Test-Time Adaptation for EEG Foundation Models: A Systematic Study under Real-World Distribution Shifts

Gabriel Jason Lee* GJLEE4@ILLINOIS.EDU

Jathurshan Pradeepkumar* JP65@ILLINOIS.EDU

Jimeng Sun JIMENG@ILLINOIS.EDU

University of Illinois Urbana-Champaign, Urbana, IL, USA

Abstract

Electroencephalography (EEG) foundation models have shown strong potential for learning generalizable representations from large-scale neural data, yet their clinical deployment is hindered by distribution shifts across clinical settings, devices, and populations. Test-time adaptation (TTA) offers a promising solution by enabling models to adapt to unlabeled target data during inference without access to source data, a valuable property in healthcare settings constrained by privacy regulations and limited labeled data. However, its effectiveness for EEG remains largely underexplored. In this work, we introduce *NeuroAdapt-Bench*, a systematic benchmark for evaluating test-time adaptation methods on EEG foundation models under realistic distribution shifts. We evaluate representative TTA approaches from other domains across multiple pretrained foundation models, diverse downstream tasks, and heterogeneous datasets spanning in-distribution, out-of-distribution, and extreme modality shifts (e.g., Ear-EEG). Our results show that the evaluated TTA methods yield inconsistent gains and often degrade performance, with gradient-based approaches particularly prone to heavy degradation and optimization-free methods showing greater stability. For the evaluated EEG foundation models and representative TTA methods, these findings highlight the limitations of directly applying existing TTA techniques to EEG and underscore the need for domain-specific adaptation strategies.

Code is available at https://github.com/leegabriel/NeuroAdapt-Bench.

1. Introduction

Electroencephalography (EEGs) offer high-resolution measurements of neuronal activity, capturing brain dynamics at the millisecond scale, making them essential for a wide range of clinical applications, including sleep staging (Phan et al., 2022; Pradeepkumar et al., 2024) and epilepsy diagnosis (Sundaram et al., 1999; Jia et al., 2026). Recent advances in self-supervised learning have led to the development of EEG foundation models (Ouahidi et al., 2025; Wang et al., 2025a), which are large neural networks trained on diverse, large-scale EEG corpora to learn generalizable representations. Despite their success, a key barrier to clinical deployment remains: *distribution shift*, where models trained on a given dataset often fail to generalize to new hospitals, acquisition devices, or patient populations.

As illustrated in Figure 1, distribution shifts are especially severe in EEG analysis. Unlike natural images, where domain gaps are often stylistic, EEG signals exhibit complex,

patient-specific dynamics and diverse acquisition protocols that vary substantially across sessions, tasks, and clinical sites (Jayaram et al., 2016; Yang et al., 2023b). For example, (Kastrati et al., 2025) reports substantial performance degradation of EEG foundation models on out-of-distribution tasks such as sleep staging.

Test-time adaptation (TTA) offers a promising solution by enabling models to adapt to target-domain data without labeled samples or access to source data, unlike traditional domain adaptation. This source-free property is particularly valuable in healthcare, where access to source data is often restricted by privacy regulations, limited labeled data, and the computational overhead of model fine-tuning. Prior TTA work in computer vision and speech has introduced a range of strategies, including entropy minimization, continual self-training, prototype adjustment, and source-free pseudo-label refinement (Wang et al., 2021, 2022; Iwasawa and Matsuo, 2021; Liang et al., 2020; Liu et al., 2024; Wang et al., 2025b).

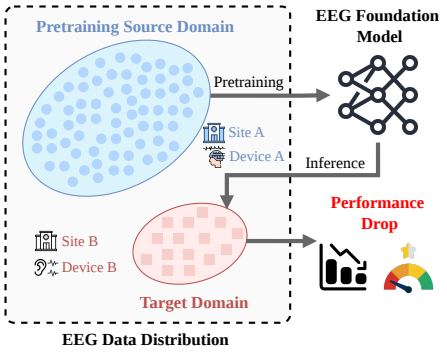

Figure 1: Distribution shift in EEG foundation model deployment. Pretrained EEG models often degrade when applied to new sites and devices, motivating the need for label and source-free test-time adaptation.

Despite growing interest in TTA across computer vision and speech recognition, its application to EEG remains underexplored. Existing studies typically focus on a single task or architecture, such as driver drowsiness detection or multimodal sleep staging (Jang et al., 2025; Guo et al., 2025; Jia et al., 2024), which makes it difficult to tell whether observed gains generalize across settings. At the same time, EEG foundation models are explicitly motivated by transfer across datasets and downstream tasks, yet there is still little evidence on how standard TTA methods behave when these models are deployed under realistic EEG distribution shifts. This leaves an important practical gap between pretrained EEG representation learning and reliable deployment.

To address this gap, we conduct a systematic benchmark of test-time adaptation for EEG foundation models. We evaluate representative TTA methods across multiple pretrained EEG foundation models, diverse downstream datasets, and heterogeneous deployment settings, encompassing a range of distribution shifts and tasks, including event detection, abnormality screening, seizure detection, and sleep staging. Overall our contributions are summarized as follows:

- **NeuroAdapt-Bench:** We introduce NeuroAdapt-Bench, a unified benchmark for evaluating test-time adaptation methods on EEG foundation models under realistic distribution shifts.
- **Comprehensive experiments under diverse distribution shifts:** We systematically evaluate TTA methods across a range of EEG tasks and deployment scenarios, including both online and offline adaptation regimes. Our experiments explicitly cover (1) *in-distribution* settings capturing subject-level variability, (2) *out-of-distribution* cross-dataset shifts involving changes in tasks, populations, and acquisition protocols,

and (3) *extreme distribution shifts* arising from unseen modalities and recording configurations (e.g., Ear-EEG Tabar et al. (2024)). We quantify performance relative to a No-TTA baseline across all settings.

- **Key insights on TTA for EEG:** Across the evaluated foundation models and TTA methods, we show that the evaluated TTA methods yield inconsistent gains and often degrade performance under distribution shift. We find that evaluated optimization-free methods (e.g., prototype-based approaches) are generally more stable than gradient-based alternatives, highlighting stability as a central consideration for deployment. In particular, T3A is the only method with a positive mean balanced-accuracy improvement across the in-distribution, out-of-distribution, and extreme-shift Ear-EEG settings. Its largest mean gain is a +18.9 percentage-point improvement in balanced accuracy for REVE-Base on CHB-MIT (Shoeb and Guttag, 2010). Detailed per-dataset deltas, relative changes, and statistical significance tests are provided in Appendices A.5,A.6, and A.7.
- **Open-source benchmark framework:** We release code and evaluation pipelines to facilitate benchmarking of future EEG foundation models and TTA methods. The benchmark will be integrated into an existing Python library to facilitate reproducibility and future research in this domain.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our study highlights several generalizable insights for deploying machine learning in healthcare. First, methods such as test-time adaptation that perform well in domains like computer vision do not necessarily transfer well to EEG signals, where they can introduce instability and degrade performance under realistic distribution shifts. Second, stability and robustness are as critical as accuracy for clinical deployment, and the evaluated optimization-free approaches are more reliable than gradient-based methods in our benchmark. Third, the type and severity of distribution shift, ranging from subject variability to cross-dataset and modality-level differences, strongly impact model performance, underscoring the need for evaluation under realistic conditions. Finally, standardized and reproducible benchmarking frameworks are essential for identifying failure modes and guiding the development of reliable healthcare AI systems.

2. Related Work

2.1. EEG Foundation Models

Recent advances in large-scale self-supervised pretraining have driven the rapid development of EEG foundation models. These models are motivated by the need to generalize across heterogeneous EEG settings, including differences in subjects, channel configurations, acquisition protocols, and task definitions. However, recent reviews emphasize that current EEG foundation models remain highly heterogeneous in their pretraining data, architectures, and evaluation, leaving their robustness under realistic deployment shift only partially understood (Yao et al., 2025; Kuruppu et al., 2026).

Broadly, existing EEG foundation models can be categorized into encoder-only and generative models. Encoder-only models, including BIOT (Yang et al., 2023a), LaBraM (Jiang et al., 2024), CBraMod (Wang et al., 2024b), REVE (Ouahidi et al., 2025), EEGPT (Wang et al., 2024a), and TFM-Tokenizer (Pradeepkumar et al., 2026b) are primarily optimized for discriminative tasks such as classification. In contrast, generative EEG foundation models (Pradeepkumar et al., 2026a; Xu et al., 2026) focus on language alignment and generative objectives. In this work, we benchmark TTA methods on encoder-based models.

2.2. Test-Time Adaptation

Test-time adaptation considers the setting in which a model trained on labeled source-domain data is deployed on unlabeled target-domain data drawn from a shifted distribution. Since deployment-time shift can substantially degrade performance, TTA aims to adapt the source-trained model during inference using only target samples available at test time, often without access to source data or target labels (Wang et al., 2025c). Prior work in computer vision has established several common TTA families, including entropy minimization, continual self-training, prototype-based adjustment, and source-free pseudo-label refinement (Wang et al., 2021; Niu et al., 2022; Wang et al., 2022; Iwasawa and Matsuo, 2021; Liang et al., 2020). Similar approaches have recently emerged in speech and audio applications under noisy and mismatched deployment conditions (Lin et al., 2024; Liu et al., 2024; Wang et al., 2025b; Dong et al., 2025).

Despite progress, TTA for biosignals remains relatively limited and largely task-specific. Recent works have explored approaches such as personalized calibration, teacher–student adaptation, and memory-based stabilization in applications including sleep staging, rPPG, and ECG (Jo et al., 2025; Guo et al., 2025; Jia et al., 2024; Huang et al., 2026; Wu et al., 2026). EEG is particularly challenging in this context due to its highly non-stationary nature across subjects and sessions, weakly structured relative to signals such as ECG, and sensitivity to artifacts and acquisition variability (Raj et al., 2025). While initial EEG-specific TTA studies show promising gains in narrowly defined tasks, such as driver drowsiness classification (Jang et al., 2025), they do not provide a comprehensive understanding of how standard TTA methods generalize across EEG foundation models and downstream tasks.

Our work bridges two emerging directions: EEG foundation models and test-time adaptation. Here, we systematically benchmark representative TTA approaches across multiple EEG foundation models under diverse downstream settings. This enables us to assess not only when adaptation improves performance, but also when it fails, which methods are most stable, and the implications for clinical deployment.

3. NeuroAdapt-Bench

This section introduces *NeuroAdapt*, our benchmark for systematically evaluating representative TTA methods on EEG foundation models across diverse tasks. Section 3.1 presents the problem formulation and describes the TTA methods considered. Sections 3.2 and 3.3 detail the benchmark design and experiment setups.

Table 1: Classification of the evaluated TTA methods by adaptation regime and update mechanism.

Method	Online	Batch Adaptation	Gradient Based
Tent	✓		✓
T3A	✓		
SHOT		✓	✓

3.1. Preliminary

In test-time adaptation, a model trained on labeled source-domain data is deployed on unlabeled target-domain data whose distribution may differ from that of the source domain (Sun et al., 2020). The objective is to adapt the source-trained model using only unlabeled target samples observed during inference. In this work, we consider three representative methods for this setting (Table 1): Tent (Wang et al., 2021), SHOT (Liang et al., 2020), and T3A (Iwasawa and Matsuo, 2021).

Under a unified formulation, we represent a source-trained classifier as:

$$f_\theta(x) = h_w(g_\phi(x)) \tag{1}$$

where g_ϕ denotes the feature extractor parameterized by ϕ, h_w is the classifier head parameterized by w, and $\theta = (\phi, w)$ represents the full set of model parameters. Given an input x, the model outputs logits $f_\theta(x)$, from which the predictive distribution $p_\theta(y \mid x)$ is obtained via the softmax function. The model's predictive entropy for a sample x is defined as:

$$H(p_\theta(\cdot \mid x)) = -\sum_{k=1}^{K} p_\theta(y = k \mid x) \log p_\theta(y = k \mid x). \tag{2}$$

Tent (Wang et al., 2021) performs test-time adaptation by minimizing prediction entropy on an unlabeled target batch B_t:

$$\mathcal{L}_{\text{Tent}} = \frac{1}{|B_t|} \sum_{x \in B_t} H(p_\theta(\cdot \mid x)) \,. \tag{3}$$

By minimizing predictive entropy, Tent encourages confident predictions on target samples under distribution shift. At test time, adaptation is restricted to the affine parameters of normalization layers, while the remaining network parameters are held fixed.

SHOT (Liang et al., 2020) assumes source-free adaptation and keeps the classifier h_w fixed while adapting only the target feature extractor g_ϕ. Its objective consists of three terms:

$$\mathcal{L}_{\text{SHOT}} = \mathcal{L}_{\text{ent}} + \mathcal{L}_{\text{div}} + \beta \mathcal{L}_{\text{PL}}, \tag{4}$$

where,

$$\mathcal{L}_{\text{ent}} = -\mathbb{E}_{x \sim \mathcal{X}_t} \sum_{k=1}^{K} p_k(x) \log p_k(x), \tag{5}$$

$$\mathcal{L}_{\text{div}} = \sum_{k=1}^{K} \hat{p}_k \log \hat{p}_k, \qquad \hat{p}_k = \mathbb{E}_{x \sim \mathcal{X}_t}[p_k(x)]. \tag{6}$$

Here, $\mathcal{L}_{\text{ent}}$ encourages confident predictions on target samples, $\mathcal{L}_{\text{div}}$ promotes diversity in the marginal output distribution and prevents degenerate collapse to a single class, and $\mathcal{L}_{\text{PL}}$ denotes the pseudo-label loss.

