q = -\sum_c p_{q,c}^0 \log p_{q,c}^0$, $k_j = \arg\max_c a_c^j$, and $m_j = a_{(1)}^j - a_{(2)}^j$ be the top-two score margin. The implemented gate is

$$u_q = \frac{\max_c \widehat{w}_c^{\text{src}} + \max_c \widehat{w}_c^{\text{ctx}}}{2}, \qquad d_q = \frac{1 + \mathbf{1}[k_{\text{src}} = k_{\text{ctx}}]}{2},$$
$$\rho_q = \text{clip}\left[\left(1 - \frac{H_q}{\log C}\right) u_q d_q \, \sigma(8(m_{\text{src}} + m_{\text{ctx}})), \rho_{\min}, \rho_{\max}\right]. \tag{8}$$

where σ is the sigmoid, $\widehat{w}^j = \text{softmax}(a^j/0.1)$, $\rho_{\min} = 0.02$, and $\rho_{\max} = 0.8$. Agreement is between *source and translated prototype predictions*, each already using both spaces. The gate neither filters context windows nor directly tests agreement between the two spaces. Its role is to modulate the correction of each query, rather than to reject or relabel context windows.

Calibration strengths. Default values are $\lambda = 0.35$, $\tau = 0.1$, $\beta = 0.2$, $\gamma = 0.05$, $\rho_{\min} = 0.02$ and $\rho_{\max} = 0.8$. The retrieval mixing weight is the sigmoid of a trainable scalar initialized to zero, hence initially 0.5. The implementation also supports source-validation selection of $\beta \in \{0, 0.05, 0.1, 0.15, 0.2\}$, $\gamma \in \{0, 0.02, 0.05\}$ and $\rho_{\max} \in \{0.3, 0.5, 0.8\}$, with criterion accuracy plus macro-F1 plus 0.05 times AUPRC. This search, when used, does not use target-subject labels. The three optional deployment operations—online adapter updates, source-prior alignment and temporal smoothing—are disabled in the main experiments.

Deployment order. The following operations are repeated for each query block.

1. Read the preceding K embedding and covariance descriptors; compute the pooled context state and the means of the two raw descriptor spaces.
2. Translate each source prototype by the corresponding shared offset in Eq. (1). Keep the original source prototypes as a second reference set.
3. For each query, compute the two-view similarity scores against both reference sets, followed by retrieval weights, the retrieved embedding and prototype logits.

4. Compute the frozen classifier's base probabilities and the reliability gate from certainty, source/context-reference agreement and their score margins.
5. Predict the feature and bias residuals, scale them with the gate, and apply Eq. (6). No gradients or target labels are needed.
6. After the block, advance the context boundary. The next block may use these now-past query descriptors as history.

Storage and computation. The source memory stores $C(d+p)$ descriptor values; its translated copy has the same size. It therefore scales with the number of classes, not the number of training windows. A direct history buffer stores $K(d+p)$ values. Encoded covariance features and context statistics can be shared across all queries in a block. Query retrieval compares against $2C$ prototypes per view, followed by small MLP evaluations. Forward-only deployment removes test-time backpropagation, but adds retrieval and calibration computation relative to NoAdapt. No end-to-end latency or FLOPs claim is inferred from these operation counts.

B Interpretation of shared prototype translation

Equation (2) follows because the same offset is subtracted when comparing two translated prototypes. Under an illustrative additive model $h = \mu_y + b_s + \epsilon$, with zero-mean noise conditional on class and subject and source prototypes $p_c^h = \mu_c + b_{\mathrm{src}}$,

$$\mathbb{E}[\bar{h}_t - \bar{p}^h] = b_s - b_{\mathrm{src}} + \sum_{c=1}^{C} (\pi_{t,c} - 1/C)\mu_c, \tag{9}$$

where $\pi_{t,c}$ is the class proportion in the context. The first term represents subject shift, while the second arises from a context whose class proportions differ from uniform. The translation is therefore an estimate of a context offset, not an identifiable estimate of a physiological subject effect. Shrinkage and retention of the source reference moderate its use but do not remove the second term. Likewise, the gate measures consistency and confidence rather than correctness. These distinctions explain why class preservation is an architectural property whereas adaptation benefit is an empirical question.

C Full main results

The two methods share the mdJPT representation and scored query suffix. Results are means and population standard deviations across target-subject folds. All six metrics improve with calibration in each task. In each comparison group, bold marks the best and underline the second-best distinct mean; ties share the same mark.

Table 4: Full main results. Each cell gives mean and SD (%).

Method	Acc	Precision	Recall	F1	AUROC	AUPRC
SEED						
NoAdapt	<u>72.97</u> ±8.57	<u>74.31</u> ±8.13	<u>72.99</u> ±8.55	<u>72.37</u> ±9.08	<u>88.45</u> ±5.33	<u>80.43</u> ±8.62
NeuroContext-TTA	**74.05** ±7.82	**75.25** ±7.21	**74.06** ±7.83	**73.72** ±7.92	**88.76** ±5.04	**80.68** ±8.39
SEED-IV						
NoAdapt	<u>52.18</u> ±4.57	<u>52.96</u> ±4.73	<u>52.50</u> ±4.56	<u>51.54</u> ±4.50	<u>76.52</u> ±3.41	<u>54.79</u> ±5.71
NeuroContext-TTA	**52.56** ±4.74	**53.12** ±5.00	**52.86** ±4.79	**52.08** ±4.68	**76.61** ±3.36	**55.02** ±5.69

Continued on next page

Table 4 continued

Method	Acc	Precision	Recall	F1	AUROC	AUPRC
FACED-3						
NoAdapt	<u>48.60</u> ±3.64	<u>33.08</u> ±5.19	<u>36.00</u> ±2.68	<u>30.99</u> ±5.05	<u>58.68</u> ±5.09	<u>40.28</u> ±4.14
NeuroContext-TTA	**50.61** ±4.49	**38.91** ±9.59	**37.71** ±3.38	**36.08** ±3.39	**59.59** ±5.88	**41.28** ±5.02
FACED-9						
NoAdapt	<u>18.70</u> ±2.86	<u>17.21</u> ±3.33	<u>17.12</u> ±2.71	<u>15.71</u> ±2.63	<u>60.06</u> ±3.37	<u>17.26</u> ±2.29
NeuroContext-TTA	**19.05** ±3.01	**17.69** ±3.81	**17.45** ±2.83	**16.23** ±2.73	**60.16** ±3.46	**17.38** ±2.35

D STANDALONE METHOD COMPARISONS

The tables report the evaluated implementations, not numbers copied from their original papers. DGCNN and EEGNet use the LibEER raw-window implementations; EMOD uses its pretrained raw-window implementation. Their representation training and adaptation mechanisms differ, so these tables compare complete systems rather than isolated TTA modules. ShareLink on FACED-3 and FACED-9 uses raw EEG converted to differential-entropy features, eight training epochs, 123 held-out-subject folds and 256 context windows. The target context is excluded from the scored query set. In each comparison group, bold marks the best mean; ties share the same mark.

Table 5: Standalone method comparisons. Each cell gives mean and SD (%). NC denotes NeuroContext-TTA.

Method	Acc	Precision	Recall	F1	AUROC	AUPRC
SEED						
REVE	39.58 ±2.40	41.67 ±3.58	39.26 ±2.53	36.69 ±4.34	58.44 ±3.57	41.07 ±3.70
DGCNN	49.30 ±6.70	53.24 ±4.78	49.39 ±6.47	47.02 ±7.22	71.24 ±5.79	55.03 ±6.03
EEGNet	37.01 ±4.20	44.16 ±13.46	36.77 ±4.06	26.71 ±4.69	62.73 ±11.22	46.19 ±12.13
EMOD	53.37 ±8.69	59.37 ±6.21	53.31 ±8.58	50.09 ±10.23	76.54 ±6.56	62.87 ±7.97
MS-DCDA	54.99 ±9.39	61.37 ±5.14	55.17 ±9.14	51.81 ±11.20	77.66 ±6.09	64.31 ±7.66
SDC-Net	56.30 ±6.84	60.06 ±5.12	56.44 ±6.65	55.43 ±6.89	77.38 ±6.52	62.60 ±8.01
ShareLink	53.20 ±7.26	59.29 ±6.65	53.08 ±7.04	49.95 ±7.82	75.35 ±7.49	60.81 ±8.60
mdJPT (NoAdapt)	72.97 ±8.57	74.31 ±8.13	72.99 ±8.55	72.37 ±9.08	88.45 ±5.33	80.43 ±8.62
mdJPT + NC	**74.05** ±7.82	**75.25** ±7.21	**74.06** ±7.83	**73.72** ±7.92	**88.76** ±5.04	**80.68** ±8.39
SEED-IV						
REVE	26.50 ±1.03	14.06 ±8.89	24.92 ±0.37	15.48 ±5.90	49.93 ±0.35	25.05 ±0.18
DGCNN	28.55 ±1.44	15.92 ±11.07	26.05 ±1.82	15.71 ±6.12	51.60 ±2.73	26.20 ±2.04
EEGNet	27.16 ±0.76	13.05 ±7.95	25.19 ±0.65	13.61 ±3.90	50.45 ±0.84	25.34 ±0.58
EMOD	26.63 ±0.85	7.84 ±4.23	25.00 ±0.00	11.19 ±2.28	50.54 ±1.93	25.61 ±1.44
MS-DCDA	27.03 ±2.66	10.20 ±4.66	25.53 ±1.28	11.66 ±2.79	52.30 ±3.93	27.15 ±3.04
SDC-Net	26.85 ±0.00	6.71 ±0.00	25.00 ±0.00	10.58 ±0.00	49.97 ±0.12	25.00 ±0.01
ShareLink	28.43 ±2.41	15.55 ±11.72	26.92 ±2.90	16.54 ±7.96	53.14 ±4.54	27.53 ±3.73
mdJPT (NoAdapt)	52.18 ±4.57	52.96 ±4.73	52.50 ±4.56	51.54 ±4.50	76.52 ±3.41	54.79 ±5.71
mdJPT + NC	**52.56** ±4.74	**53.12** ±5.00	**52.86** ±4.79	**52.08** ±4.68	**76.61** ±3.36	**55.02** ±5.69

Continued on next page

Table 5 continued

Method	Acc	Precision	Recall	F1	AUROC	AUPRC
FACED-3						
REVE	45.33 ±2.27	31.64 ±2.74	33.80 ±1.35	30.50 ±2.80	51.19 ±2.10	34.55 ±1.29
DGCNN	45.29 ±2.12	30.16 ±1.51	33.46 ±1.54	31.14 ±1.57	50.99 ±2.16	34.55 ±1.44
EEGNet	49.47 ±3.30	34.60 ±4.58	36.35 ±2.49	31.54 ±4.78	**60.22** ±4.83	41.16 ±4.00
EMOD	46.61 ±3.17	31.23 ±3.12	34.39 ±2.36	30.39 ±3.46	54.85 ±4.42	36.92 ±3.14
MS-DCDA	48.65 ±4.49	36.29 ±10.88	36.16 ±3.34	30.82 ±5.92	59.01 ±5.78	40.82 ±4.92
SDC-Net	46.81 ±3.12	31.27 ±6.18	34.32 ±2.32	27.58 ±4.78	55.34 ±5.21	37.50 ±3.60
ShareLink	48.41 ±4.37	33.42 ±5.44	35.77 ±3.23	31.50 ±4.94	57.43 ±6.01	39.28 ±4.60
mdJPT (NoAdapt)	48.60 ±3.64	33.08 ±5.19	36.00 ±2.68	30.99 ±5.05	58.68 ±5.09	40.28 ±4.14
mdJPT + NC	**50.61** ±4.49	**38.91** ±9.59	**37.71** ±3.38	**36.08** ±3.39	59.59 ±5.88	**41.28** ±5.02
FACED-9						
REVE	12.78 ±1.72	10.33 ±3.42	11.65 ±0.98	8.01 ±1.18	51.48 ±1.48	12.47 ±0.64
DGCNN	14.26 ±1.40	9.39 ±4.44	11.18 ±0.80	6.69 ±1.27	51.25 ±1.41	12.44 ±0.63
EEGNet	17.71 ±2.58	15.44 ±6.05	14.83 ±2.27	9.46 ±2.34	59.43 ±4.32	16.61 ±2.57
EMOD	17.29 ±2.26	11.61 ±5.70	13.95 ±2.08	8.35 ±2.41	59.24 ±3.21	16.53 ±1.99
MS-DCDA	15.94 ±2.62	14.20 ±5.23	14.48 ±2.20	9.93 ±3.46	58.99 ±3.39	16.70 ±1.92
SDC-Net	15.98 ±0.58	4.02 ±3.02	11.32 ±0.49	3.74 ±0.73	53.50 ±3.44	13.72 ±1.54
ShareLink	17.11 ±2.34	13.38 ±5.82	14.06 ±1.99	8.93 ±2.40	58.52 ±3.70	16.30 ±2.11
mdJPT (NoAdapt)	18.70 ±2.86	17.21 ±3.33	17.12 ±2.71	15.71 ±2.63	60.06 ±3.37	17.26 ±2.29
mdJPT + NC	**19.05** ±3.01	**17.69** ±3.81	**17.45** ±2.83	**16.23** ±2.73	**60.16** ±3.46	**17.38** ±2.35

E GENERALITY ACROSS FROZEN EEG BACKBONES

Each backbone has its own NoAdapt reference and separately trained calibration modules. Values are not obtained by transferring one calibrator checkpoint between different embedding spaces. Best and second-best marks compare the two rows within each backbone pair, not unrelated backbones. In each comparison group, bold marks the best and underline the second-best distinct mean; ties share the same mark.

Table 6: Generality across frozen EEG backbones. Each cell gives mean and SD (%). NC denotes NeuroContext-TTA.

Method	Acc	Precision	Recall	F1	AUROC	AUPRC
SEED						
mdJPT (NoAdapt)	72.97 ±8.57	74.31 ±8.13	72.99 ±8.55	72.37 ±9.08	88.45 ±5.33	80.43 ±8.62
mdJPT + NC	**74.05** ±7.82	**75.25** ±7.21	**74.06** ±7.83	**73.72** ±7.92	**88.76** ±5.04	**80.68** ±8.39
REVE	39.58 ±2.40	41.67 ±3.58	39.26 ±2.53	36.69 ±4.34	58.44 ±3.57	41.07 ±3.70
REVE + NC	**44.08** ±4.61	**43.94** ±4.42	**43.99** ±4.57	**43.89** ±4.47	**62.64** ±5.38	**45.91** ±6.02
DGCNN	49.30 ±6.70	**53.24** ±4.78	49.39 ±6.47	47.02 ±7.22	71.24 ±5.79	55.03 ±6.03
DGCNN + NC	**52.55** ±5.04	52.63 ±5.10	**52.62** ±4.98	**52.08** ±4.92	**71.69** ±5.71	**55.52** ±5.93
EEGNet	37.01 ±4.20	44.16 ±13.46	36.77 ±4.06	26.71 ±4.69	62.73 ±11.22	46.19 ±12.13
EEGNet + NC	**48.26** ±7.26	**48.31** ±7.09	**48.08** ±7.26	**46.64** ±7.66	**67.26** ±7.94	**50.69** ±8.66
EMOD	53.37 ±8.69	59.37 ±6.21	53.31 ±8.58	50.09 ±10.23	76.54 ±6.56	62.87 ±7.97
EMOD + NC	**58.21** ±6.97	**59.83** ±5.93	**58.09** ±6.92	**57.64** ±6.54	**77.04** ±6.35	**63.38** ±7.77
SEED-IV						
mdJPT (NoAdapt)	52.18 ±4.57	52.96 ±4.73	52.50 ±4.56	51.54 ±4.50	76.52 ±3.41	54.79 ±5.71
mdJPT + NC	**52.56** ±4.74	**53.12** ±5.00	**52.86** ±4.79	**52.08** ±4.68	**76.61** ±3.36	**55.02** ±5.69
REVE	**26.50** ±1.03	**14.06** ±8.89	**24.92** ±0.37	15.48 ±5.90	**49.93** ±0.35	25.05 ±0.18
REVE + NC	26.47 ±1.07	14.01 ±8.83	24.90 ±0.43	**15.54** ±5.98	49.75 ±0.80	**25.23** ±0.54
DGCNN	28.55 ±1.44	**15.92** ±11.07	26.05 ±1.82	15.71 ±6.12	51.60 ±2.73	26.20 ±2.04
DGCNN + NC	**28.63** ±1.59	15.89 ±11.03	**26.07** ±1.82	**15.79** ±6.20	**51.70** ±2.71	**26.33** ±2.00
EEGNet	27.16 ±0.76	**13.05** ±7.95	25.19 ±0.65	13.61 ±3.90	50.45 ±0.84	25.34 ±0.58
EEGNet + NC	**27.28** ±0.97	12.94 ±8.04	**25.26** ±0.83	**13.76** ±3.98	**50.65** ±0.77	**25.61** ±0.56
EMOD	26.63 ±0.85	**7.84** ±4.23	25.00 ±0.00	11.19 ±2.28	**50.54** ±1.93	25.61 ±1.44
EMOD + NC	**26.70** ±0.56	7.82 ±4.13	**25.03** ±0.10	**11.29** ±2.64	50.51 ±1.94	**25.63** ±1.43

Continued on next page

Table 6 continued

Method	Acc	Precision	Recall	F1	AUROC	AUPRC
FACED-3						
mdJPT (NoAdapt)	48.60 ±3.64	33.08 ±5.19	36.00 ±2.68	30.99 ±5.05	58.68 ±5.09	40.28 ±4.14
mdJPT + NC	**50.61** ±4.49	**38.91** ±9.59	**37.71** ±3.38	**36.08** ±3.39	**59.59** ±5.88	**41.28** ±5.02
REVE	45.33 ±2.27	**31.64** ±2.74	**33.80** ±1.35	30.50 ±2.80	**51.19** ±2.10	34.55 ±1.29
REVE + NC	**45.56** ±2.01	30.45 ±1.32	33.72 ±1.33	**31.81** ±1.30	**51.19** ±2.08	**34.56** ±1.30
DGCNN	45.29 ±2.12	30.16 ±1.51	33.46 ±1.54	**31.14** ±1.57	50.99 ±2.16	34.55 ±1.44
DGCNN + NC	**45.66** ±1.90	**30.24** ±1.62	**33.54** ±1.41	30.54 ±1.93	**51.08** ±2.13	**34.62** ±1.44
EEGNet	49.47 ±3.30	**34.60** ±4.58	36.35 ±2.49	31.54 ±4.78	60.22 ±4.83	41.16 ±4.00
EEGNet + NC	**50.41** ±3.65	34.50 ±3.90	**37.07** ±2.73	**33.52** ±3.94	**60.34** ±4.90	**41.27** ±4.09
EMOD	46.61 ±3.17	31.23 ±3.12	34.39 ±2.36	**30.39** ±3.46	54.85 ±4.42	36.92 ±3.14
EMOD + NC	**47.05** ±2.91	**31.62** ±3.90	**34.46** ±2.21	29.99 ±3.74	**54.87** ±4.43	**36.94** ±3.16
FACED-9						
mdJPT (NoAdapt)	18.70 ±2.86	17.21 ±3.33	17.12 ±2.71	15.71 ±2.63	60.06 ±3.37	17.26 ±2.29
mdJPT + NC	**19.05** ±3.01	**17.69** ±3.81	**17.45** ±2.83	**16.23** ±2.73	**60.16** ±3.46	**17.38** ±2.35
REVE	12.78 ±1.72	10.33 ±3.42	11.65 ±0.98	8.01 ±1.18	51.48 ±1.48	12.47 ±0.64
REVE + NC	**13.20** ±1.27	**11.85** ±1.70	**11.67** ±1.14	**10.77** ±1.25	**51.49** ±1.49	**12.48** ±0.65
DGCNN	14.26 ±1.40	**9.39** ±4.44	11.18 ±0.80	**6.69** ±1.27	51.25 ±1.41	12.44 ±0.63
DGCNN + NC	**15.18** ±0.93	8.86 ±4.46	**11.27** ±0.69	5.86 ±1.14	**51.37** ±1.42	**12.49** ±0.64
EEGNet	17.71 ±2.58	**15.44** ±6.05	14.83 ±2.27	**9.46** ±2.34	59.43 ±4.32	16.61 ±2.57
EEGNet + NC	**18.48** ±2.46	14.59 ±5.32	**14.95** ±2.43	9.34 ±2.67	**59.57** ±4.32	**16.70** ±2.60
EMOD	17.29 ±2.26	11.61 ±5.70	13.95 ±2.08	**8.35** ±2.41	59.24 ±3.21	16.53 ±1.99
EMOD + NC	**17.87** ±2.21	**11.72** ±6.02	**13.97** ±2.05	8.11 ±2.20	**59.26** ±3.23	**16.54** ±2.00

F COMPONENT ABLATIONS

Without the context encoder, its output is zero but the shared translation remains. Without translated memory, the context reference is replaced by the source reference. Without SPD retrieval, the covariance-derived context features are zero and similarity uses only embeddings. Without embedding retrieval, similarity uses only encoded covariance and the retrieved embedding value is zero. Without the reliability gate, its value is fixed to one. Without the residual adapter, feature and bias residuals are zero but the gated prototype-logit contribution remains. These are pathway interventions in the implemented model. In each comparison group, bold marks the best mean; ties share the same mark. The complete results show dataset-dependent effects: removing the gate slightly increases macro-F1 on SEED-IV (52.14 versus 52.08) and FACED-9 (16.28 versus 16.23), whereas the full model performs best on SEED and FACED-3. Thus the SEED ablations do not establish uniform component benefits across tasks.

Table 7: Component ablations. Each cell gives mean and SD (%).