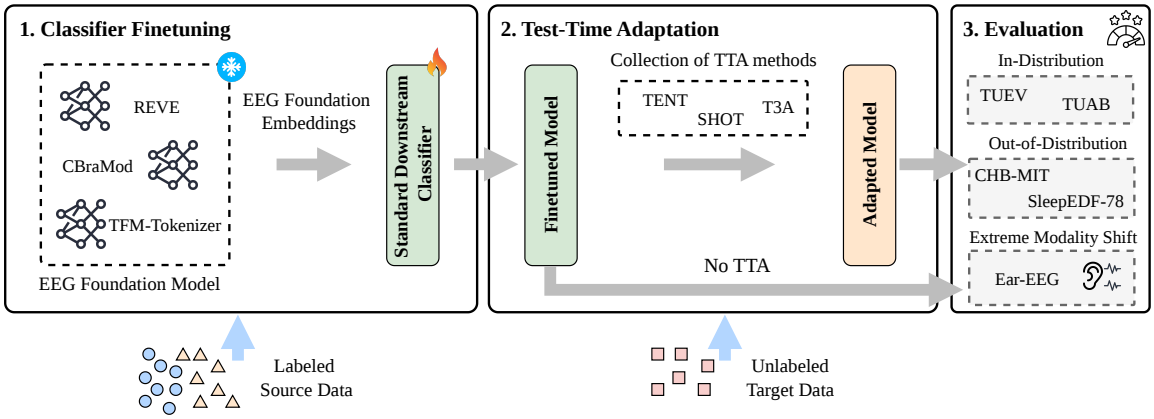

Figure 2: Overview of NeuroAdapt-Bench. The benchmark consists of three stages: (1) supervised finetuning of an EEG foundation model (e.g., REVE, CBRaMod, or TFM-Tokenizer) with a classification head on labeled source-domain data; (2) optional test-time adaptation on unlabeled target-domain data using methods such as TENT, SHOT, or T3A, alongside a no-adaptation baseline; and (3) evaluation on target-domain samples to measure performance and robustness under distribution shift.

T3A (Iwasawa and Matsuo, 2021) is an optimization-free method that keeps the feature extractor g_ϕ fixed and adapts the classifier online using target features. Formally, let $z = g_\phi(x)$ denote the feature representation of x, and let $\{\omega_k\}_{k=1}^{K}$ denote the source classifier weight vectors. For each class k, T3A maintains a support set $S_k^{(t)}$ at test-time step t, consisting of target feature vectors assigned to that class. The class template is then computed as the mean of the corresponding support set:

$$c_k^{(t)} = \frac{1}{|S_k^{(t)}|} \sum_{z \in S_k^{(t)}} z. \tag{7}$$

Prediction is then performed using the adjusted classifier

$$p(y = k \mid z) \propto \exp\left(z^\top c_k^{(t)}\right), \tag{8}$$

where $c_k^{(t)}$ is the class prototype for class k at time t. In this way, T3A adapts the classifier geometry directly by refining class prototypes from target test features, without updating network parameters through gradient-based optimization.

3.2. NeuroAdapt-Bench Design

As illustrated in Figure 2, NeuroAdapt-Bench follows a three-stage pipeline: (1) classifier fine-tuning, (2) test-time adaptation, and (3) evaluation.

Stage 1: Classifier Fine-Tuning Each foundation model is paired with the same lightweight classification head, replacing any model-specific heads used in prior work, to control for classifier architecture as a confounding factor in cross-model comparison. This

design ensures that downstream performance differences are more cleanly attributable to the pretrained encoder representations rather than to gains induced by model-specific classification layers. The encoder backbone is frozen, and only the classification head is trained, further standardizing the optimization setting across models and preserving a consistent initialization for subsequent test-time adaptation. Architectural details of the shared classifier head are provided in Appendix A.2. Model selection is performed exclusively on a held-out validation split, with no access to test data.

Stage 2: Test-Time Adaptation The held-out test split is treated as unlabeled target data. During adaptation, only EEG signals are provided to the model, and ground-truth labels are strictly withheld. As mentioned previously, we evaluate three representative TTA methods alongside a No-TTA baseline, selected to span two orthogonal axes of clinical deployment constraints: whether the full target set must be available before adaptation begins (offline vs. online), and whether the method requires gradient computation (Table 1). **No-TTA** performs inference with the frozen fine-tuned checkpoint and serves as the unadapted baseline. **Tent** updates the affine parameters of normalization layers (batch normalization, layer normalization, and group normalization) through entropy minimization on each incoming batch, accumulating state across the test stream without requiring a prior pass over the full target set. **T3A** maintains a per-class support set of low-entropy prototype features that is updated incrementally with each batch, and no gradient computation is required. **SHOT** first performs a full pass over the target set to construct refined feature centroids via mutual information maximization and pseudo-labeling, and then adapts the encoder through gradient descent. It therefore requires the complete target set to be available before adaptation begins.

Online methods (Tent, T3A) are suitable for streaming scenarios such as continuous bedside monitoring, whereas offline methods (SHOT) is better suited to settings in which a batch of recordings can be collected before deployment (e.g. sleep studies).

Stage 3: Evaluation After adaptation, ground-truth labels are used to compute standard classification metrics, including accuracy, balanced accuracy, ROC-AUC, PR-AUC, Cohen's κ, and weighted F_1. For each (method, model, dataset) combination, we report the mean and standard deviation across five random seeds. To isolate the effect of adaptation, we additionally report the relative improvement:

$$\Delta_{\text{TTA}} = \text{metric}_{\text{TTA}} - \text{metric}_{\text{No-TTA}}, \tag{9}$$

computed per seed prior to aggregation. This ensures that the reported variability reflects differences in adaptation performance rather than absolute model accuracy. Appendix A.6 reports the same comparisons as relative changes, and Appendix A.7 reports statistical significance tests against No-TTA. Appendix A.8 reports measured adaptation runtime and peak GPU memory.

3.3. Experiment Setup

We evaluate TTA for EEG foundation models using a standardized downstream pipeline that isolates the effect of deployment-time adaptation from model-specific classifier design. The benchmark spans four foundation-model variants, five EEG datasets, both binary and multiclass tasks, and patient-disjoint evaluation.

Datasets, Tasks and Metrics. We evaluate on five EEG datasets: TUEV (Harati et al., 2015), TUAB (Lopez et al., 2015), CHB-MIT (Shoeb and Guttag, 2010) from PhysioNet (Goldberger et al., 2000), EarEEG (Bjarke Mikkelsen et al., 2025; Tabar et al., 2024), and SleepEDF-78 (Eldele, 2022). TUEV and TUAB are treated as in-distribution datasets, as they are included in the pretraining corpora of the evaluated EEG foundation models and correspond to event classification and abnormality detection tasks. In contrast, CHB-MIT and SleepEDF-78 are considered out-of-distribution because they are not part of the pretraining data for most models, except for TFM-Tokenizer, which includes CHB-MIT during pretraining. CHB-MIT focuses on epilepsy seizure detection, and SleepEDF-78 focuses on sleep staging. To further study robustness to extreme distributional shift, we include an ear-EEG setting that differs substantially in signal modality, channel configuration, and acquisition setup from those seen during pretraining for all models. This setting focuses on sleep staging using EarEEG data and represents a challenging out-of-domain evaluation scenario.

We analyze these settings as subject-level, cross-dataset/task, and cross-modality/device shifts rather than as one uniform distribution shift. All evaluations use patient-disjoint splits to avoid subject leakage. To illustrate the evaluated distribution shifts, Appendix A.9 provides a qualitative t-SNE diagnostic of source-target structure in raw EEG and frozen foundation-model embeddings before TTA.

We report balanced accuracy, ROC-AUC, and PR-AUC for binary classification tasks, and balanced accuracy, Cohen's κ, and weighted F_1 for multi-class classification.

EEG Foundation Models. We evaluate three EEG foundation-model families, including CBraMod (Wang et al., 2024b), TFM-Tokenizer (Pradeepkumar et al., 2026b), and REVE (Ouahidi et al., 2025). For REVE, we consider both the Base and Large variants, as it is pretrained on one of the largest EEG corpora to date. CBraMod provides an efficient architecture with demonstrated generalization across multiple EEG tasks. In contrast to these continuous embedding-based models, TFM-Tokenizer introduces a discrete tokenization framework for EEG, enabling evaluation across fundamentally different representation paradigms. All models are integrated through a shared interface while preserving their native input processing and representational assumptions. We attach a common lightweight classification head to the publicly available and fixed pretrained backbones and fine-tune one classifier per (model, dataset) pair. Test-time adaptation is then applied during inference.

Standard Downstream Classifier. For fair comparison, each pretrained backbone is paired with the same lightweight downstream classifier. The encoder produces a latent representation, which is pooled when needed, passed through a shared feature adapter, and mapped to task logits by a linear classification layer. This removes backbone-specific downstream engineering as a major confound.

4. Results and Discussion

4.1. Does test-time adaptation improve performance for EEG foundation models?

Figure 3 shows the relative performance of TTA methods compared to the No-TTA baseline on TUEV and TUAB, which are in-distribution datasets included in the pretraining

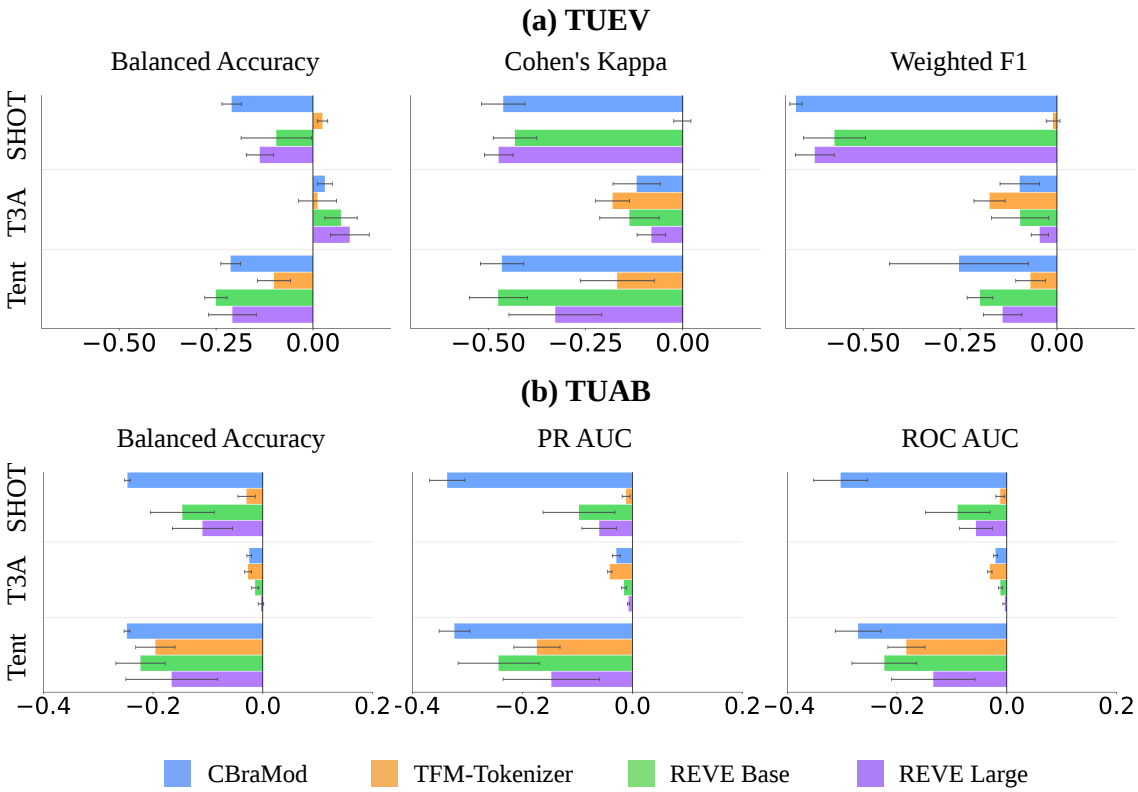

Figure 3: TTA relative performance on in-distribution datasets (TUEV and TUAB). (a) Δ_{TTA} on TUEV relative to the No-TTA baseline; (b) Δ_{TTA} on TUAB relative to the No-TTA baseline.

corpora of the EEG foundation models. In this setting, the primary source of variability is subject-level differences between the training and test splits. Across models and datasets, the evaluated gradient-based methods (Tent and SHOT) consistently degrade performance, often substantially. In contrast, T3A exhibits the most stable behavior and provides modest improvements in balanced accuracy on TUEV, with lower variability across seeds and batch sizes. On TUAB, however, all evaluated TTA methods degrade performance, with Tent showing the largest drop.

These results show limited benefit from the evaluated TTA methods on the in-distribution datasets. We avoid attributing this pattern to a specific mechanism, since we do not directly measure representation alignment or adaptation-induced drift in this experiment. The relative stability of T3A is therefore treated as an empirical observation rather than evidence that its optimization-free design is the causal factor.