Variant	Acc	Precision	Recall	F1	AUROC	AUPRC
SEED						
NoAdapt	72.97 ±8.57	74.31 ±8.13	72.99 ±8.55	72.37 ±9.08	88.45 ±5.33	80.43 ±8.62
w/o context encoder	73.79 ±8.23	75.00 ±7.57	73.79 ±8.24	73.49 ±8.23	88.56 ±5.33	80.52 ±8.53
w/o translated memory	73.64 ±8.89	74.74 ±8.37	73.64 ±8.90	73.34 ±8.93	88.63 ±5.46	80.67 ±8.84
w/o SPD retrieval	73.62 ±8.43	74.82 ±7.67	73.63 ±8.44	73.38 ±8.36	88.54 ±5.24	80.51 ±8.52
w/o embedding retrieval	73.59 ±8.02	74.76 ±7.50	73.61 ±8.04	73.26 ±8.06	**88.76** ±5.16	80.64 ±8.53
w/o reliability gate	73.28 ±8.34	74.24 ±8.00	73.31 ±8.35	72.85 ±8.61	88.41 ±5.33	80.13 ±8.69
w/o residual adapter	73.67 ±8.12	74.90 ±7.48	73.67 ±8.13	73.32 ±8.20	88.63 ±5.26	80.62 ±8.56
NeuroContext-TTA	**74.05** ±7.82	**75.25** ±7.21	**74.06** ±7.83	**73.72** ±7.92	**88.76** ±5.04	**80.68** ±8.39
SEED-IV						
NoAdapt	52.18 ±4.57	52.96 ±4.73	52.50 ±4.56	51.54 ±4.50	76.52 ±3.41	54.79 ±5.71
w/o context encoder	52.56 ±4.78	53.11 ±5.03	52.86 ±4.81	52.08 ±4.71	76.59 ±3.39	54.96 ±5.78
w/o translated memory	52.55 ±4.76	53.09 ±5.02	52.84 ±4.79	52.06 ±4.69	76.60 ±3.39	54.94 ±5.83
w/o SPD retrieval	52.57 ±4.75	53.12 ±5.01	52.86 ±4.80	52.09 ±4.68	76.58 ±3.39	54.97 ±5.76
w/o embedding retrieval	52.55 ±4.77	53.10 ±5.04	52.84 ±4.81	52.06 ±4.70	76.57 ±3.39	54.91 ±5.82
w/o reliability gate	**52.61** ±4.94	**53.15** ±5.10	**52.90** ±5.01	**52.14** ±4.84	**76.61** ±3.43	54.89 ±5.79
w/o residual adapter	52.57 ±4.76	53.12 ±5.03	52.86 ±4.81	52.08 ±4.70	76.59 ±3.39	54.95 ±5.80

Continued on next page

Table 7 continued

Variant	Acc	Precision	Recall	F1	AUROC	AUPRC
NeuroContext-TTA	52.56 ±4.74	53.12 ±5.00	52.86 ±4.79	52.08 ±4.68	**76.61** ±3.36	**55.02** ±5.69
FACED-3						
NoAdapt	48.60 ±3.64	33.08 ±5.19	36.00 ±2.68	30.99 ±5.05	58.68 ±5.09	40.28 ±4.14
w/o context encoder	49.55 ±3.60	33.56 ±3.26	36.59 ±2.67	33.36 ±3.65	58.69 ±5.09	40.29 ±4.14
w/o translated memory	49.55 ±3.57	33.56 ±3.24	36.59 ±2.65	33.36 ±3.64	58.69 ±5.09	40.29 ±4.15
w/o SPD retrieval	49.57 ±3.58	33.57 ±3.25	36.60 ±2.66	33.37 ±3.64	58.69 ±5.09	40.29 ±4.15
w/o embedding retrieval	49.56 ±3.59	33.57 ±3.25	36.60 ±2.66	33.37 ±3.64	58.69 ±5.09	40.29 ±4.15
w/o reliability gate	49.54 ±3.65	33.56 ±3.34	36.60 ±2.70	33.44 ±3.59	58.65 ±5.08	40.27 ±4.14
w/o residual adapter	49.52 ±3.60	33.53 ±3.23	36.57 ±2.67	33.33 ±3.67	58.69 ±5.09	40.29 ±4.15
NeuroContext-TTA	**50.61** ±4.49	**38.91** ±9.59	**37.71** ±3.38	**36.08** ±3.39	**59.59** ±5.88	**41.28** ±5.02
FACED-9						
NoAdapt	18.70 ±2.86	17.21 ±3.33	17.12 ±2.71	15.71 ±2.63	60.06 ±3.37	17.26 ±2.29
w/o context encoder	18.74 ±2.80	17.16 ±3.36	17.13 ±2.68	15.94 ±2.68	60.08 ±3.38	17.28 ±2.30
w/o translated memory	18.75 ±2.80	17.16 ±3.35	17.14 ±2.69	15.95 ±2.69	60.08 ±3.38	17.28 ±2.30
w/o SPD retrieval	18.76 ±2.79	17.14 ±3.34	17.16 ±2.67	15.96 ±2.68	60.13 ±3.37	17.30 ±2.30
w/o embedding retrieval	18.74 ±2.81	17.15 ±3.37	17.13 ±2.70	15.93 ±2.70	60.08 ±3.38	17.27 ±2.30
w/o reliability gate	18.94 ±2.82	17.53 ±3.46	17.33 ±2.73	**16.28** ±2.71	**60.29** ±3.45	**17.42** ±2.36
w/o residual adapter	18.73 ±2.79	17.18 ±3.35	17.13 ±2.68	15.93 ±2.68	60.07 ±3.37	17.27 ±2.30
NeuroContext-TTA	**19.05** ±3.01	**17.69** ±3.81	**17.45** ±2.83	16.23 ±2.73	60.16 ±3.46	17.38 ±2.35

G CONTEXT-SIZE SENSITIVITY

We vary the number of preceding windows. The first K windows are excluded from scoring, so changing K can change the scored suffix as well as the context. The primary settings are 256 for SEED and both FACED tasks and 512 for SEED-IV. Results do not imply monotonic improvement with longer history. In each comparison group, bold marks the best and underline the second-best distinct mean; ties share the same mark.

Table 8: Context-size sensitivity. Each cell gives mean and SD (%).

Context size K	Acc	Precision	Recall	F1	AUROC	AUPRC
SEED						
64	73.68 ±9.35	74.63 ±8.56	73.35 ±9.27	73.12 ±9.30	88.20 ±5.83	<u>80.10</u> ±9.18
128	**74.17** ±9.49	<u>75.24</u> ±8.92	<u>73.84</u> ±9.38	<u>73.56</u> ±9.49	<u>88.23</u> ±5.72	79.95 ±9.21
256	<u>74.05</u> ±7.82	**75.25** ±7.21	**74.06** ±7.83	**73.72** ±7.92	**88.76** ±5.04	**80.68** ±8.39
512	73.40 ±8.67	74.80 ±7.87	73.05 ±8.50	72.80 ±8.57	88.15 ±5.82	79.84 ±9.19
SEED-IV						
128	<u>52.26</u> ±5.04	<u>52.93</u> ±5.42	<u>52.68</u> ±5.28	<u>51.91</u> ±5.06	<u>76.50</u> ±3.61	55.01 ±6.06
256	<u>52.26</u> ±5.09	52.92 ±5.47	<u>52.68</u> ±5.32	<u>51.91</u> ±5.10	76.48 ±3.62	54.99 ±6.09
512	**52.56** ±4.74	**53.12** ±5.00	**52.86** ±4.79	**52.08** ±4.68	**76.61** ±3.36	<u>55.02</u> ±5.69
1024	<u>52.26</u> ±5.06	52.91 ±5.44	<u>52.68</u> ±5.30	<u>51.91</u> ±5.08	76.49 ±3.60	**55.06** ±6.00
FACED-3						
64	<u>50.47</u> ±3.56	<u>34.14</u> ±3.27	36.85 ±2.59	<u>33.78</u> ±3.64	58.78 ±5.18	40.24 ±4.02
128	50.36 ±3.37	34.01 ±3.22	<u>36.87</u> ±2.46	33.76 ±3.45	<u>58.80</u> ±5.17	40.30 ±4.13
256	**50.61** ±4.49	**38.91** ±9.59	**37.71** ±3.38	**36.08** ±3.39	**59.59** ±5.88	**41.28** ±5.02
512	49.63 ±3.95	33.77 ±3.57	36.59 ±2.88	33.35 ±3.93	58.44 ±5.39	<u>40.33</u> ±4.46
FACED-9						
64	<u>19.10</u> ±2.66	**17.71** ±3.32	<u>17.35</u> ±2.54	16.09 ±2.48	<u>60.29</u> ±3.30	17.20 ±2.20
128	19.05 ±2.83	17.59 ±3.18	<u>17.35</u> ±2.67	<u>16.20</u> ±2.60	**60.31** ±3.48	17.35 ±2.45
256	19.05 ±3.01	<u>17.69</u> ±3.81	**17.45** ±2.83	**16.23** ±2.73	60.16 ±3.46	<u>17.38</u> ±2.35
512	**19.51** ±3.12	17.21 ±3.97	17.28 ±3.00	<u>16.20</u> ±3.09	59.98 ±3.42	**17.89** ±2.48

H FEATURE AND COVARIANCE PERTURBATIONS

For a query descriptor matrix with final dimension d_u, we choose $m = \mathrm{round}(rd_u)$ dimensions at random at corruption rate r. Gaussian noise is added to those dimensions, with scale equal to the empirical standard deviation of the selected entries times `corrupt_std`; the default multiplier is 1.0. This operation is applied to both query embeddings and query log-covariance vectors. Context descriptors remain clean, and no feature is deliberately masked to zero. The table reports NeuroContext-TTA only; it supports a degradation analysis, not a paired robustness advantage over NoAdapt. In each comparison group, bold marks the best and underline the second-best distinct mean; ties share the same mark.

Table 9: Feature and covariance perturbations. Each cell gives mean and SD (%).

Corrupt rate	Acc	Precision	Recall	F1	AUROC	AUPRC
SEED						
0.0	**74.05** ±7.82	**75.25** ±7.21	**74.06** ±7.83	**73.72** ±7.92	**88.76** ±5.04	**80.68** ±8.39
0.1	<u>72.84</u> ±7.29	<u>73.80</u> ±6.84	<u>72.84</u> ±7.28	<u>72.44</u> ±7.47	<u>88.25</u> ±5.06	<u>79.87</u> ±8.26
0.2	72.42 ±7.32	73.27 ±6.91	72.41 ±7.29	72.01 ±7.49	87.71 ±5.16	78.97 ±8.30
0.3	71.88 ±7.25	72.66 ±6.94	71.87 ±7.23	71.47 ±7.44	87.31 ±5.19	78.29 ±8.30
SEED-IV						
0.0	**52.56** ±4.74	**53.12** ±5.00	**52.86** ±4.79	**52.08** ±4.68	**76.61** ±3.36	**55.02** ±5.69
0.1	<u>51.95</u> ±4.13	<u>52.37</u> ±4.50	<u>52.26</u> ±4.26	<u>51.48</u> ±4.13	<u>76.23</u> ±3.36	<u>54.13</u> ±5.68
0.2	51.43 ±3.83	51.75 ±4.18	51.76 ±3.99	50.96 ±3.86	75.79 ±3.36	53.36 ±5.53
0.3	50.94 ±3.62	51.29 ±3.95	51.27 ±3.82	50.50 ±3.63	75.39 ±3.37	52.68 ±5.45
FACED-3						
0.0	**50.61** ±4.49	**38.91** ±9.59	**37.71** ±3.38	**36.08** ±3.39	**59.59** ±5.88	**41.28** ±5.02
0.1	<u>50.10</u> ±4.17	38.04 ±7.54	<u>37.41</u> ±3.11	<u>35.89</u> ±3.14	<u>58.94</u> ±5.54	<u>40.64</u> ±4.58
0.2	49.51 ±4.10	38.06 ±8.47	36.96 ±3.16	35.46 ±3.27	58.18 ±5.33	39.99 ±4.35
0.3	49.35 ±3.90	<u>38.23</u> ±7.96	36.89 ±2.94	35.45 ±2.99	57.65 ±4.85	39.48 ±3.90
FACED-9						
0.0	**19.05** ±3.01	**17.69** ±3.81	**17.45** ±2.83	**16.23** ±2.73	**60.16** ±3.46	**17.38** ±2.35
0.1	<u>18.68</u> ±2.84	<u>17.17</u> ±3.11	<u>17.09</u> ±2.60	<u>15.96</u> ±2.47	<u>59.60</u> ±3.28	<u>16.99</u> ±2.16
0.2	18.08 ±2.80	16.59 ±3.33	16.53 ±2.59	15.46 ±2.51	58.85 ±3.16	16.48 ±2.01
0.3	17.62 ±2.39	16.21 ±2.82	16.10 ±2.19	15.10 ±2.07	58.42 ±3.00	16.13 ±1.79

I CONTEXT-SOURCE AND CONTEXT-QUALITY INTERVENTIONS

These interventions test sensitivity to the conditioning input. Target and source-derived contexts often yield similar values; shuffling gives identical results, as expected for permutation-invariant statistics. Zero and source-mean contexts can also match or exceed the target context on some metrics. Accordingly, the evidence does not establish that subject-matched context is necessary for the gains, or that the reliability gate detects mismatched context. The gate acts on query confidence and reference consistency. In each comparison group, bold marks the best and underline the second-best distinct mean; ties share the same mark.

Table 10: Context-source and context-quality interventions. Each cell gives mean and SD (%).

Context	Acc	Precision	Recall	F1	AUROC	AUPRC
SEED						
Target previous context	**74.05** ±7.82	**75.25** ±7.21	**74.06** ±7.83	**73.72** ±7.92	**88.76** ±5.04	**80.68** ±8.39
Random source same-index	73.87 ±7.96	75.10 ±7.29	73.88 ±7.97	73.55 ±8.04	88.72 ±5.04	80.65 ±8.35
Random source random-window	74.00 ±7.86	75.18 ±7.28	74.01 ±7.87	73.67 ±7.97	88.71 ±5.05	80.63 ±8.34
Shuffled target context	**74.05** ±7.82	**75.25** ±7.21	**74.06** ±7.83	**73.72** ±7.92	**88.76** ±5.04	**80.68** ±8.39
Zero context	73.72 ±7.76	74.92 ±7.22	73.72 ±7.78	73.38 ±7.86	88.58 ±4.97	80.44 ±8.19
Source-mean context	73.96 ±7.89	75.17 ±7.31	73.97 ±7.90	73.63 ±7.99	88.72 ±5.05	80.67 ±8.35
SEED-IV						
Target previous context	52.56 ±4.74	53.12 ±5.00	52.86 ±4.79	**52.08** ±4.68	76.61 ±3.36	55.02 ±5.69
Random source same-index	**52.57** ±4.73	53.12 ±4.99	52.86 ±4.78	**52.08** ±4.67	76.62 ±3.36	55.02 ±5.68
Random source random-window	**52.57** ±4.73	53.12 ±4.99	52.86 ±4.78	**52.08** ±4.66	76.62 ±3.35	55.04 ±5.66
Shuffled target context	52.56 ±4.74	53.12 ±5.00	52.86 ±4.79	**52.08** ±4.68	76.61 ±3.36	55.02 ±5.69
Zero context	**52.57** ±4.71	53.12 ±4.97	**52.87** ±4.76	**52.08** ±4.65	**76.63** ±3.35	**55.06** ±5.65
Source-mean context	**52.57** ±4.72	**53.13** ±4.97	**52.87** ±4.76	**52.08** ±4.65	**76.63** ±3.35	55.05 ±5.66
FACED-3						
Target previous context	50.61 ±4.49	38.91 ±9.59	37.71 ±3.38	36.08 ±3.39	59.59 ±5.88	41.28 ±5.02

Continued on next page

Table 10 continued

Context	Acc	Precision	Recall	F1	AUROC	AUPRC
Random source same-index	50.60 ±4.49	38.90 ±9.59	37.70 ±3.38	36.07 ±3.39	59.59 ±5.88	41.28 ±5.02
Random source random-window	50.63 ±4.49	38.92 ±9.58	37.72 ±3.38	36.09 ±3.39	59.59 ±5.87	41.28 ±5.01
Shuffled target context	50.61 ±4.49	38.91 ±9.59	37.71 ±3.38	36.08 ±3.39	59.59 ±5.88	41.28 ±5.02
Zero context	**50.64** ±4.49	38.92 ±9.59	37.72 ±3.38	**36.10** ±3.39	**59.60** ±5.87	**41.29** ±5.02
Source-mean context	**50.64** ±4.49	**38.93** ±9.59	**37.73** ±3.38	**36.10** ±3.39	**59.60** ±5.87	**41.29** ±5.02
FACED-9						
Target previous context	**19.05** ±3.01	**17.69** ±3.81	**17.45** ±2.83	**16.23** ±2.73	**60.16** ±3.46	**17.38** ±2.35
Random source same-index	**19.05** ±3.01	**17.69** ±3.81	17.44 ±2.83	**16.23** ±2.74	**60.16** ±3.46	**17.38** ±2.35
Random source random-window	**19.05** ±3.01	**17.69** ±3.81	**17.45** ±2.83	**16.23** ±2.74	**60.16** ±3.46	**17.38** ±2.35
Shuffled target context	**19.05** ±3.01	**17.69** ±3.81	**17.45** ±2.83	**16.23** ±2.73	**60.16** ±3.46	**17.38** ±2.35
Zero context	**19.05** ±3.01	**17.69** ±3.80	**17.45** ±2.83	**16.23** ±2.74	**60.16** ±3.46	**17.38** ±2.35
Source-mean context	**19.05** ±3.01	**17.69** ±3.80	**17.45** ±2.83	**16.23** ±2.74	**60.16** ±3.46	**17.38** ±2.35

J SEPARATE COMPARISON WITH FIXED-FEATURE TTA MODULES

This experiment uses a separate configuration series. Its NoAdapt and calibrated values differ from the main series, so it is evaluated only within this table group and is not used to claim cross-table superiority. The series includes 15 target-subject folds for SEED and SEED-IV and 123 for each FACED task. TENT, EATA, SAR and DeYO denote the evaluated EEG adaptations of those methods. Within this series, NeuroContext-TTA improves macro-F1 over NoAdapt by 1.35, 0.54, 3.00 and 0.37 points on SEED, SEED-IV, FACED-3 and FACED-9. The four alternative TTA modules remain close to NoAdapt, with small differences on SEED-IV. This comparison describes these configurations rather than a general failure of entropy-based adaptation. In each comparison group, bold marks the best and underline the second-best distinct mean; ties share the same mark.

Table 11: Separate comparison with fixed-feature TTA modules. Each cell gives mean and SD (%). NC denotes NeuroContext-TTA.

Method	Acc	Precision	Recall	F1	AUROC	AUPRC
SEED						
mdJPT (NoAdapt)	<u>72.97</u> ±8.57	<u>74.31</u> ±8.13	<u>72.99</u> ±8.55	<u>72.37</u> ±9.08	<u>88.45</u> ±5.33	<u>80.43</u> ±8.62
TENT-EEG	<u>72.97</u> ±8.56	<u>74.31</u> ±8.13	72.98 ±8.55	<u>72.37</u> ±9.07	<u>88.45</u> ±5.33	<u>80.43</u> ±8.62
EATA-EEG	<u>72.97</u> ±8.56	<u>74.31</u> ±8.13	<u>72.99</u> ±8.55	<u>72.37</u> ±9.07	<u>88.45</u> ±5.33	<u>80.43</u> ±8.62
SAR-EEG	<u>72.97</u> ±8.56	<u>74.31</u> ±8.13	<u>72.99</u> ±8.55	<u>72.37</u> ±9.07	<u>88.45</u> ±5.33	<u>80.43</u> ±8.62
DeYO-EEG	<u>72.97</u> ±8.56	<u>74.31</u> ±8.13	72.98 ±8.55	<u>72.37</u> ±9.07	<u>88.45</u> ±5.33	<u>80.43</u> ±8.62
mdJPT + NC	**74.05** ±7.82	**75.25** ±7.21	**74.06** ±7.83	**73.72** ±7.92	**88.76** ±5.04	**80.68** ±8.39
SEED-IV						
mdJPT (NoAdapt)	52.18 ±4.57	52.96 ±4.73	52.50 ±4.56	51.54 ±4.50	<u>76.52</u> ±3.41	<u>54.79</u> ±5.71
TENT-EEG	52.18 ±4.56	52.97 ±4.73	52.51 ±4.56	51.54 ±4.50	<u>76.52</u> ±3.41	<u>54.79</u> ±5.71
EATA-EEG	52.17 ±4.57	52.96 ±4.74	52.50 ±4.57	51.54 ±4.51	<u>76.52</u> ±3.41	<u>54.79</u> ±5.71
SAR-EEG	<u>52.19</u> ±4.56	52.97 ±4.73	<u>52.52</u> ±4.56	51.55 ±4.50	<u>76.52</u> ±3.41	<u>54.79</u> ±5.71
DeYO-EEG	<u>52.19</u> ±4.56	<u>52.98</u> ±4.73	<u>52.52</u> ±4.56	<u>51.56</u> ±4.50	<u>76.52</u> ±3.41	<u>54.79</u> ±5.72
mdJPT + NC	**52.56** ±4.74	**53.12** ±5.00	**52.86** ±4.79	**52.08** ±4.68	**76.61** ±3.36	**55.02** ±5.69
FACED-3						
mdJPT (NoAdapt)	<u>48.13</u> ±3.24	<u>33.97</u> ±4.87	<u>35.65</u> ±2.36	<u>30.10</u> ±4.90	<u>58.68</u> ±5.00	<u>40.35</u> ±3.96
TENT-EEG	<u>48.13</u> ±3.24	**33.98** ±4.87	<u>35.65</u> ±2.36	<u>30.10</u> ±4.90	<u>58.68</u> ±5.00	<u>40.35</u> ±3.96
EATA-EEG	<u>48.13</u> ±3.24	<u>33.97</u> ±4.87	<u>35.65</u> ±2.36	<u>30.10</u> ±4.90	<u>58.68</u> ±5.00	<u>40.35</u> ±3.96
SAR-EEG	<u>48.13</u> ±3.24	<u>33.97</u> ±4.87	<u>35.65</u> ±2.36	<u>30.10</u> ±4.90	<u>58.68</u> ±5.00	<u>40.35</u> ±3.96