4.2. How does TTA behave under cross-dataset and task shifts?

To evaluate TTA under realistic distribution shifts, we consider out-of-distribution datasets that are not included in the pretraining corpora of most evaluated models. These settings introduce compounded shifts that vary simultaneously across cohort, acquisition protocol,

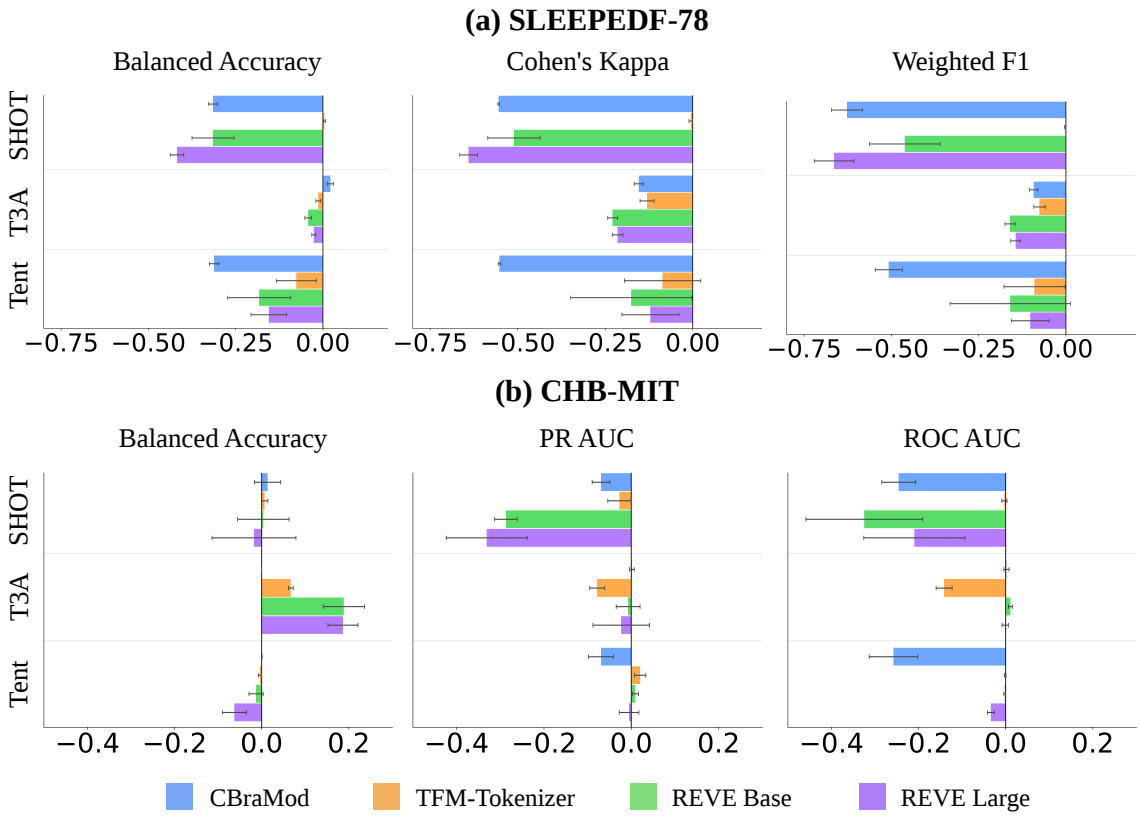

Figure 4: TTA relative performance on out-of-distribution datasets (SLEEPEDF-78 and CHB-MIT). (a) Δ_{TTA} on SLEEPEDF-78 relative to the No-TTA baseline; (b) Δ_{TTA} on CHB-MIT relative to the No-TTA baseline.

hardware, and sensor configuration. We therefore characterize robustness under these combined shifts rather than attempting to isolate a single dominant cause, which would require controlled, metadata-specific analyses beyond the scope of this benchmark.

Figure 4 summarizes the relative performance of TTA methods compared to the No-TTA baseline on SleepEDF-78 and CHB-MIT. On CHB-MIT, T3A provides consistent improvements in balanced accuracy across most models, showing trends similar to those observed in the in-distribution setting. We do not test the mechanism underlying this pattern directly. We therefore treat the T3A result as an empirical trend rather than evidence of a specific causal mechanism.

The REVE family, in particular, benefits from T3A on balanced accuracy and shows less to no degradation in ROC-AUC and PR-AUC. However, REVE shows higher degradation in PR-AUC and ROC AUC under SHOT.

In contrast, TTA on the more challenging SleepEDF-78 dataset shows greater degradation across nearly all methods and metrics. This dataset differs substantially in task (sleep staging) and channel configuration, making adaptation particularly difficult. T3A offers only marginal gains (e.g., slight improvements in balanced accuracy for CBraMod),

Table 2: TTA performance on the Ear-EEG sleep staging task (EESM23). We report Δ_{TTA} relative to the No-TTA baseline for each foundation model, aggregated across random seeds and adaptation batch sizes. Values are shown as mean ± standard deviation.

TTA Method	Foundation Model	Balanced Acc.Δ	Cohen's KappaΔ	Weighted F1Δ
SHOT	CBraMod	-0.065 ± 0.020	-0.087 ± 0.027	-0.224 ± 0.015
	TFM-Tokenizer	$+0.000 \pm 0.000$	-0.000 ± 0.001	-0.000 ± 0.000
	REVE-Base	-0.018 ± 0.032	-0.085 ± 0.064	-0.119 ± 0.076
	REVE-Large	-0.068 ± 0.056	-0.156 ± 0.083	-0.150 ± 0.087
T3A	CBraMod	$\mathbf{+0.048} \pm 0.012$	$\mathbf{+0.064} \pm 0.016$	$\mathbf{+0.018} \pm 0.009$
	TFM-Tokenizer	-0.005 ± 0.007	-0.042 ± 0.006	-0.009 ± 0.006
	REVE-Base	$+0.037 \pm 0.007$	$+0.001 \pm 0.015$	$+0.001 \pm 0.017$
	REVE-Large	$+0.022 \pm 0.007$	-0.010 ± 0.014	-0.009 ± 0.010
Tent	CBraMod	-0.064 ± 0.022	-0.084 ± 0.030	-0.189 ± 0.060
	TFM-Tokenizer	-0.001 ± 0.001	$+0.000 \pm 0.001$	-0.000 ± 0.001
	REVE-Base	-0.032 ± 0.022	-0.047 ± 0.035	-0.046 ± 0.028
	REVE-Large	-0.018 ± 0.014	-0.025 ± 0.018	-0.019 ± 0.015

while Tent and SHOT consistently degrade performance. Notably, TFM-Tokenizer shows relatively greater robustness across all TTA approaches, with smaller performance drops compared to other models. Overall, these results demonstrate that the evaluated TTA methods struggle to generalize under cross-dataset shifts.

4.3. Can TTA handle unseen EEG modalities such as EarEEG?

In this section, we evaluate TTA under extreme distribution shift by considering an unseen EEG modality, such as ear-EEG. Most EEG foundation models are pretrained on scalp EEG data following the standard 10-20 system, whereas ear-EEG differs substantially in signal characteristics, channel configuration, and acquisition setup. With the growing adoption of wearable EEG technologies (Bjarke Mikkelsen et al., 2025), understanding cross-modality generalization from scalp-EEG to wearable EEG has become increasingly important (Anandakumar et al., 2023).

In this setting, we assess whether the evaluated TTA methods can adapt pretrained scalp-EEG foundation models to ear-EEG sleep staging and the relative improvement results are summarized in Table 2. Overall, the evaluated TTA methods are unstable under this modality shift. The evaluated gradient-based approaches (SHOT and Tent) consistently degrade performance across models and metrics. In contrast, T3A is more stable and yields improvements for some models, notably for CBraMod across all metrics, with moderate gains for REVE in balanced accuracy.

4.4. How does adaptation batch size affect performance?

Figure 5 shows the effect of adaptation batch size on balanced accuracy across all models and datasets. Overall, increasing batch size for the evaluated methods does not provide consistent performance gains. For the evaluated gradient-based methods (SHOT and Tent), increasing batch size can reduce degradation in some settings, but it does not provide consistent gains across datasets.

In contrast, T3A is relatively insensitive to batch size, as it updates class prototypes without gradient-based optimization. These results suggest that simply increasing batch size is insufficient to stabilize or improve the evaluated TTA methods for EEG.

4.5. Which TTA methods are most stable?

Across the evaluated TTA methods, T3A was by far the most stable, as when used with foundation models, it caused the least degradation relative to the No-TTA baseline and occasionally even yielded benefits. We hypothesize that, as an optimization-free approach, T3A avoids updating model parameters, which may help preserve pretrained representations under heterogeneous EEG deployment settings. However, more controlled experiments across other approaches would be necessary to validate this mechanism.

In contrast, the evaluated gradient-based methods (SHOT and Tent) frequently lead to substantial performance degradation, often outweighing their occasional benefits on specific model architectures. We hypothesize that their objectives may perturb pretrained representations, which would be consistent with the negative transfer observed under distribution shift, although we similarly do not verify this mechanism directly. Validating the dominant mechanism is left to future work.

Overall, our ablation suggests that increasing batch size generally helps reduce degradation for the evaluated TTA methods. However, the performance benefits do not seem large enough to justify increasing memory or computational requirements without a significant gain.

How do online and offline adaptation strategies compare? In terms of the online versus offline characteristics outlined in Table 1, we find that, for the evaluated TTA methods, the distinction between online and offline adaptation plays a less significant role

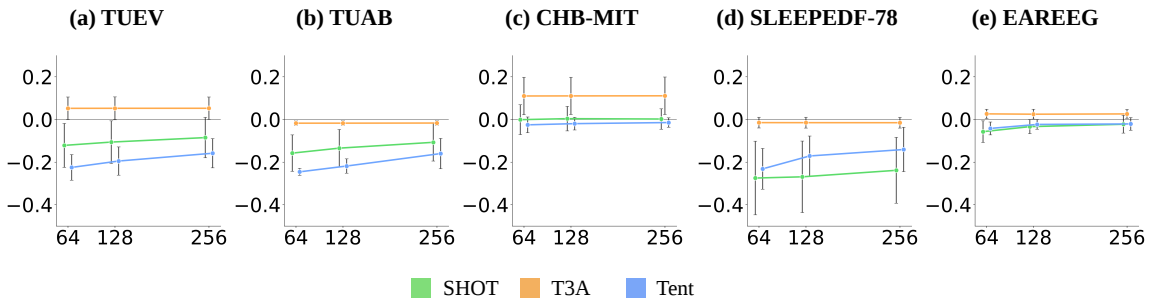

Figure 5: Balanced accuracy improvements relative to the No-TTA baseline across adaptation batch sizes (64, 128, 256). Results are shown for (a) TUEV, (b) TUAB, (c) CHB-MIT, (d) SleepEDF-78, and (e) EarEEG.

than the nature of the update mechanism. While both Tent and T3A operate in an online setting, their behaviors differ. Tent updates internal normalization parameters, often introducing instability, whereas T3A modifies class prototypes without updating model weights, resulting in more stable performance and occasional improvements. These results suggest that the where and how of adaptation are more critical for these methods than whether it is performed in a streaming or batch setting.

Does representation type affect TTA behavior? Overall, our experiments on the evaluated models and methods suggest responses to adaptation vary between continuous-embedding and discrete-tokenization models. However, we present this as a hypothesis suggested by the empirical trends, and leave a controlled test to future work.

For instance, TFM-Tokenizer shows more resistance to degradation for the evaluated TTA methods, particularly under the SHOT method, across all datasets in both in-distribution and out-of-distribution settings. Other continuous embedding-based models, such as REVE, seem to benefit the most from the T3A method, especially on the CHB-MIT dataset.

4.6. Discussion and Implications for Future Works

Our results highlight several important considerations for deploying TTA with EEG foundation models. First, the evaluated foundation models are not plug-and-play in real-world clinical settings, while they perform well on in-distribution datasets (e.g., TUAB, TUEV), performance degrades substantially under out-of-distribution conditions, particularly under extreme shifts such as EarEEG. This underscores the distribution shift as a primary barrier to reliable deployment. Second, we observe that the evaluated optimization-free TTA methods exhibit greater stability than the evaluated gradient-based approaches, which exhibit performance degradation. This suggests that future work should prioritize robustness and explore alternative adaptation strategies that minimize disruptive updates.

4.7. Limitations

Our benchmark covers three representative test-time adaptation methods across batch and streaming settings, four EEG foundation model variants, and five downstream datasets. However, it does not span the full space of TTA approaches and model families, so our conclusions are scoped to the evaluated EEG foundation models and TTA categories. We benchmark generic TTA methods rather than physiology-aware adaptation methods designed for EEG-specific nonstationarity, subject variability, spectral structure, and low signal-to-noise ratio.

Our work provides a reproducible evaluation framework to test and study the effectiveness of TTA methods with EEG foundation models. We do not evaluate calibration or uncertainty, so reliability diagrams, expected calibration error, and uncertainty-aware evaluation remain important before adapted models can be considered clinically trustworthy.

Although our settings include patient-disjoint, cross-dataset, cross-task/institutional, and cross-modality shifts, stronger real-world shifts such as temporal drift and center-specific acquisition protocols remain future stress tests.

We also restrict this study to EEG foundation models, and do not include analysis of non-foundation or general-purpose time-series foundation model baselines which require

broader experimental controls. Such broader analyses could provide the basis for future work, but are not considered in this benchmark.

Finally, computational cost is another limitation. Larger models, such as REVE-Large, can make adaptation memory-intensive, limiting the feasible batch sizes and hardware accessibility. Evaluating computational efficiency and the feasibility of clinical deployment more directly is an important direction for future work.

5. Conclusion

In this work, we present *NeuroAdapt-Bench*, a systematic benchmark for evaluating test-time adaptation methods with EEG foundation models under realistic distribution shifts. Across diverse datasets and tasks, we find that the evaluated TTA methods yield inconsistent gains and often degrade performance relative to the No-TTA baseline. In particular, the evaluated optimization-free approaches such as T3A demonstrate greater stability and often positive gains, whereas the evaluated gradient-based methods are more prone to degradation.