Continued on next page

Table 11 continued

Method	Acc	Precision	Recall	F1	AUROC	AUPRC
DeYO-EEG	48.13 ±3.24	33.97 ±4.87	35.65 ±2.36	30.10 ±4.90	58.68 ±5.00	40.35 ±3.96
mdJPT + NC	**49.55** ±3.62	33.69 ±3.59	**36.59** ±2.70	**33.10** ±3.70	**58.69** ±5.00	**40.36** ±3.96
FACED-9						
mdJPT (NoAdapt)	18.84 ±3.07	17.49 ±3.66	17.28 ±2.85	15.86 ±2.66	60.15 ±3.45	17.37 ±2.33
TENT-EEG	18.84 ±3.07	17.49 ±3.66	17.27 ±2.85	15.86 ±2.66	60.15 ±3.45	17.37 ±2.33
EATA-EEG	18.84 ±3.07	17.49 ±3.66	17.28 ±2.85	15.86 ±2.66	60.15 ±3.45	17.37 ±2.33
SAR-EEG	18.84 ±3.07	17.49 ±3.66	17.28 ±2.85	15.86 ±2.66	60.15 ±3.45	17.37 ±2.33
DeYO-EEG	18.84 ±3.07	17.49 ±3.66	17.28 ±2.85	15.86 ±2.66	60.15 ±3.45	17.37 ±2.33
mdJPT + NC	**19.05** ±2.99	**17.69** ±3.79	**17.45** ±2.81	**16.23** ±2.72	**60.16** ±3.44	**17.38** ±2.34

K TRUE-CLASS MARGIN ANALYSIS ACROSS TARGET FOLDS

For a test sample with true class y, the true-class margin is

$$m(\ell, y) = \ell_y - \max_{c \neq y} \ell_c, \qquad \Delta m = m(\ell^{\text{after}}, y) - m(\ell^{\text{before}}, y). \tag{10}$$

Here the labels enter only the evaluation, never the adaptation procedure. Positive and negative margins correspond to correct and incorrect top-1 predictions, respectively, except at exact ties. The before and after values refer to the same query sample.

Each fold has three panels. The margin distribution overlays the before and after histograms. The paired plot places the before margin on the horizontal axis and the after margin on the vertical axis; points above the diagonal have increased margin. Crossings from negative to positive margins correspond to corrected predictions, and reverse crossings to new errors, apart from ties. The gain histogram displays Δm. Its annotations report the mean gain, the percentage with $\Delta m > 0$, and counts of incorrect-to-correct (fixed) and correct-to-incorrect (broken) predictions.

The main text selects two positive examples rather than a representative random sample. The following pages retain all 15 folds from each dataset, including negative and mixed outcomes. Each original vector figure retains its three-panel horizontal layout; plotted observations, axes and annotations are unchanged. Table 12 transcribes their displayed statistics. Counts are summed only over the supplied visualized query sets; no pooled accuracy percentage or significance test is inferred without sample denominators and the paired prediction arrays.

The diagnostics capture different aspects of prediction change. SEED has positive mean gain in 10/15 folds and more fixed than broken predictions in 10/15, with two ties. SEED-IV has positive mean gain in 13/15 folds but more fixed than broken predictions in 10/15. For example, SEED Fold 07 has a negative mean gain (-0.317) but 284 fixed versus 37 broken predictions. Conversely, SEED-IV Fold 04 has positive mean gain (0.076) but 36 fixed versus 106 broken predictions. Margin magnitudes depend on logit scale and need not track decision-boundary crossings. These plots assess discriminative correction, not probabilistic calibration, physiological alignment or uniform benefit to every subject.

Table 12: Statistics transcribed from all 30 supplied fold visualizations. Mean gain is in logit units; improved is the percentage of samples with increased margin. Net equals fixed minus broken.

Dataset	Fold	Mean gain	Improved (%)	Fixed	Broken	Net
SEED	00	0.086	52.9	39	29	+10
SEED	01	−0.023	51.0	62	6	+56
SEED	02	0.046	56.8	0	0	+0
SEED	03	0.568	61.2	290	6	+284
SEED	04	0.015	44.0	44	28	+16
SEED	05	−0.085	42.6	52	28	+24
SEED	06	−0.081	52.4	79	137	−58
SEED	07	−0.317	36.4	284	37	+247
SEED	08	0.039	60.4	2	0	+2
SEED	09	0.541	61.5	138	18	+120
SEED	10	0.046	58.6	0	9	−9
SEED	11	−0.239	45.7	41	7	+34
SEED	12	0.097	64.4	0	0	+0
SEED	13	0.024	60.6	43	0	+43
SEED	14	0.149	70.6	0	2	−2
SEED-IV	00	0.006	54.2	29	15	+14
SEED-IV	01	−0.011	51.8	2	110	−108
SEED-IV	02	0.123	60.1	102	28	+74
SEED-IV	03	0.129	53.8	275	5	+270
SEED-IV	04	0.076	59.9	36	106	−70

Continued on next page

Table 12 continued

Dataset	Fold	Mean gain	Improved (%)	Fixed	Broken	Net
SEED-IV	05	0.072	56.7	44	34	+10
SEED-IV	06	0.025	58.3	96	5	+91
SEED-IV	07	−0.015	45.8	90	72	+18
SEED-IV	08	0.027	52.2	32	34	−2
SEED-IV	09	0.100	62.0	40	44	−4
SEED-IV	10	0.059	59.2	80	4	+76
SEED-IV	11	0.029	55.0	57	92	−35
SEED-IV	12	0.047	57.9	61	9	+52
SEED-IV	13	0.116	63.4	73	53	+20
SEED-IV	14	0.148	56.4	159	9	+150

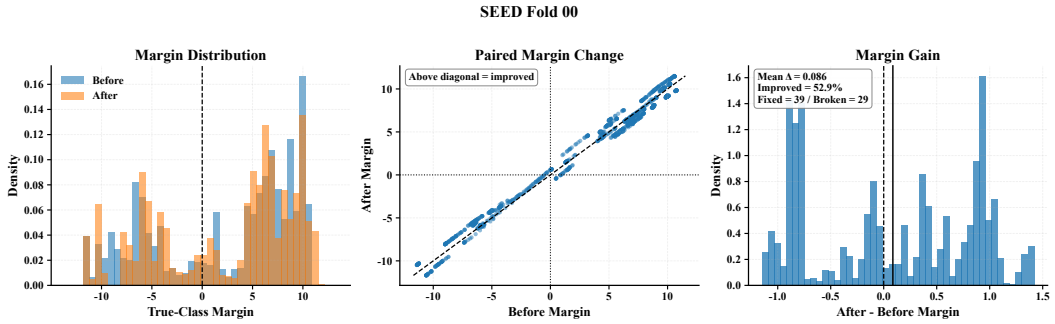


Figure 5: SEED, target Fold 00. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain 0.086; improved 52.9%; fixed 39, broken 29.

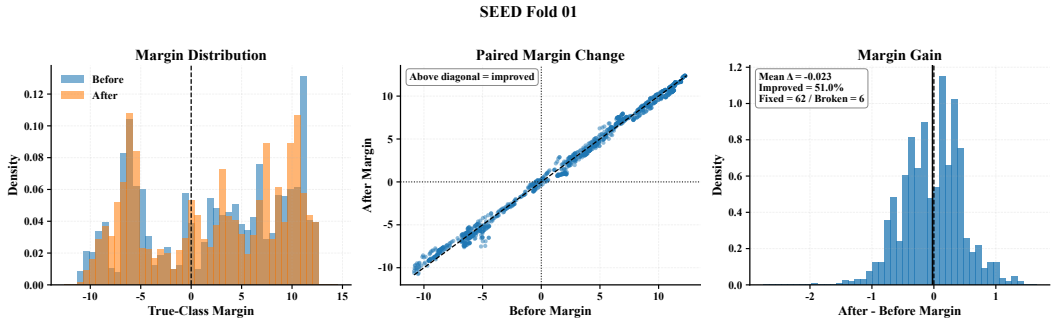


Figure 6: SEED, target Fold 01. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain -0.023; improved 51.0%; fixed 62, broken 6.

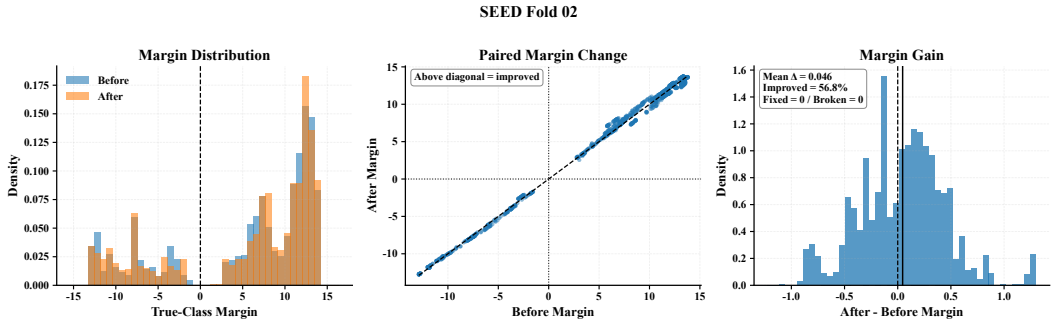


Figure 7: SEED, target Fold 02. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain 0.046; improved 56.8%; fixed 0, broken 0.

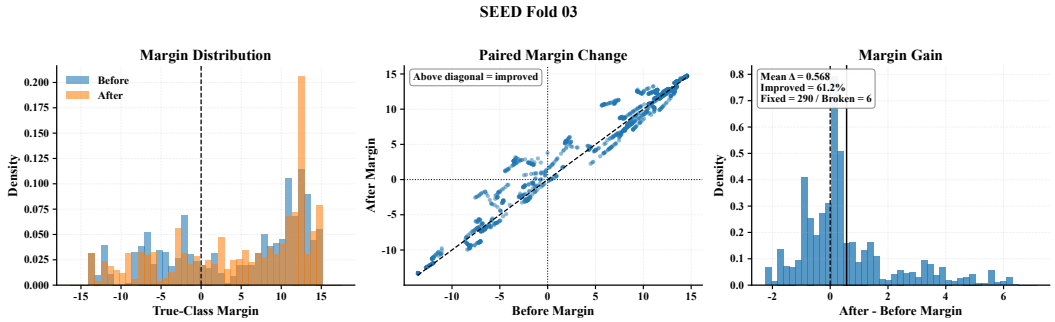


Figure 8: SEED, target Fold 03. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain 0.568; improved 61.2%; fixed 290, broken 6.