Our results highlight that distribution shift remains a fundamental challenge for deploying the evaluated EEG foundation models, and that the evaluated TTA methods from other domains do not transfer reliably in this benchmark, thereby motivating EEG-specific test-time adaptation approaches. We frame these results as diagnostic, not as a final rejection of TTA for EEG. Overall, our study establishes a reproducible evaluation framework and provides empirical insights to guide the development of reliable and robust adaptation strategies for EEG.

References

Mithunjha Anandakumar, Jathurshan Pradeepkumar, Simon L Kappel, Chamira US Edussooriya, and Anjula C De Silva. A knowledge distillation framework for enhancing ear-eeg based sleep staging with scalp-eeg data. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 514–519. IEEE, 2023.

Kaare Bjarke Mikkelsen, Yousef Rezai Tabar, Laura Rævsbæk Birch, Simon Lind Kappel, Christian Bech Christensen, Lars Dalskov Mosgaard, Marit Otto, Martin Christian Hemmsen, Mike Lind Rank, and Preben Kidmose. Ear-eeg sleep monitoring data sets. *Scientific Data*, 12(1):301, 2025.

Jiaheng Dong, Hong Jia, Soumyajit Chatterjee, Abhirup Ghosh, James Bailey, and Ting Dang. E-bats: Efficient backpropagation-free test-time adaptation for speech foundation models. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=WwzurufeFN.

Emadeldeen Eldele. Preprocessed Sleep-EDF-78 dataset, 2022. URL https://doi.org/10.21979/N9/EUHGHS.

Ary L Goldberger, Luis A N Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, C-K Peng, and H Eugene Stanley.

Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

Hanfei Guo, Hai Guo, Chang Li, Hu Peng, Zhihui Han, and Xun Chen. Test time adaptation for cross-domain sleep stage classification. *Biomedical Signal Processing and Control*, 100:106980, 2025. doi: 10.1016/j.bspc.2024.106980. URL https://doi.org/10.1016/j.bspc.2024.106980.

Amir Harati, Meysam Golmohammadi, Silvia Lopez, Iyad Obeid, and Joseph Picone. Improved eeg event classification using differential energy. In *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–4. IEEE, 2015.

Pei-Kai Huang, Tzu-Hsien Chen, Ya-Ting Chan, Kuan-Wen Chen, and Chiou-Ting Hsu. Fully test-time rppg estimation via synthetic signal-guided feature learning. *Pattern Recognition*, 170:112102, 2026. doi: 10.1016/j.patcog.2025.112102. URL https://doi.org/10.1016/j.patcog.2025.112102.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Geun-Deok Jang, Dong-Kyun Han, and Seong-Whan Lee. Test-time adaptation for eeg-based driver drowsiness classification. In *Pattern Recognition and Artificial Intelligence*, volume 14892 of *Lecture Notes in Computer Science*, pages 410–424. Springer, Singapore, 2025. doi: 10.1007/978-981-97-8702-9_28. URL https://doi.org/10.1007/978-981-97-8702-9_28.

Vinay Jayaram, Morteza Alamgir, Yasemin Altun, Bernhard Scholkopf, and Moritz Grosse-Wentrup. Transfer learning in brain-computer interfaces. *IEEE Computational Intelligence Magazine*, 11(1):20–31, 2016.

Haohui Jia, Zheng Chen, Lingwei Zhu, Rikuto Kotoge, Jathurshan Pradeepkumar, Yasuko Matsubara, Jimeng Sun, Yasushi Sakurai, and Takashi Matsubara. ODEBrain: Continuous-time EEG graph for modeling dynamic brain networks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=9EjZGs8ube.

Ziyu Jia, Xihao Yang, Chenyang Zhou, Haoyang Deng, and Tianzi Jiang. Atta: Adaptive test-time adaptation for multi-modal sleep stage classification. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 5882–5890, 2024. doi: 10.24963/ijcai.2024/650. URL https://www.ijcai.org/proceedings/2024/650.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*, 2024.

Yong-Yeon Jo, Byeong Tak Lee, Jeong-Ho Hong, Hak Seung Lee, Joon-myoung Kwon, and Beom Joon Kim. Test-time calibration: A framework for personalized test-time adaptation in real-world biosignals. In *Proceedings of the Sixth Conference on Health,*

Inference, and Learning, volume 287 of *Proceedings of Machine Learning Research*, pages 381–394. PMLR, 2025. URL https://proceedings.mlr.press/v287/jo25a.html.

Ard Kastrati, Josua Bürki, Jonas Lauer, Cheng Xuan, Raffaele Iaquinto, and Roger Wattenhofer. Eeg-bench: A benchmark for eeg foundation models in clinical applications. *arXiv preprint arXiv:2512.08959*, 2025.

Gayal Kuruppu, Neeraj Wagh, Vaclav Kremen, and Yogatheesan Varatharajah. Eeg foundation models: a critical review of current progress and future directions. *Journal of Neural Engineering*, 23(2), 2026. doi: 10.1088/1741-2552/ae4455. URL https://doi.org/10.1088/1741-2552/ae4455.

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 (ICML)*, pages 6028–6039, 2020.

Guan-Ting Lin, Wei Ping Huang, and Hung-yi Lee. Continual test-time adaptation for end-to-end speech recognition on noisy speech. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20003–20015, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1116. URL https://aclanthology.org/2024.emnlp-main.1116/.

Hongfu Liu, Hengguan Huang, and Ye Wang. Advancing test-time adaptation in wild acoustic test settings. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7138–7155, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.405. URL https://aclanthology.org/2024.emnlp-main.405/.

Sebas Lopez, G Suarez, D Jungreis, I Obeid, and Joseph Picone. Automated identification of abnormal adult eegs. In *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–5. IEEE, 2015.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16888–16905. PMLR, 2022. URL https://proceedings.mlr.press/v162/niu22a.html.

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. *arXiv preprint arXiv:2510.21585*, 2025.

Huy Phan, Kaare Mikkelsen, Oliver Y Chén, Philipp Koch, Alfred Mertins, and Maarten De Vos. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering*, 69(8):2456–2467, 2022.

Jathurshan Pradeepkumar, Mithunjha Anandakumar, Vinith Kugathasan, Dhinesh Suntharalingham, Simon L Kappel, Anjula C De Silva, and Chamira US Edussooriya. Toward

interpretable sleep stage classification using cross-modal transformers. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:2893–2904, 2024.

Jathurshan Pradeepkumar, Zheng Chen, and Jimeng Sun. Neural signals generate clinical notes in the wild. *arXiv preprint arXiv:2601.22197*, 2026a.

Jathurshan Pradeepkumar, Xihao Piao, Zheng Chen, and Jimeng Sun. Tokenizing single-channel EEG with time-frequency motif learning. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=2sPmWHZ8Ir.

Vandana Akshath Raj, Tejasvi Parupudi, Ananthakrishna Thalengala, and Subramanya G. Nayak. A comprehensive review of deep learning models for denoising eeg signals: challenges, advances, and future directions. *Discover Applied Sciences*, 7, 2025. doi: 10.1007/s42452-025-07808-2. URL https://doi.org/10.1007/s42452-025-07808-2.

Ali H. Shoeb and John V. Guttag. Application of machine learning to epileptic seizure detection. In *International Conference on Machine Learning*, 2010. URL https://api.semanticscholar.org/CorpusID:11141395.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In Hal Daume III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9229–9248. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/sun20b.html.

M Sundaram, RM Sadler, GB Young, and N Pillay. Eeg in epilepsy: current perspectives. *Canadian Journal of Neurological Sciences*, 26(4):255–262, 1999.

Yousef Rezaei Tabar, Kaare Mikkelsen, Laura Birch, Nelly Shenton, Simon L Kappel, Astrid R Bertelsen, Reza Nikbakht, Hans O Toft, Chris H Henriksen, Martin C Hemmsen, Mike L Rank, Marit Otto, and Preben Kidmose. "ear-eeg sleep monitoring 2023 (eesm23)", 2024.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

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

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. *arXiv preprint arXiv:2412.07236*, 2024b.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBramod: A criss-cross brain foundation model for EEG decoding.

In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=NPNUHgHF2w.

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*, pages 7201–7211, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Continual_Test-Time_Domain_Adaptation_CVPR_2022_paper.html.

Yanshuo Wang, Yanghao Zhou, Yukang Lin, Haoxing Chen, Jin Zhang, Wentao Zhu, Jie Hong, and Xuesong Li. Dynamic model-bank test-time adaptation for automatic speech recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21831–21841, Suzhou, China, 2025b. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1107. URL https://aclanthology.org/2025.emnlp-main.1107/.

Zixin Wang, Yadan Luo, Liang Zheng, Zhuoxiao Chen, Sen Wang, and Zi Huang. In search of lost online test-time adaptation: A survey. *International Journal of Computer Vision*, 133:1106–1139, 2025c. doi: 10.1007/s11263-024-02213-5. URL https://doi.org/10.1007/s11263-024-02213-5.

Xinyi Wu, Zouheir Rezki, Bhavan Balusu, Jiaming Zhou, and Jun Bai. One model, dual tasks: a novel distributionally adaptive learning framework for ecg classification and generation addressing intra- and inter-patient variability. *Expert Systems with Applications*, 299:130096, 2026. doi: 10.1016/j.eswa.2025.130096. URL https://doi.org/10.1016/j.eswa.2025.130096.

Zongzhe Xu, Zitao Shuai, Eideen Mozaffari, Ravi S Aysola, Rajesh Kumar, and Yuzhe Yang. Sleeplm: Natural-language intelligence for human sleep. *arXiv preprint arXiv:2602.23605*, 2026.

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

Chaoqi Yang, M Brandon Westover, and Jimeng Sun. Manydg: Many-domain generalization for healthcare applications. *arXiv preprint arXiv:2301.08834*, 2023b.

Yuxuan Yao, Hongbo Wang, Li Chen, Yiheng Peng, and Jingjing Luo. Foundation models for eeg decoding: current progress and prospective research. *Journal of Neural Engineering*, 22(6), 2025. doi: 10.1088/1741-2552/ae17e9. URL https://doi.org/10.1088/1741-2552/ae17e9.

Appendix A.

A.1. Dataset Preprocessing and Split Policy

Table 3 summarizes the dataset-specific preprocessing and split policy used in the benchmark. Across all datasets, preprocessing standardizes temporal support and amplitude scaling while preserving dataset-specific channel geometry rather than forcing all recordings into a single global montage. For each foundation model, we reproduce the preprocessing expected by the original model implementation, including its resampling, windowing, normalization, and channel-handling conventions.

Table 3: Dataset preprocessing and split policy used in the benchmark.

Dataset	Task	Classes	Channels	Window	Rate	Normalization	Split policy
TUEV	Multiclass	6	16	5 s	200 Hz	Per-channel 95th percentile	Official train/eval pools; seeded patient-level 80/20 split of the train pool into train/val
TUAB	Binary	2	16	10 s	200 Hz	Per-channel 95th percentile	Official train/eval pools; seeded patient-level 80/20 split of the train pool into train/val
CHB-MIT	Binary	2	16	10 s	256 Hz	Per-channel 95th percentile	Precomputed train/val/test split
SleepEDF-78	Multiclass	5	2	30 s	100 Hz	Per-channel 95th percentile	Precomputed train/val/test split
EarEEG	Multiclass	6	4	30 s	250 Hz	Per-channel 95th percentile	Precomputed train/val/test split; final stored channel dropped before preprocessing

Table 4 summarizes the benchmark datasets, tasks, and shift categories. We separate these categories to avoid treating subject-level variation, cross-dataset/task transfer, and device/modality transfer as one uniform shift.

Table 4: Dataset-task summary and shift category.

Dataset	Task	Classes	Shift category
TUEV	Event classification	6	Subject-level (in-distribution)
TUAB	Abnormality detection	2	Subject-level (in-distribution)
CHB-MIT	Seizure detection	2	Cross-dataset/task (out-of-distribution)
SleepEDF-78	Sleep staging	5	Cross-dataset/task (out-of-distribution)
EarEEG	Sleep staging	6	Cross-modality/device (extreme modality shift)

A.2. Shared Downstream Classifier and Fine-Tuning

To compare pretrained backbones under a common downstream protocol, each foundation model is paired with the same lightweight classifier head. The shared head first applies LayerNorm to the pooled feature representation, then projects features to a 128-dimensional hidden space, applies GELU and dropout, and finally maps to task logits with a linear classification layer. Encoder backbones are frozen during the reported downstream fine-tuning experiments. REVE is the only backbone that additionally consumes channel-position information, while the other backbones operate only on the waveform input. For REVE, channel positions are constructed from a shared 3D electrode position bank; for bipolar channels, the benchmark uses the mean of the two endpoint electrode coordinates. For

EarEEG, the four channels are mapped to the aliases A2, T8, A1, and T7 before position lookup. CBraMod additionally applies a fixed input scaling factor of 100 to match the expected amplitude range of the pretrained backbone.

Table 5: Shared downstream classifier and fine-tuning configuration.