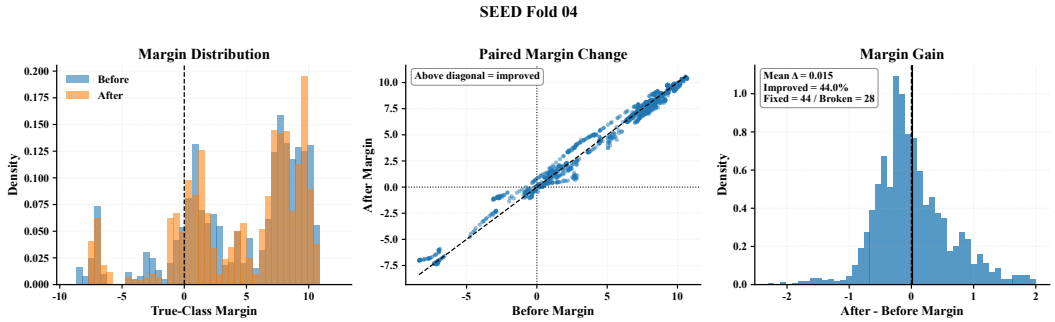


Figure 9: SEED, target Fold 04. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain 0.015; improved 44.0%; fixed 44, broken 28.

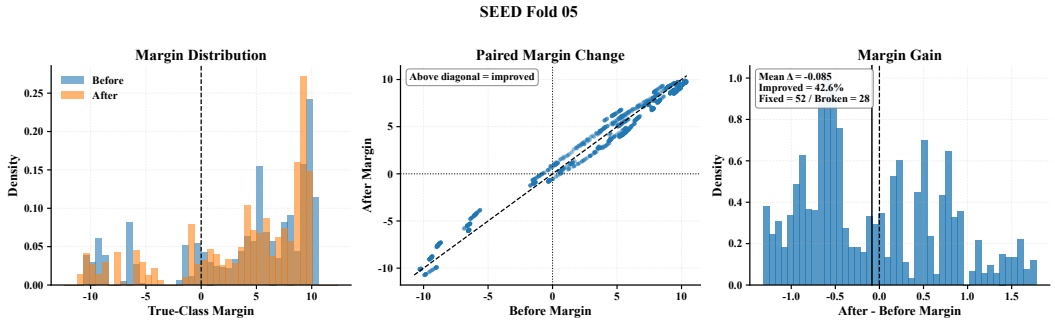


Figure 10: SEED, target Fold 05. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain -0.085; improved 42.6%; fixed 52, broken 28.

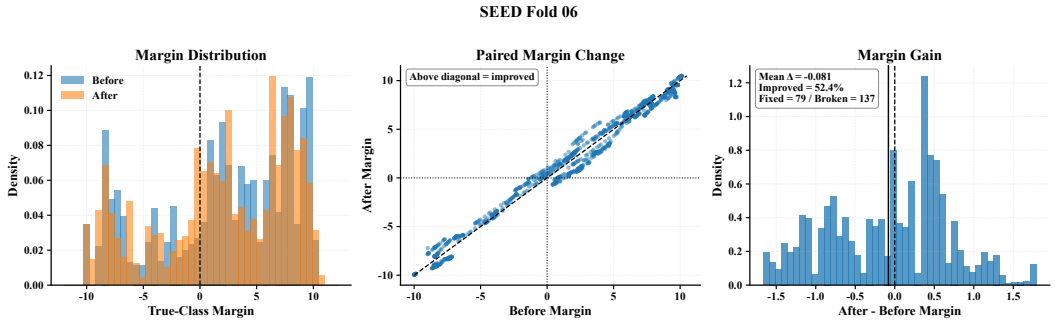


Figure 11: SEED, target Fold 06. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain -0.081; improved 52.4%; fixed 79, broken 137.

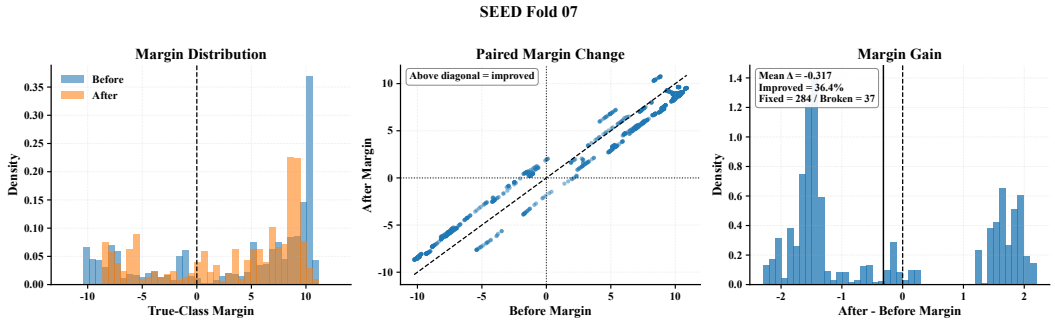


Figure 12: SEED, target Fold 07. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain -0.317; improved 36.4%; fixed 284, broken 37.

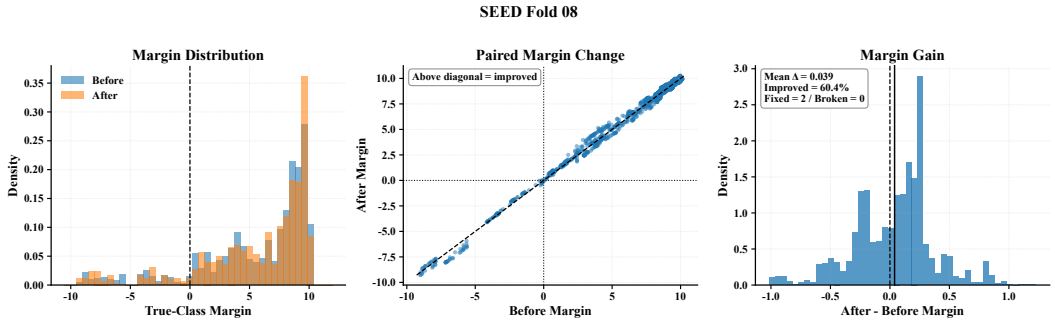


Figure 13: SEED, target Fold 08. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain 0.039; improved 60.4%; fixed 2, broken 0.

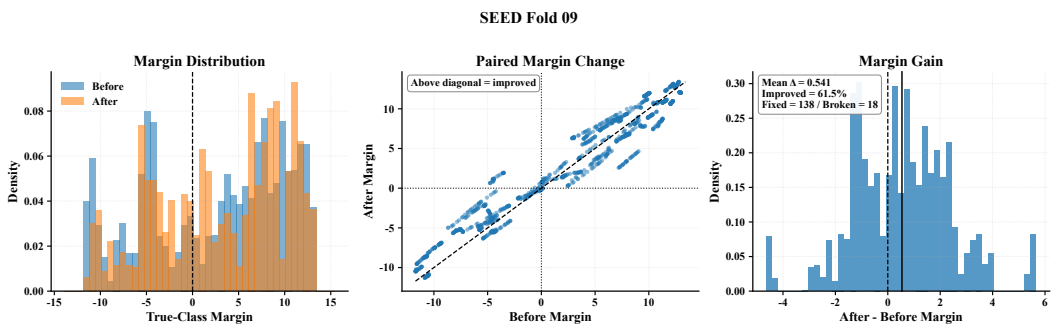


Figure 14: SEED, target Fold 09. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain 0.541; improved 61.5%; fixed 138, broken 18.

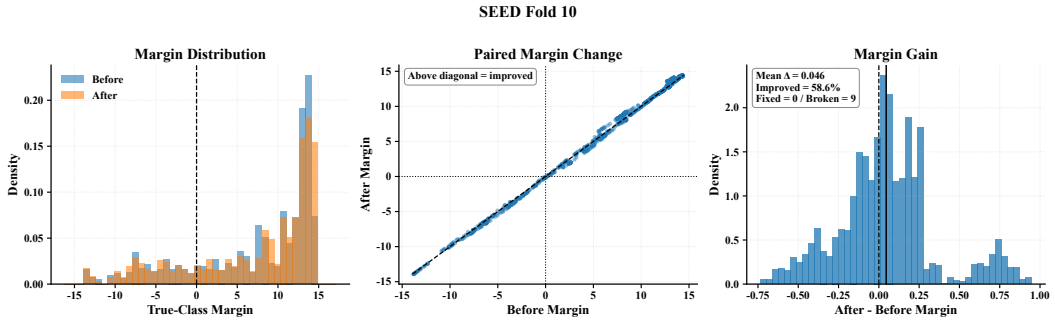


Figure 15: SEED, target Fold 10. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain 0.046; improved 58.6%; fixed 0, broken 9.

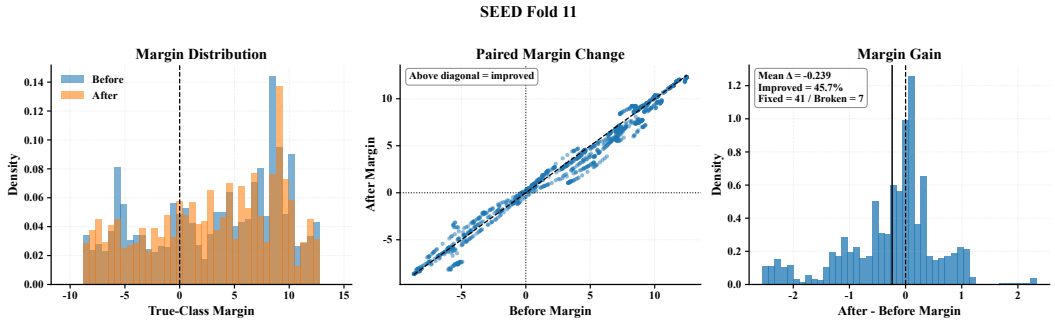


Figure 16: SEED, target Fold 11. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain -0.239; improved 45.7%; fixed 41, broken 7.