Component	**Setting**
Foundation-model variants	CBraMod, TFM, REVE-Base, REVE-Large
Encoder training during fine-tuning	Frozen
Shared downstream head	LayerNorm $\rightarrow$ Linear(128) $\rightarrow$ GELU $\rightarrow$ Dropout(0.1) $\rightarrow$ Linear(C)
Pooling policy	Mean pooling for sequence outputs; native pooled outputs used when provided by the backbone
Loss	Cross-entropy objective
Optimizer	AdamW
Learning rate	10^{-3}
Weight decay	10^{-4}
Epochs	10
Classifier training batch size	512
Data augmentation	None
Model selection for binary tasks	Best validation ROC-AUC
Model selection for multi-class tasks	Best validation Cohen's κ
Reported study seeds	Five random seeds used in the reported study

A.3. Test-Time Adaptation Configuration

Table 6 summarizes the operational configuration of the evaluated test-time adaptation methods. We report the specific settings used in the benchmark rather than the full space of possible variants for each method family. We do not tune TTA hyperparameters per dataset because target labels are withheld and per-dataset tuning would leak target information. Instead, we use fixed settings based on the original method defaults and vary only adaptation batch size systematically in the batch-size ablation.

Table 6: Test-time adaptation configuration used in the benchmark.

Method	**Regime**	**Updated component**	**Optimizer**	**Key settings**
No-TTA	None	None	None	Frozen checkpoint inference
Tent	Online	Affine parameters of normalization layers	SGD	lr = 10^{-3}, momentum = 0.9, steps = 1, episodic = False
SHOT	Offline	Trainable feature modules only; classifier fixed	SGD	lr = 10^{-4}, wd = 10^{-4}, steps = 1, episodic = False, MI weight = 1.0, PL weight = 1.0
T3A	Online	Classifier supports / prototypes	None	filter_k = 20, episodic = False

A.4. Evaluation and Aggregation

Table 7 summarizes the evaluation settings used throughout the benchmark.

Table 7: Evaluation and aggregation settings.

Setting	**Value**
Primary binary metrics	Balanced accuracy, ROC-AUC, PR-AUC
Primary multiclass metrics	Balanced accuracy, Cohen's κ, weighted F_1
Additional logged metric	Accuracy
Aggregation	Mean $\pm$ standard deviation over study seeds
Relative adaptation metric	$\Delta_{\text{TTA}} = \text{metric}_{\text{TTA}} - \text{metric}_{\text{No-TTA}}$, computed per seed before averaging
TTA evaluation batch sizes	64, 128, 256

A.5. Per-Dataset Delta Tables

Our quantitative results are summarized in the following tables. Each table reports performance deltas relative to the No-TTA baseline for a single dataset, separated by foundation model and aggregated across seeds and adaptation batch sizes. Values are mean $\pm$ standard deviation, and bold values indicate the largest mean for each metric in each table.

Table 8: TUEV delta relative to the no-TTA baseline.

TTA Method	**Foundation Model**	**Bal. Acc. Δ**	**Cohen's κ Δ**	**Weighted F1 Δ**
SHOT	CBraMod	-0.210 ± 0.025	-0.462 ± 0.056	-0.673 ± 0.016
	TFM	$+0.024 \pm 0.014$	$\mathbf{-0.001} \pm 0.022$	$\mathbf{-0.010} \pm 0.018$
	REVE-Base	-0.095 ± 0.090	-0.431 ± 0.056	-0.574 ± 0.080
	REVE-Large	-0.137 ± 0.035	-0.473 ± 0.037	-0.624 ± 0.050
T3A	CBraMod	$+0.031 \pm 0.019$	-0.118 ± 0.061	-0.096 ± 0.051
	TFM	$+0.012 \pm 0.049$	-0.181 ± 0.044	-0.174 ± 0.040
	REVE-Base	$+0.072 \pm 0.042$	-0.137 ± 0.076	-0.095 ± 0.074
	REVE-Large	$\mathbf{+0.095} \pm 0.050$	-0.081 ± 0.037	-0.044 ± 0.022
Tent	CBraMod	-0.212 ± 0.026	-0.465 ± 0.056	-0.253 ± 0.179
	TFM	-0.101 ± 0.043	-0.169 ± 0.095	-0.068 ± 0.038
	REVE-Base	-0.250 ± 0.029	-0.474 ± 0.075	-0.198 ± 0.033
	REVE-Large	-0.208 ± 0.061	-0.328 ± 0.119	-0.140 ± 0.050

Table 9: TUAB delta relative to the no-TTA baseline.

TTA Method	Foundation Model	Bal. Acc. Δ	ROC AUC Δ	PR AUC Δ
SHOT	CBraMod	-0.247 ± 0.005	-0.303 ± 0.049	-0.337 ± 0.032
	TFM	-0.030 ± 0.016	-0.012 ± 0.008	-0.011 ± 0.007
	REVE-Base	-0.147 ± 0.058	-0.090 ± 0.059	-0.097 ± 0.065
	REVE-Large	-0.110 ± 0.055	-0.056 ± 0.030	-0.060 ± 0.032
T3A	CBraMod	-0.025 ± 0.004	-0.021 ± 0.004	-0.029 ± 0.007
	TFM	-0.027 ± 0.006	-0.031 ± 0.004	-0.041 ± 0.004
	REVE-Base	-0.014 ± 0.006	-0.012 ± 0.004	-0.016 ± 0.004
	REVE-Large	$\mathbf{-0.003} \pm 0.005$	$\mathbf{-0.004} \pm 0.003$	$\mathbf{-0.007} \pm 0.002$
Tent	CBraMod	-0.248 ± 0.005	-0.271 ± 0.042	-0.324 ± 0.028
	TFM	-0.196 ± 0.036	-0.183 ± 0.034	-0.174 ± 0.042
	REVE-Base	-0.223 ± 0.045	-0.223 ± 0.059	-0.243 ± 0.074
	REVE-Large	-0.166 ± 0.084	-0.134 ± 0.076	-0.147 ± 0.088

Table 10: CHB-MIT delta relative to the no-TTA baseline.

TTA Method	Foundation Model	Bal. Acc. Δ	ROC AUC Δ	PR AUC Δ
SHOT	CBraMod	$+0.014 \pm 0.030$	-0.246 ± 0.039	-0.069 ± 0.020
	TFM	$+0.007 \pm 0.007$	-0.003 ± 0.006	-0.027 ± 0.026
	REVE-Base	$+0.004 \pm 0.059$	-0.325 ± 0.134	-0.287 ± 0.026
	REVE-Large	-0.018 ± 0.096	-0.210 ± 0.116	-0.331 ± 0.093
T3A	CBraMod	$+0.000 \pm 0.000$	$+0.002 \pm 0.006$	$+0.001 \pm 0.005$
	TFM	$+0.067 \pm 0.006$	-0.141 ± 0.019	-0.078 ± 0.017
	REVE-Base	$\mathbf{+0.189} \pm 0.048$	$\mathbf{+0.011} \pm 0.005$	-0.007 ± 0.027
	REVE-Large	$+0.187 \pm 0.035$	-0.001 ± 0.008	-0.023 ± 0.065
Tent	CBraMod	$+0.000 \pm 0.001$	-0.257 ± 0.056	-0.070 ± 0.029
	TFM	-0.004 ± 0.004	-0.001 ± 0.002	$\mathbf{+0.020} \pm 0.013$
	REVE-Base	-0.013 ± 0.016	-0.002 ± 0.002	$+0.009 \pm 0.007$
	REVE-Large	-0.063 ± 0.028	-0.034 ± 0.008	-0.005 ± 0.022

Table 11: Sleep-EDF delta relative to the no-TTA baseline.

TTA Method	Foundation Model	Bal. Acc. Δ	Cohen's κ Δ	Weighted F1 Δ
SHOT	CBraMod	-0.315 ± 0.013	-0.554 ± 0.003	-0.627 ± 0.044
	TFM	$+0.004 \pm 0.003$	$\mathbf{-0.006} \pm 0.005$	$\mathbf{-0.002} \pm 0.002$
	REVE-Base	-0.315 ± 0.060	-0.511 ± 0.075	-0.461 ± 0.101
	REVE-Large	-0.418 ± 0.019	-0.640 ± 0.025	-0.664 ± 0.057
T3A	CBraMod	$\mathbf{+0.022} \pm 0.009$	-0.154 ± 0.013	-0.093 ± 0.012
	TFM	-0.014 ± 0.007	-0.130 ± 0.020	-0.076 ± 0.017
	REVE-Base	-0.042 ± 0.010	-0.229 ± 0.014	-0.160 ± 0.014
	REVE-Large	-0.026 ± 0.006	-0.215 ± 0.015	-0.144 ± 0.014
Tent	CBraMod	-0.312 ± 0.014	-0.552 ± 0.004	-0.508 ± 0.039
	TFM	-0.076 ± 0.058	-0.086 ± 0.109	-0.090 ± 0.087
	REVE-Base	-0.183 ± 0.090	-0.176 ± 0.173	-0.160 ± 0.172
	REVE-Large	-0.155 ± 0.051	-0.121 ± 0.081	-0.102 ± 0.054

Table 12: EAR-EEG delta relative to the no-TTA baseline.

TTA Method	Foundation Model	Bal. Acc. Δ	Cohen's κ Δ	Weighted F1 Δ
SHOT	CBraMod	-0.065 ± 0.020	-0.087 ± 0.027	-0.224 ± 0.015
	TFM	$+0.000 \pm 0.000$	-0.000 ± 0.001	-0.000 ± 0.000
	REVE-Base	-0.018 ± 0.032	-0.085 ± 0.064	-0.119 ± 0.076
	REVE-Large	-0.068 ± 0.056	-0.156 ± 0.083	-0.150 ± 0.087
T3A	CBraMod	$\mathbf{+0.048} \pm 0.012$	$\mathbf{+0.064} \pm 0.016$	$\mathbf{+0.018} \pm 0.009$
	TFM	-0.005 ± 0.007	-0.042 ± 0.006	-0.009 ± 0.006
	REVE-Base	$+0.037 \pm 0.007$	$+0.001 \pm 0.015$	$+0.001 \pm 0.017$
	REVE-Large	$+0.022 \pm 0.007$	-0.010 ± 0.014	-0.009 ± 0.010
Tent	CBraMod	-0.064 ± 0.022	-0.084 ± 0.030	-0.189 ± 0.060
	TFM	-0.001 ± 0.001	$+0.000 \pm 0.001$	-0.000 ± 0.001
	REVE-Base	-0.032 ± 0.022	-0.047 ± 0.035	-0.046 ± 0.028
	REVE-Large	-0.018 ± 0.014	-0.025 ± 0.018	-0.019 ± 0.015

A.6. Relative Performance Change

We report TTA-minus-No-TTA deltas together with the corresponding relative change. Relative change is computed as $100 \times$ mean Δ/mean No-TTA. Values use the same seed and batch-size aggregation as Appendix A.5.

Table 13: TUEV relative performance change from No-TTA. Entries show mean delta ± standard deviation, with relative change in parentheses.

TTA Method	Foundation Model	Bal. Acc. Δ (%)	Cohen's κ Δ (%)	Weighted F1 Δ (%)
SHOT	CBraMod	-0.210 ± 0.025 (-55%)	-0.462 ± 0.056 (-99%)	-0.673 ± 0.016 (-93%)
	TFM	$+0.024 \pm 0.014$ ($+7\%$)	-0.001 ± 0.022 (-0%)	-0.010 ± 0.018 (-1%)
	REVE-Base	-0.095 ± 0.090 (-21%)	-0.431 ± 0.056 (-77%)	-0.574 ± 0.080 (-74%)
	REVE-Large	-0.137 ± 0.035 (-28%)	-0.473 ± 0.037 (-81%)	-0.624 ± 0.050 (-79%)
T3A	CBraMod	$+0.031 \pm 0.019$ ($+8\%$)	-0.118 ± 0.061 (-26%)	-0.096 ± 0.051 (-13%)
	TFM	$+0.012 \pm 0.049$ ($+3\%$)	-0.181 ± 0.044 (-45%)	-0.174 ± 0.040 (-25%)
	REVE-Base	$+0.072 \pm 0.042$ ($+16\%$)	-0.137 ± 0.076 (-25%)	-0.095 ± 0.074 (-12%)
	REVE-Large	$+0.095 \pm 0.050$ ($+19\%$)	-0.081 ± 0.037 (-14%)	-0.044 ± 0.022 (-6%)
Tent	CBraMod	-0.212 ± 0.026 (-56%)	-0.465 ± 0.056 (-100%)	-0.253 ± 0.179 (-35%)
	TFM	-0.101 ± 0.043 (-27%)	-0.169 ± 0.095 (-42%)	-0.068 ± 0.038 (-10%)
	REVE-Base	-0.250 ± 0.029 (-55%)	-0.474 ± 0.075 (-85%)	-0.198 ± 0.033 (-26%)
	REVE-Large	-0.208 ± 0.061 (-42%)	-0.328 ± 0.119 (-56%)	-0.140 ± 0.050 (-18%)

Table 14: TUAB relative performance change from No-TTA. Entries show mean delta ± standard deviation, with relative change in parentheses.

TTA Method	Foundation Model	Bal. Acc. Δ (%)	ROC AUC Δ (%)	PR AUC Δ (%)
SHOT	CBraMod	-0.247 ± 0.005 (-33%)	-0.303 ± 0.049 (-37%)	-0.337 ± 0.032 (-41%)
	TFM	-0.030 ± 0.016 (-4%)	-0.012 ± 0.008 (-1%)	-0.011 ± 0.007 (-1%)
	REVE-Base	-0.147 ± 0.058 (-18%)	-0.090 ± 0.059 (-10%)	-0.097 ± 0.065 (-11%)
	REVE-Large	-0.110 ± 0.055 (-14%)	-0.056 ± 0.030 (-6%)	-0.060 ± 0.032 (-7%)
T3A	CBraMod	-0.025 ± 0.004 (-3%)	-0.021 ± 0.004 (-3%)	-0.029 ± 0.007 (-4%)
	TFM	-0.027 ± 0.006 (-4%)	-0.031 ± 0.004 (-4%)	-0.041 ± 0.004 (-5%)
	REVE-Base	-0.014 ± 0.006 (-2%)	-0.012 ± 0.004 (-1%)	-0.016 ± 0.004 (-2%)
	REVE-Large	-0.003 ± 0.005 (-0%)	-0.004 ± 0.003 (-0%)	-0.007 ± 0.002 (-1%)
Tent	CBraMod	-0.248 ± 0.005 (-33%)	-0.271 ± 0.042 (-33%)	-0.324 ± 0.028 (-39%)
	TFM	-0.196 ± 0.036 (-26%)	-0.183 ± 0.034 (-22%)	-0.174 ± 0.042 (-21%)
	REVE-Base	-0.223 ± 0.045 (-28%)	-0.223 ± 0.059 (-25%)	-0.243 ± 0.074 (-27%)
	REVE-Large	-0.166 ± 0.084 (-21%)	-0.134 ± 0.076 (-15%)	-0.147 ± 0.088 (-16%)

Table 15: CHB-MIT relative performance change from No-TTA. Entries show mean delta ± standard deviation, with relative change in parentheses.

TTA Method	Foundation Model	Bal. Acc. Δ (%)	ROC AUC Δ (%)	PR AUC Δ (%)
SHOT	CBraMod	$+0.014 \pm 0.030$ (+3%)	-0.246 ± 0.039 (−33%)	-0.069 ± 0.020 (−75%)
	TFM	$+0.007 \pm 0.007$ (+1%)	-0.003 ± 0.006 (−0%)	-0.027 ± 0.026 (−8%)
	REVE-Base	$+0.004 \pm 0.059$ (+1%)	-0.325 ± 0.134 (−39%)	-0.287 ± 0.026 (−90%)
	REVE-Large	-0.018 ± 0.096 (−3%)	-0.210 ± 0.116 (−24%)	-0.331 ± 0.093 (−82%)
T3A	CBraMod	$+0.000 \pm 0.000$ (+0%)	$+0.002 \pm 0.006$ (+0%)	$+0.001 \pm 0.005$ (+1%)
	TFM	$+0.067 \pm 0.006$ (+12%)	-0.141 ± 0.019 (−16%)	-0.078 ± 0.017 (−24%)
	REVE-Base	$+0.189 \pm 0.048$ (+34%)	$+0.011 \pm 0.005$ (+1%)	-0.007 ± 0.027 (−2%)
	REVE-Large	$+0.187 \pm 0.035$ (+31%)	-0.001 ± 0.008 (−0%)	-0.023 ± 0.065 (−6%)
Tent	CBraMod	$+0.000 \pm 0.001$ (+0%)	-0.257 ± 0.056 (−34%)	-0.070 ± 0.029 (−75%)
	TFM	-0.004 ± 0.004 (−1%)	-0.001 ± 0.002 (−0%)	$+0.020 \pm 0.013$ (+6%)
	REVE-Base	-0.013 ± 0.016 (−2%)	-0.002 ± 0.002 (−0%)	$+0.009 \pm 0.007$ (+3%)
	REVE-Large	-0.063 ± 0.028 (−10%)	-0.034 ± 0.008 (−4%)	-0.005 ± 0.022 (−1%)

Table 16: Sleep-EDF relative performance change from No-TTA. Entries show mean delta ± standard deviation, with relative change in parentheses.

TTA Method	Foundation Model	Bal. Acc. Δ (%)	Cohen's κ Δ (%)	Weighted F1 Δ (%)
SHOT	CBraMod	-0.315 ± 0.013 (−61%)	-0.554 ± 0.003 (−100%)	-0.627 ± 0.044 (−95%)
	TFM	$+0.004 \pm 0.003$ (+1%)	-0.006 ± 0.005 (−1%)	-0.002 ± 0.002 (−0%)
	REVE-Base	-0.315 ± 0.060 (−51%)	-0.511 ± 0.075 (−80%)	-0.461 ± 0.101 (−63%)
	REVE-Large	-0.418 ± 0.019 (−64%)	-0.640 ± 0.025 (−94%)	-0.664 ± 0.057 (−87%)
T3A	CBraMod	$+0.022 \pm 0.009$ (+4%)	-0.154 ± 0.013 (−28%)	-0.093 ± 0.012 (−14%)
	TFM	-0.014 ± 0.007 (−2%)	-0.130 ± 0.020 (−23%)	-0.076 ± 0.017 (−11%)
	REVE-Base	-0.042 ± 0.010 (−7%)	-0.229 ± 0.014 (−36%)	-0.160 ± 0.014 (−22%)
	REVE-Large	-0.026 ± 0.006 (−4%)	-0.215 ± 0.015 (−32%)	-0.144 ± 0.014 (−19%)
Tent	CBraMod	-0.312 ± 0.014 (−61%)	-0.552 ± 0.004 (−100%)	-0.508 ± 0.039 (−77%)
	TFM	-0.076 ± 0.058 (−14%)	-0.086 ± 0.109 (−15%)	-0.090 ± 0.087 (−13%)
	REVE-Base	-0.183 ± 0.090 (−29%)	-0.176 ± 0.173 (−28%)	-0.160 ± 0.172 (−22%)
	REVE-Large	-0.155 ± 0.051 (−24%)	-0.121 ± 0.081 (−18%)	-0.102 ± 0.054 (−13%)

Table 17: EAR-EEG relative performance change from No-TTA. Entries show mean delta ± standard deviation, with relative change in parentheses.

TTA Method	Foundation Model	Bal. Acc. Δ (%)	Cohen's κ Δ (%)	Weighted F1 Δ (%)
SHOT	CBraMod	-0.065 ± 0.020 (-27%)	-0.087 ± 0.027 (-93%)	-0.224 ± 0.015 (-75%)
	TFM	$+0.000 \pm 0.000$ ($+0\%$)	-0.000 ± 0.001 (-0%)	-0.000 ± 0.000 (-0%)
	REVE-Base	-0.018 ± 0.032 (-5%)	-0.085 ± 0.064 (-28%)	-0.119 ± 0.076 (-25%)
	REVE-Large	-0.068 ± 0.056 (-17%)	-0.156 ± 0.083 (-42%)	-0.150 ± 0.087 (-29%)
T3A	CBraMod	$+0.048 \pm 0.012$ ($+20\%$)	$+0.064 \pm 0.016$ ($+69\%$)	$+0.018 \pm 0.009$ ($+6\%$)
	TFM	-0.005 ± 0.007 (-1%)	-0.042 ± 0.006 (-15%)	-0.009 ± 0.006 (-2%)
	REVE-Base	$+0.037 \pm 0.007$ ($+10\%$)	$+0.001 \pm 0.015$ ($+0\%$)	$+0.001 \pm 0.017$ ($+0\%$)
	REVE-Large	$+0.022 \pm 0.007$ ($+6\%$)	-0.010 ± 0.014 (-3%)	-0.009 ± 0.010 (-2%)
Tent	CBraMod	-0.064 ± 0.022 (-27%)	-0.084 ± 0.030 (-90%)	-0.189 ± 0.060 (-63%)
	TFM	-0.001 ± 0.001 (-0%)	$+0.000 \pm 0.001$ ($+0\%$)	-0.000 ± 0.001 (-0%)
	REVE-Base	-0.032 ± 0.022 (-9%)	-0.047 ± 0.035 (-15%)	-0.046 ± 0.028 (-10%)
	REVE-Large	-0.018 ± 0.014 (-5%)	-0.025 ± 0.018 (-7%)	-0.019 ± 0.015 (-4%)

A.7. Statistical Significance Testing

We compare each TTA method with the matched No-TTA baseline for every method, foundation model, dataset, and metric. Deltas are computed within each (seed, batch size) run as $\Delta_{\text{TTA}} = \text{metric}_{\text{TTA}} - \text{metric}_{\text{No-TTA}}$. Because batch sizes reuse the same seed/split/model, we average the three batch-size deltas within each seed, yielding 5 seed-level deltas for each of 180 comparisons. We run two-sided paired t-tests against zero and apply Benjamini–Hochberg FDR correction over all 180 tests. Tables 18–22 report the FDR-adjusted q-values by dataset. After correction, 145 comparisons are significant ($q < 0.05$), with 17 improvements and 128 degradations. Significant effects are mostly harmful, SHOT and Tent mainly degrade performance, and T3A accounts for most improvements while remaining mixed across settings.

Table 18: TUEV paired seed-level t-tests comparing each TTA method against No-TTA, with Benjamini–Hochberg FDR correction. Asterisks denote FDR-adjusted significance levels, with $^{*}q < 0.05$, $^{**}q < 0.01$, and $^{***}q < 0.001$. Unmarked comparisons are not significant.

TTA Method	Foundation Model	Metric	Mean Δ	95% CI	q
SHOT	CBraMod	Bal. Acc.	−0.210	[−0.243, −0.176]	$< 0.001^{***}$
		Cohen's κ	−0.462	[−0.537, −0.387]	$< 0.001^{***}$
		Weighted F1	−0.673	[−0.693, −0.653]	$< 0.001^{***}$
	TFM	Bal. Acc.	+0.024	[+0.008, +0.041]	0.020^{*}
		Cohen's κ	−0.001	[−0.021, +0.018]	0.881
		Weighted F1	−0.010	[−0.025, +0.004]	0.142
	REVE-Base	Bal. Acc.	−0.095	[−0.190, +0.001]	0.063
		Cohen's κ	−0.431	[−0.482, −0.380]	$< 0.001^{***}$
		Weighted F1	−0.574	[−0.630, −0.517]	$< 0.001^{***}$
	REVE-Large	Bal. Acc.	−0.137	[−0.170, −0.104]	$< 0.001^{***}$
		Cohen's κ	−0.473	[−0.516, −0.430]	$< 0.001^{***}$
		Weighted F1	−0.624	[−0.676, −0.573]	$< 0.001^{***}$
T3A	CBraMod	Bal. Acc.	+0.031	[+0.005, +0.056]	0.036^{*}
		Cohen's κ	−0.118	[−0.200, −0.037]	0.022^{*}
		Weighted F1	−0.096	[−0.165, −0.028]	0.024^{*}
	TFM	Bal. Acc.	+0.012	[−0.054, +0.077]	0.677
		Cohen's κ	−0.181	[−0.239, −0.122]	0.002^{**}
		Weighted F1	−0.174	[−0.228, −0.120]	0.002^{**}
	REVE-Base	Bal. Acc.	+0.072	[+0.016, +0.128]	0.030^{*}
		Cohen's κ	−0.137	[−0.239, −0.035]	0.027^{*}
		Weighted F1	−0.095	[−0.194, +0.004]	0.068
	REVE-Large	Bal. Acc.	+0.095	[+0.028, +0.162]	0.024^{*}
		Cohen's κ	−0.081	[−0.130, −0.032]	0.015^{*}
		Weighted F1	−0.044	[−0.074, −0.014]	0.021^{*}
Tent	CBraMod	Bal. Acc.	−0.212	[−0.246, −0.178]	$< 0.001^{***}$
		Cohen's κ	−0.465	[−0.539, −0.390]	$< 0.001^{***}$
		Weighted F1	−0.253	[−0.442, −0.063]	0.027^{*}
	TFM	Bal. Acc.	−0.101	[−0.135, −0.066]	0.002^{**}
		Cohen's κ	−0.169	[−0.250, −0.088]	0.008^{**}
		Weighted F1	−0.068	[−0.106, −0.031]	0.012^{*}
	REVE-Base	Bal. Acc.	−0.250	[−0.270, −0.231]	$< 0.001^{***}$
		Cohen's κ	−0.474	[−0.522, −0.427]	$< 0.001^{***}$
		Weighted F1	−0.198	[−0.221, −0.176]	$< 0.001^{***}$
	REVE-Large	Bal. Acc.	−0.208	[−0.229, −0.187]	$< 0.001^{***}$
		Cohen's κ	−0.328	[−0.350, −0.305]	$< 0.001^{***}$
		Weighted F1	−0.140	[−0.150, −0.130]	$< 0.001^{***}$

Table 19: TUAB paired seed-level t-tests comparing each TTA method against No-TTA, with Benjamini–Hochberg FDR correction. Asterisks denote FDR-adjusted significance levels, with $^{*}q < 0.05$, $^{**}q < 0.01$, and $^{***}q < 0.001$. Unmarked comparisons are not significant.