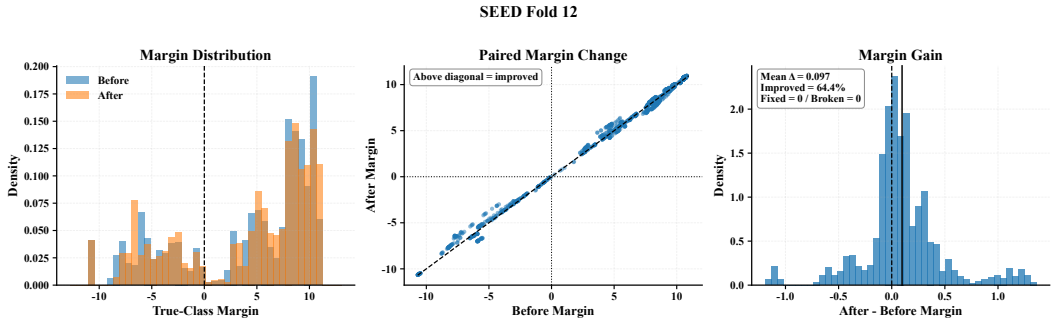


Figure 17: SEED, target Fold 12. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain 0.097; improved 64.4%; fixed 0, broken 0.

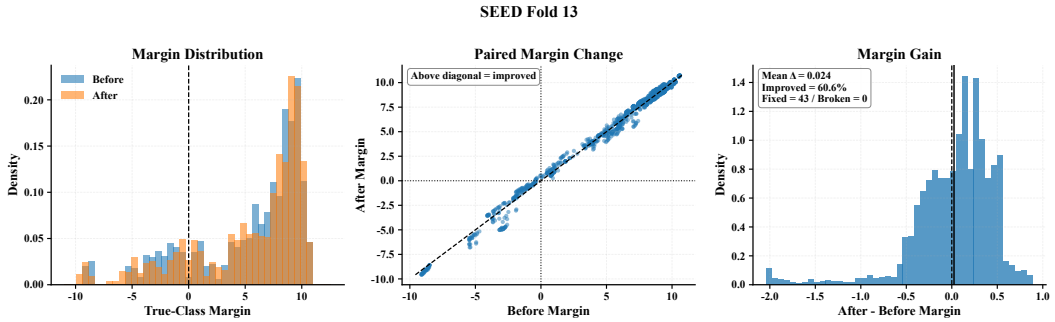


Figure 18: SEED, target Fold 13. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain 0.024; improved 60.6%; fixed 43, broken 0.

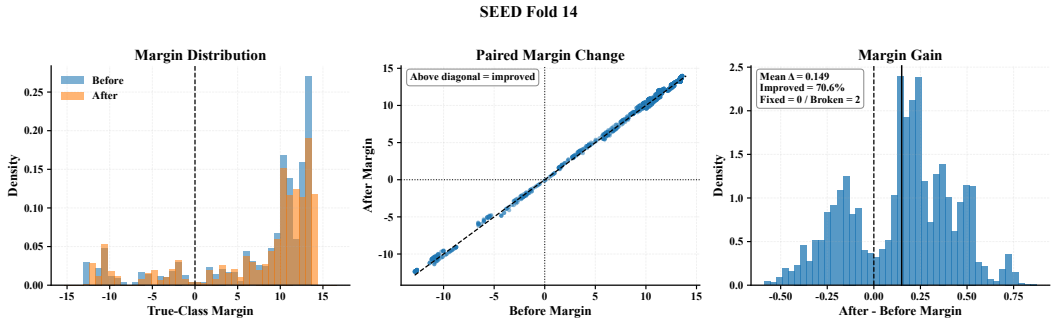


Figure 19: SEED, target Fold 14. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain 0.149; improved 70.6%; fixed 0, broken 2.

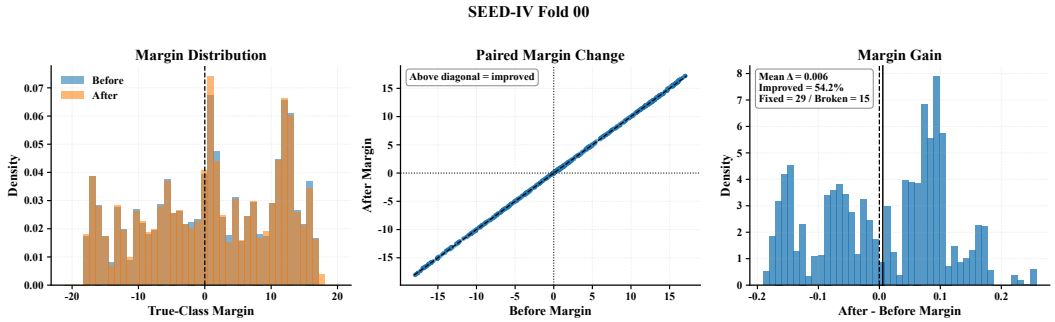


Figure 20: SEED-IV, target Fold 00. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain 0.006; improved 54.2%; fixed 29, broken 15.

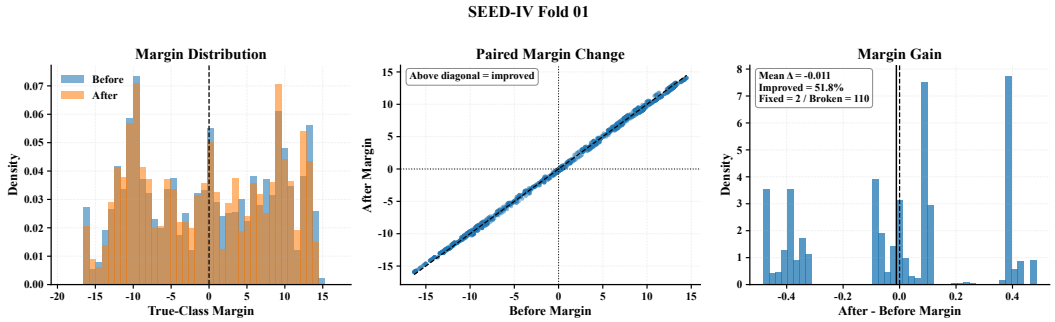


Figure 21: SEED-IV, target Fold 01. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain -0.011; improved 51.8%; fixed 2, broken 110.

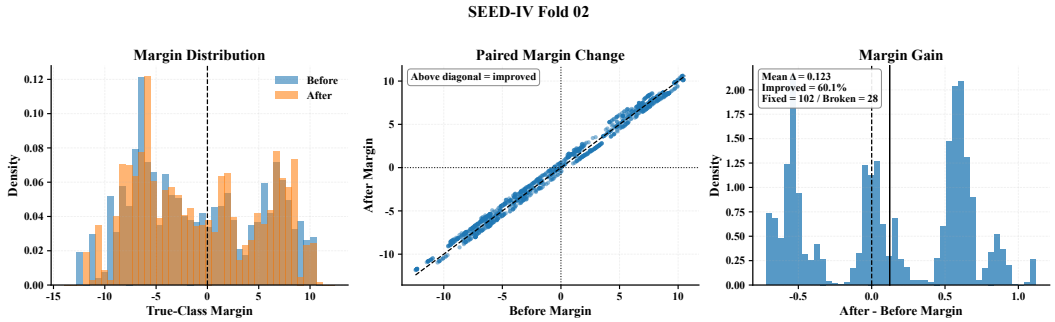


Figure 22: SEED-IV, target Fold 02. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain 0.123; improved 60.1%; fixed 102, broken 28.

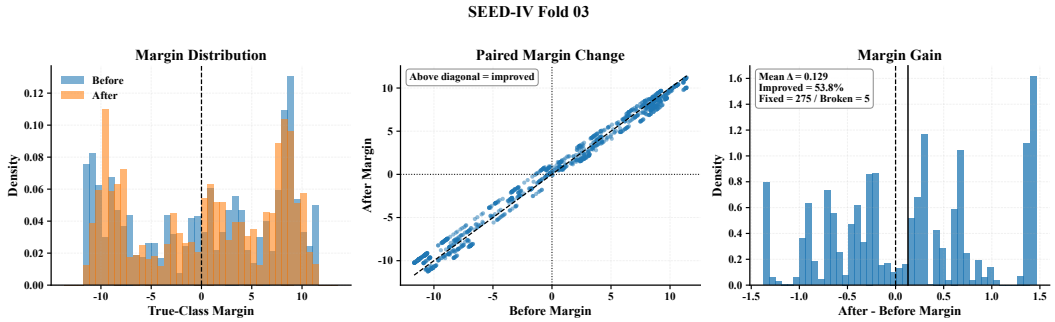


Figure 23: SEED-IV, target Fold 03. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain 0.129; improved 53.8%; fixed 275, broken 5.

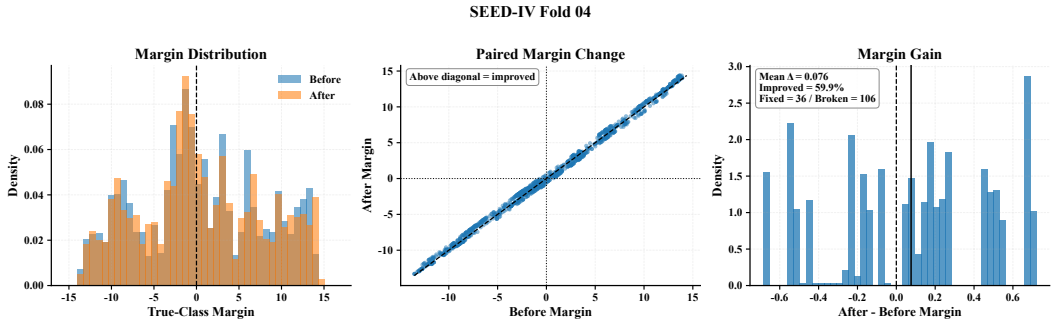


Figure 24: SEED-IV, target Fold 04. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain 0.076; improved 59.9%; fixed 36, broken 106.

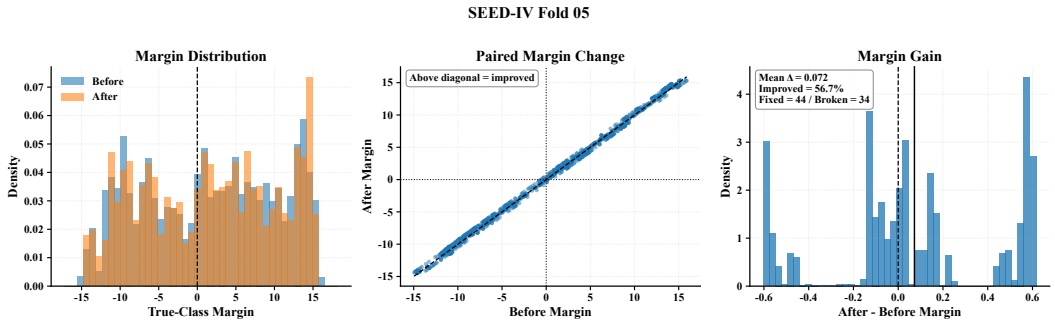


Figure 25: SEED-IV, target Fold 05. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain 0.072; improved 56.7%; fixed 44, broken 34.