TTA Method	Foundation Model	Metric	Mean Δ	95% CI	q
SHOT	CBraMod	Bal. Acc.	−0.247	[−0.251, −0.242]	$< 0.001^{***}$
		ROC AUC	−0.303	[−0.362, −0.244]	$< 0.001^{***}$
		PR AUC	−0.337	[−0.372, −0.302]	$< 0.001^{***}$
	TFM	Bal. Acc.	−0.030	[−0.048, −0.012]	0.015^{*}
		ROC AUC	−0.012	[−0.020, −0.004]	0.018^{*}
		PR AUC	−0.011	[−0.018, −0.005]	0.012^{*}
	REVE-Base	Bal. Acc.	−0.147	[−0.189, −0.105]	0.002^{**}
		ROC AUC	−0.090	[−0.135, −0.044]	0.009^{**}
		PR AUC	−0.097	[−0.148, −0.047]	0.010^{**}
	REVE-Large	Bal. Acc.	−0.110	[−0.171, −0.049]	0.012^{*}
		ROC AUC	−0.056	[−0.080, −0.032]	0.005^{**}
		PR AUC	−0.060	[−0.084, −0.037]	0.004^{**}
T3A	CBraMod	Bal. Acc.	−0.025	[−0.031, −0.019]	$< 0.001^{***}$
		ROC AUC	−0.021	[−0.026, −0.015]	$< 0.001^{***}$
		PR AUC	−0.029	[−0.039, −0.020]	0.002^{**}
	TFM	Bal. Acc.	−0.027	[−0.035, −0.018]	0.002^{**}
		ROC AUC	−0.031	[−0.036, −0.026]	$< 0.001^{***}$
		PR AUC	−0.041	[−0.047, −0.036]	$< 0.001^{***}$
	REVE-Base	Bal. Acc.	−0.014	[−0.023, −0.005]	0.016^{*}
		ROC AUC	−0.012	[−0.017, −0.007]	0.004^{**}
		PR AUC	−0.016	[−0.021, −0.010]	0.003^{**}
	REVE-Large	Bal. Acc.	−0.003	[−0.010, +0.003]	0.268
		ROC AUC	−0.004	[−0.008, +0.001]	0.110
		PR AUC	−0.007	[−0.010, −0.005]	0.003^{**}
Tent	CBraMod	Bal. Acc.	−0.248	[−0.254, −0.241]	$< 0.001^{***}$
		ROC AUC	−0.271	[−0.303, −0.239]	$< 0.001^{***}$
		PR AUC	−0.324	[−0.339, −0.308]	$< 0.001^{***}$
	TFM	Bal. Acc.	−0.196	[−0.220, −0.172]	$< 0.001^{***}$
		ROC AUC	−0.183	[−0.210, −0.156]	$< 0.001^{***}$
		PR AUC	−0.174	[−0.205, −0.142]	$< 0.001^{***}$
	REVE-Base	Bal. Acc.	−0.223	[−0.244, −0.202]	$< 0.001^{***}$
		ROC AUC	−0.223	[−0.256, −0.190]	$< 0.001^{***}$
		PR AUC	−0.243	[−0.273, −0.214]	$< 0.001^{***}$
	REVE-Large	Bal. Acc.	−0.166	[−0.205, −0.128]	$< 0.001^{***}$
		ROC AUC	−0.134	[−0.165, −0.103]	$< 0.001^{***}$
		PR AUC	−0.147	[−0.182, −0.113]	$< 0.001^{***}$

Table 20: CHB-MIT paired seed-level t-tests comparing each TTA method against No-TTA, with Benjamini–Hochberg FDR correction. Asterisks denote FDR-adjusted significance levels, with $^{*}q < 0.05$, $^{**}q < 0.01$, and $^{***}q < 0.001$. Unmarked comparisons are not significant.

TTA Method	Foundation Model	Metric	Mean Δ	95% CI	q
SHOT	CBraMod	Bal. Acc.	+0.014	[−0.009, +0.036]	0.192
		ROC AUC	−0.246	[−0.291, −0.201]	$< 0.001^{***}$
		PR AUC	−0.069	[−0.092, −0.046]	0.002^{**}
	TFM	Bal. Acc.	+0.007	[−0.002, +0.015]	0.110
		ROC AUC	−0.003	[−0.009, +0.003]	0.248
		PR AUC	−0.027	[−0.052, −0.003]	0.044^{*}
	REVE-Base	Bal. Acc.	+0.004	[−0.073, +0.080]	0.909
		ROC AUC	−0.325	[−0.497, −0.152]	0.010^{*}
		PR AUC	−0.287	[−0.320, −0.254]	$< 0.001^{***}$
	REVE-Large	Bal. Acc.	−0.018	[−0.142, +0.106]	0.739
		ROC AUC	−0.210	[−0.355, −0.064]	0.022^{*}
		PR AUC	−0.331	[−0.454, −0.208]	0.003^{**}
T3A	CBraMod	Bal. Acc.	+0.000	[+0.000, +0.000]	1.000
		ROC AUC	+0.002	[−0.006, +0.010]	0.621
		PR AUC	+0.001	[−0.006, +0.009]	0.664
	TFM	Bal. Acc.	+0.067	[+0.059, +0.074]	$< 0.001^{***}$
		ROC AUC	−0.141	[−0.166, −0.116]	$< 0.001^{***}$
		PR AUC	−0.078	[−0.101, −0.055]	0.002^{**}
	REVE-Base	Bal. Acc.	+0.189	[+0.125, +0.253]	0.002^{**}
		ROC AUC	+0.011	[+0.005, +0.017]	0.014^{*}
		PR AUC	−0.007	[−0.044, +0.029]	0.642
	REVE-Large	Bal. Acc.	+0.187	[+0.140, +0.233]	$< 0.001^{***}$
		ROC AUC	−0.001	[−0.011, +0.009]	0.801
		PR AUC	−0.023	[−0.110, +0.064]	0.537
Tent	CBraMod	Bal. Acc.	+0.000	[−0.000, +0.001]	0.408
		ROC AUC	−0.257	[−0.269, −0.246]	$< 0.001^{***}$
		PR AUC	−0.070	[−0.105, −0.034]	0.009^{**}
	TFM	Bal. Acc.	−0.004	[−0.009, +0.001]	0.112
		ROC AUC	−0.001	[−0.003, +0.000]	0.142
		PR AUC	+0.020	[+0.008, +0.033]	0.017^{*}
	REVE-Base	Bal. Acc.	−0.013	[−0.034, +0.008]	0.192
		ROC AUC	−0.002	[−0.005, +0.001]	0.127
		PR AUC	+0.009	[+0.001, +0.018]	0.048^{*}
	REVE-Large	Bal. Acc.	−0.063	[−0.094, −0.031]	0.009^{**}
		ROC AUC	−0.034	[−0.041, −0.027]	$< 0.001^{***}$
		PR AUC	−0.005	[−0.023, +0.013]	0.502

Table 21: Sleep-EDF paired seed-level t-tests comparing each TTA method against No-TTA, with Benjamini–Hochberg FDR correction. Asterisks denote FDR-adjusted significance levels, with $^{*}q < 0.05$, $^{**}q < 0.01$, and $^{***}q < 0.001$. Unmarked comparisons are not significant.

TTA Method	Foundation Model	Metric	Mean Δ	95% CI	q
SHOT	CBraMod	Bal. Acc.	−0.315	[−0.332, −0.298]	$< 0.001^{***}$
		Cohen's κ	−0.554	[−0.558, −0.551]	$< 0.001^{***}$
		Weighted F1	−0.627	[−0.676, −0.578]	$< 0.001^{***}$
	TFM	Bal. Acc.	+0.004	[+0.001, +0.008]	0.026^{*}
		Cohen's κ	−0.006	[−0.010, −0.001]	0.026^{*}
		Weighted F1	−0.002	[−0.004, −0.000]	0.032^{*}
	REVE-Base	Bal. Acc.	−0.315	[−0.330, −0.301]	$< 0.001^{***}$
		Cohen's κ	−0.511	[−0.541, −0.480]	$< 0.001^{***}$
		Weighted F1	−0.461	[−0.491, −0.432]	$< 0.001^{***}$
	REVE-Large	Bal. Acc.	−0.418	[−0.435, −0.401]	$< 0.001^{***}$
		Cohen's κ	−0.640	[−0.668, −0.613]	$< 0.001^{***}$
		Weighted F1	−0.664	[−0.725, −0.603]	$< 0.001^{***}$
T3A	CBraMod	Bal. Acc.	+0.022	[+0.009, +0.034]	0.012^{*}
		Cohen's κ	−0.154	[−0.171, −0.137]	$< 0.001^{***}$
		Weighted F1	−0.093	[−0.109, −0.076]	$< 0.001^{***}$
	TFM	Bal. Acc.	−0.014	[−0.023, −0.004]	0.024^{*}
		Cohen's κ	−0.130	[−0.157, −0.103]	$< 0.001^{***}$
		Weighted F1	−0.076	[−0.098, −0.054]	0.002^{**}
	REVE-Base	Bal. Acc.	−0.042	[−0.055, −0.029]	0.002^{**}
		Cohen's κ	−0.229	[−0.248, −0.210]	$< 0.001^{***}$
		Weighted F1	−0.160	[−0.180, −0.141]	$< 0.001^{***}$
	REVE-Large	Bal. Acc.	−0.026	[−0.034, −0.018]	0.002^{**}
		Cohen's κ	−0.215	[−0.235, −0.194]	$< 0.001^{***}$
		Weighted F1	−0.144	[−0.163, −0.124]	$< 0.001^{***}$
Tent	CBraMod	Bal. Acc.	−0.312	[−0.330, −0.294]	$< 0.001^{***}$
		Cohen's κ	−0.552	[−0.556, −0.548]	$< 0.001^{***}$
		Weighted F1	−0.508	[−0.532, −0.483]	$< 0.001^{***}$
	TFM	Bal. Acc.	−0.076	[−0.111, −0.042]	0.006^{**}
		Cohen's κ	−0.086	[−0.153, −0.020]	0.030^{*}
		Weighted F1	−0.090	[−0.143, −0.037]	0.014^{*}
	REVE-Base	Bal. Acc.	−0.183	[−0.226, −0.140]	$< 0.001^{***}$
		Cohen's κ	−0.176	[−0.278, −0.074]	0.013^{*}
		Weighted F1	−0.160	[−0.263, −0.057]	0.018^{*}
	REVE-Large	Bal. Acc.	−0.155	[−0.162, −0.147]	$< 0.001^{***}$
		Cohen's κ	−0.121	[−0.154, −0.088]	0.001^{**}
		Weighted F1	−0.102	[−0.121, −0.084]	$< 0.001^{***}$

Table 22: EAR-EEG paired seed-level t-tests comparing each TTA method against No-TTA, with Benjamini–Hochberg FDR correction. Asterisks denote FDR-adjusted significance levels, with $^{*}q < 0.05$, $^{**}q < 0.01$, and $^{***}q < 0.001$. Unmarked comparisons are not significant.

TTA Method	Foundation Model	Metric	Mean Δ	95% CI	q
SHOT	CBraMod	Bal. Acc.	−0.065	[−0.089, −0.042]	0.003^{**}
		Cohen's κ	−0.087	[−0.119, −0.054]	0.003^{**}
		Weighted F1	−0.224	[−0.238, −0.210]	$< 0.001^{***}$
	TFM	Bal. Acc.	+0.000	[−0.000, +0.000]	0.777
		Cohen's κ	−0.000	[−0.001, +0.000]	0.514
		Weighted F1	−0.000	[−0.001, +0.000]	0.204
	REVE-Base	Bal. Acc.	−0.018	[−0.031, −0.006]	0.021^{*}
		Cohen's κ	−0.085	[−0.099, −0.071]	$< 0.001^{***}$
		Weighted F1	−0.119	[−0.133, −0.105]	$< 0.001^{***}$
	REVE-Large	Bal. Acc.	−0.068	[−0.118, −0.018]	0.026^{*}
		Cohen's κ	−0.156	[−0.221, −0.090]	0.005^{**}
		Weighted F1	−0.150	[−0.207, −0.094]	0.003^{**}
T3A	CBraMod	Bal. Acc.	+0.048	[+0.033, +0.063]	0.002^{**}
		Cohen's κ	+0.064	[+0.043, +0.085]	0.002^{**}
		Weighted F1	+0.018	[+0.006, +0.029]	0.018^{*}
	TFM	Bal. Acc.	−0.005	[−0.014, +0.005]	0.279
		Cohen's κ	−0.042	[−0.050, −0.034]	$< 0.001^{***}$
		Weighted F1	−0.009	[−0.016, −0.002]	0.031^{*}
	REVE-Base	Bal. Acc.	+0.037	[+0.028, +0.045]	$< 0.001^{***}$
		Cohen's κ	+0.001	[−0.018, +0.020]	0.903
		Weighted F1	+0.001	[−0.022, +0.023]	0.931
	REVE-Large	Bal. Acc.	+0.022	[+0.014, +0.031]	0.004^{**}
		Cohen's κ	−0.010	[−0.028, +0.009]	0.243
		Weighted F1	−0.009	[−0.023, +0.004]	0.150
Tent	CBraMod	Bal. Acc.	−0.064	[−0.085, −0.043]	0.002^{**}
		Cohen's κ	−0.084	[−0.116, −0.052]	0.003^{**}
		Weighted F1	−0.189	[−0.226, −0.153]	$< 0.001^{***}$
	TFM	Bal. Acc.	−0.001	[−0.001, +0.000]	0.197
		Cohen's κ	+0.000	[−0.000, +0.001]	0.267
		Weighted F1	−0.000	[−0.001, +0.000]	0.072
	REVE-Base	Bal. Acc.	−0.032	[−0.042, −0.022]	0.002^{**}
		Cohen's κ	−0.047	[−0.062, −0.031]	0.002^{**}
		Weighted F1	−0.046	[−0.059, −0.033]	0.002^{**}
	REVE-Large	Bal. Acc.	−0.018	[−0.024, −0.012]	0.003^{**}
		Cohen's κ	−0.025	[−0.033, −0.016]	0.003^{**}
		Weighted F1	−0.019	[−0.028, −0.010]	0.007^{**}

A.8. Test-Time Computational Cost

We report measured adaptation cost on the full evaluation grid. Runtime is the wall-clock time spent in the adaptation step before final evaluation. Peak memory is the maximum CUDA allocation observed during adaptation. No-TTA has no adaptation step and is not included in the table. Each entry aggregates 75 runs across five datasets (TUEV, TUAB, CHB-MIT, Sleep-EDF, EAR-EEG), five seeds (3, 5, 6, 5635, 10996), and three adaptation

batch sizes (64, 128, 256). Table 23 reports adaptation latency and memory requirements across all datasets, seeds, and adaptation batch sizes.