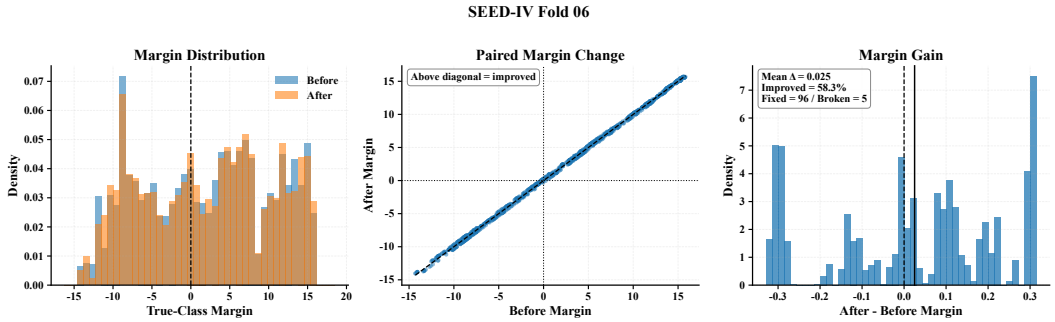


Figure 26: SEED-IV, target Fold 06. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain 0.025; improved 58.3%; fixed 96, broken 5.

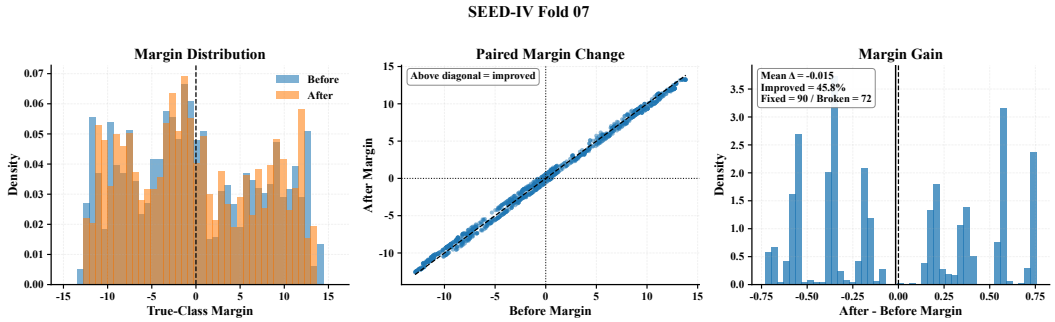


Figure 27: SEED-IV, target Fold 07. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain -0.015; improved 45.8%; fixed 90, broken 72.

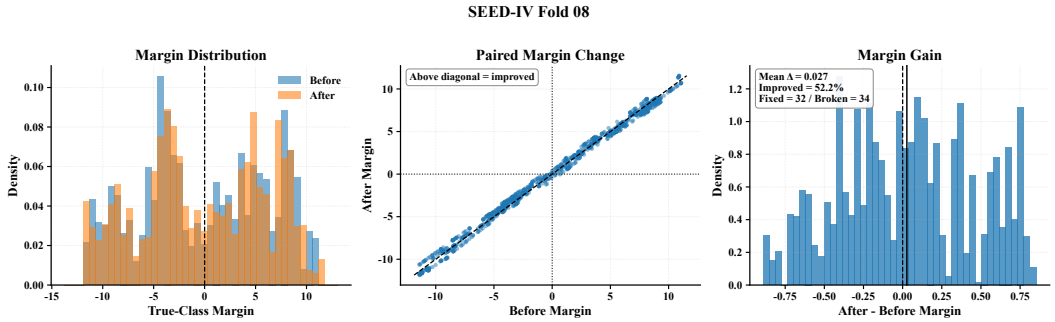


Figure 28: SEED-IV, target Fold 08. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain 0.027; improved 52.2%; fixed 32, broken 34.

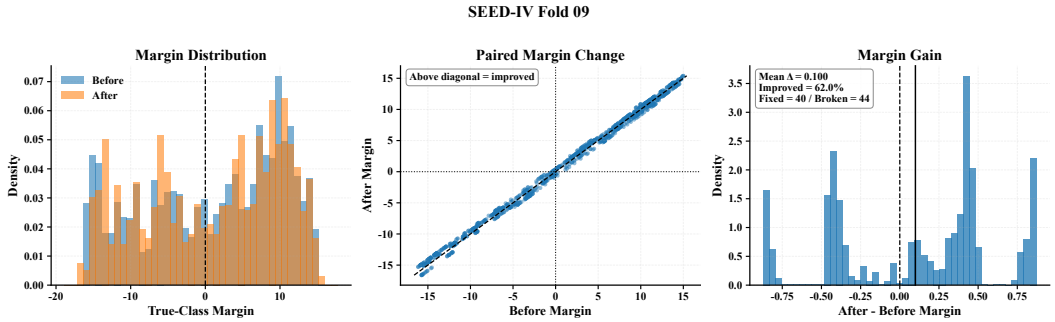


Figure 29: SEED-IV, target Fold 09. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain 0.100; improved 62.0%; fixed 40, broken 44.

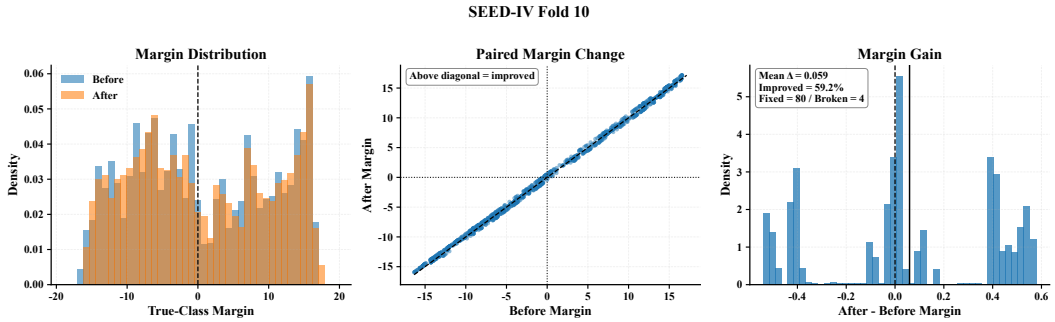


Figure 30: SEED-IV, target Fold 10. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain 0.059; improved 59.2%; fixed 80, broken 4.

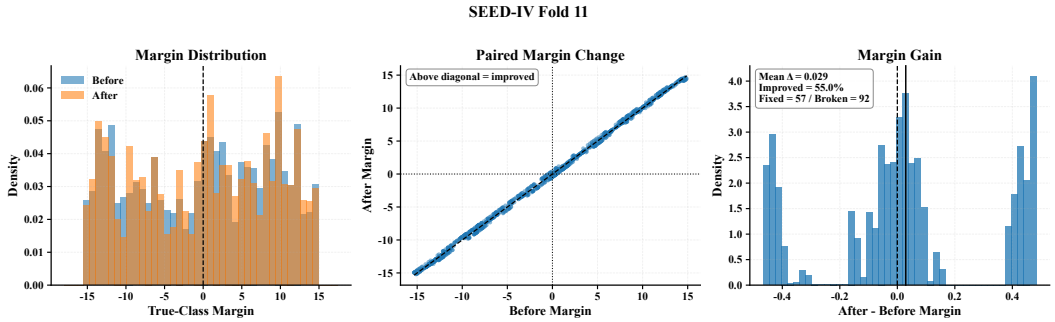


Figure 31: SEED-IV, target Fold 11. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain 0.029; improved 55.0%; fixed 57, broken 92.

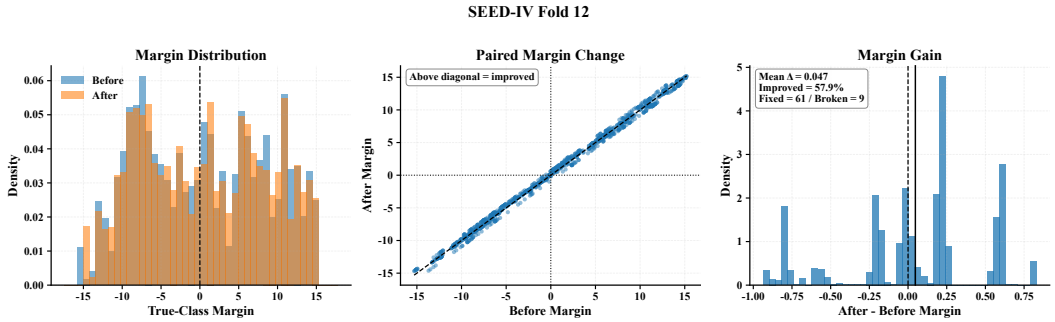


Figure 32: SEED-IV, target Fold 12. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain 0.047; improved 57.9%; fixed 61, broken 9.

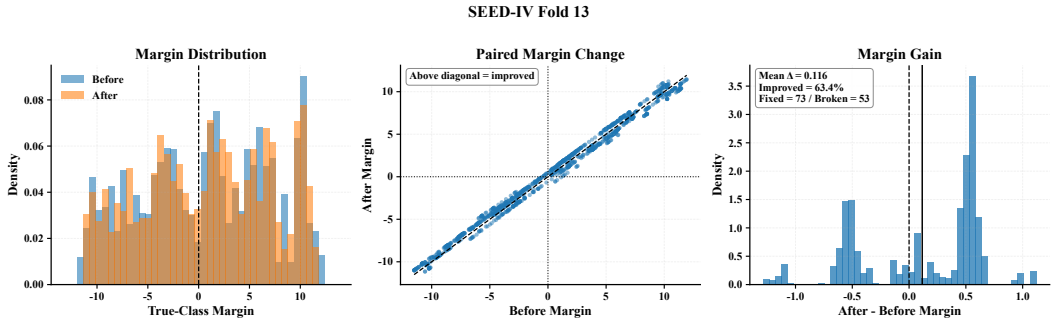


Figure 33: SEED-IV, target Fold 13. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain 0.116; improved 63.4%; fixed 73, broken 53.

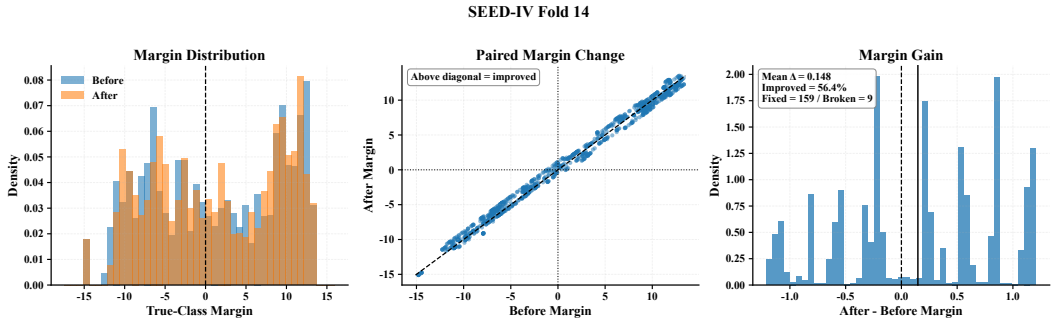


Figure 34: SEED-IV, target Fold 14. Left: before/after margin distributions. Middle: paired margins. Right: margin gains. Mean gain 0.148; improved 56.4%; fixed 159, broken 9.