Table 23: Measured adaptation cost on NVIDIA H200 GPUs, summarized by TTA method and foundation model. Runtime is reported per EEG window and peak memory is the maximum CUDA allocation observed during adaptation.

TTA Method	Foundation Model	Median μs/window	Max μs/window	Median peak GB	Max peak GB
T3A	CBraMod	0.16	1.89	0.08	0.08
	TFM	0.28	3.41	0.07	0.07
	REVE-Base	0.25	2.97	0.32	0.32
	REVE-Large	0.25	2.74	1.52	1.52
Tent	CBraMod	0.26	2.68	0.07	0.07
	TFM	0.26	2.70	0.04	0.04
	REVE-Base	0.28	3.52	0.55	0.55
	REVE-Large	0.39	5.36	2.94	2.94
SHOT	CBraMod	190.82	2496.89	0.25	0.54
	TFM	285.94	3349.99	1.61	6.22
	REVE-Base	768.53	3403.66	0.96	2.11
	REVE-Large	3183.54	4305.95	3.86	6.55

A.9. Source-Target Structure Visualization

We use t-SNE to qualitatively inspect source-target structure under the distribution shifts evaluated in the benchmark. We use source for the training split and target for the held-out test split. For each dataset, the first column shows raw EEG and the remaining columns show frozen foundation-model embeddings before the downstream classifier. The top row colors points by task class, and the bottom row colors the same points by source or target split. Visible source-target structure suggests that the representation is not fully domain-invariant in this diagnostic view. These plots motivate evaluating test-time adaptation, while quantitative conclusions come from the benchmark results and significance analyses. For EAR-EEG, this diagnostic is especially relevant because it is an unseen ear-EEG modality, while the evaluated foundation models were pretrained primarily on scalp EEG.

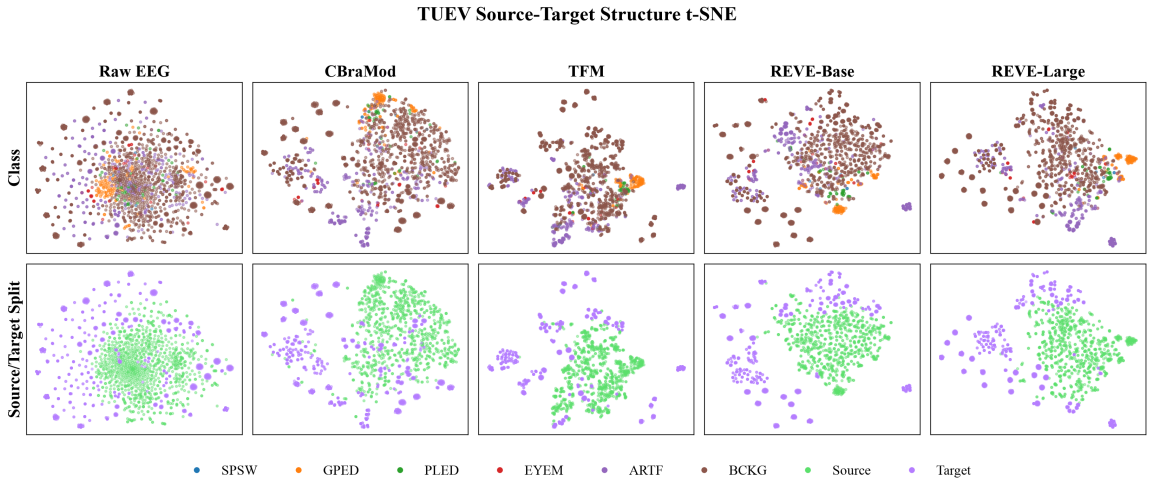

Figure 6: TUEV source-target t-SNE visualization for raw EEG and frozen foundation-model embeddings.

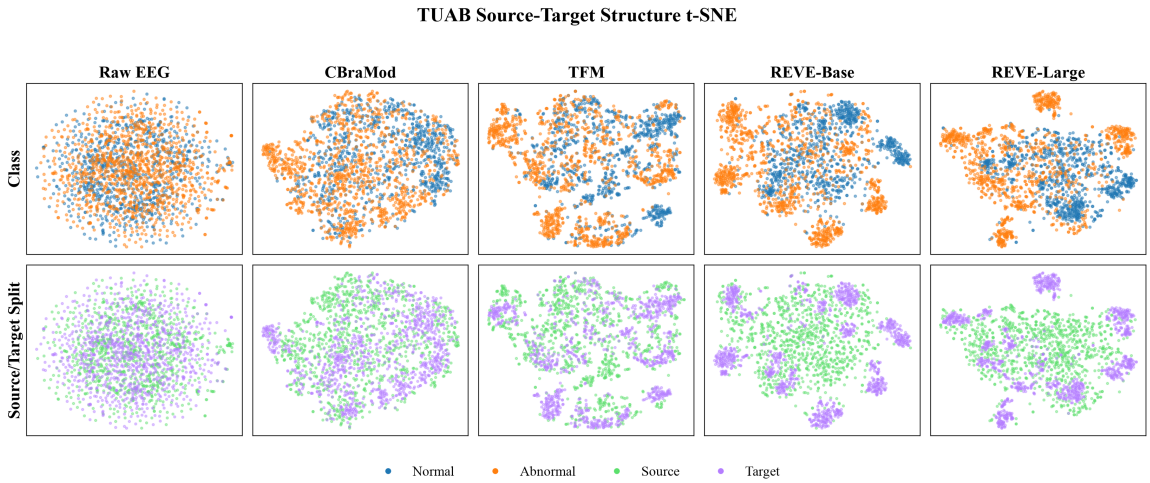

Figure 7: TUAB source-target t-SNE visualization for raw EEG and frozen foundation-model embeddings.

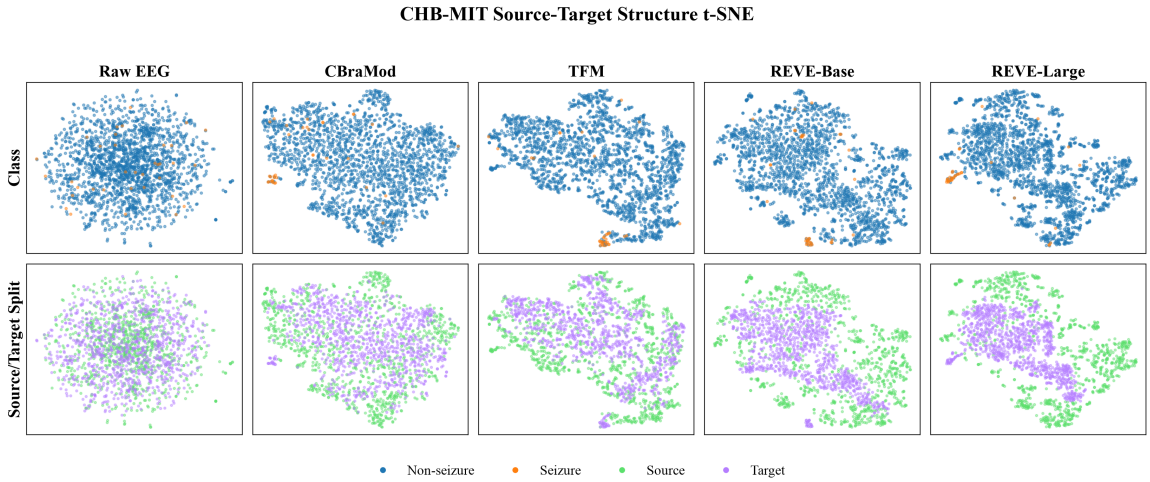

Figure 8: CHB-MIT source-target t-SNE visualization for raw EEG and frozen foundation-model embeddings.

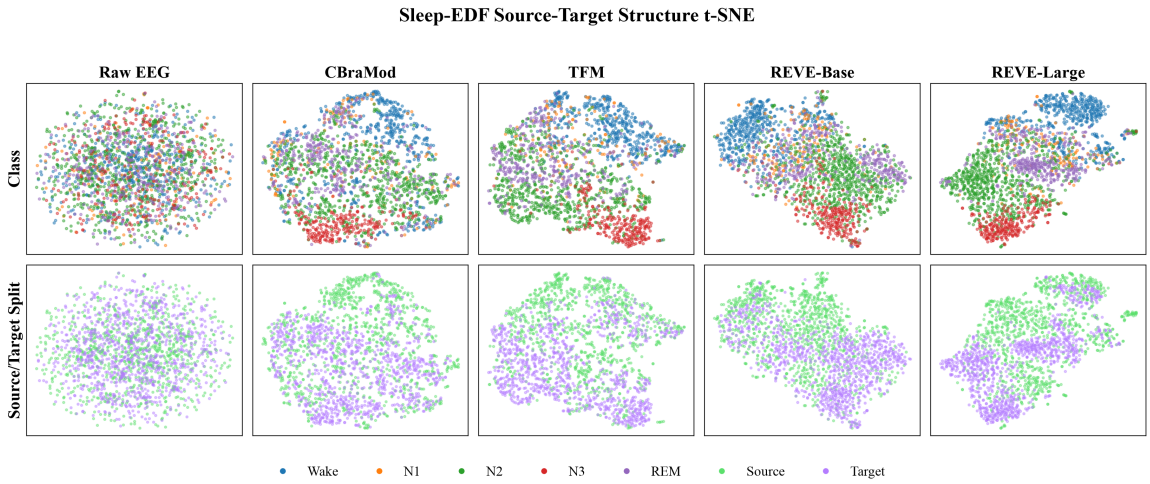

Figure 9: Sleep-EDF source-target t-SNE visualization for raw EEG and frozen foundation-model embeddings.

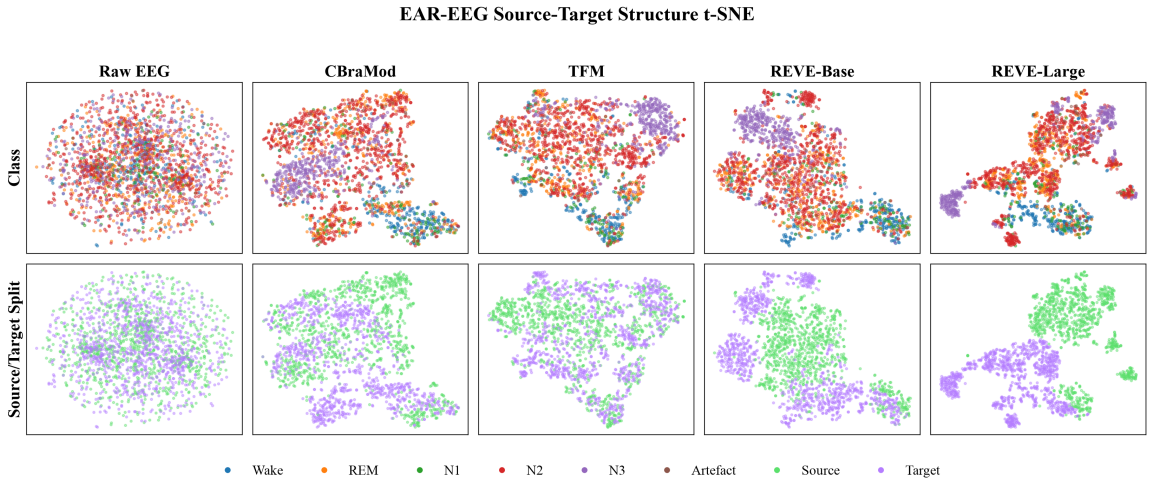

Figure 10: EAR-EEG source-target t-SNE visualization for raw EEG and frozen foundation-model embeddings.

A.10. Absolute Performance by Dataset

Values are mean ± standard deviation aggregated across adaptation batch sizes (64, 128, 256) and reported study seeds.

Table 24: CHB-MIT and TUAB

Dataset	Method	Foundation Model	Performance Metrics		
			Bal. Acc.	PR AUC	ROC AUC
CHB-MIT	No-TTA	CBraMod	0.500 ± 0.000	0.093 ± 0.015	0.750 ± 0.020
		TFM	0.534 ± 0.006	0.331 ± 0.024	0.864 ± 0.007
		REVE-Base	0.552 ± 0.039	0.318 ± 0.019	0.843 ± 0.010
		REVE-Large	0.608 ± 0.042	0.404 ± 0.020	0.890 ± 0.015
	SHOT	CBraMod	0.514 ± 0.030	0.024 ± 0.013	0.504 ± 0.025
		TFM	0.541 ± 0.012	0.304 ± 0.038	0.861 ± 0.011
		REVE-Base	0.556 ± 0.070	0.031 ± 0.010	0.518 ± 0.131
		REVE-Large	0.590 ± 0.074	0.072 ± 0.091	0.680 ± 0.126
	Tent	CBraMod	0.500 ± 0.001	0.023 ± 0.018	0.493 ± 0.057
		TFM	0.530 ± 0.003	0.351 ± 0.022	0.863 ± 0.006
		REVE-Base	0.539 ± 0.025	0.327 ± 0.016	0.840 ± 0.010
		REVE-Large	0.545 ± 0.026	0.398 ± 0.028	0.856 ± 0.015
	T3A	CBraMod	0.500 ± 0.000	0.094 ± 0.019	0.752 ± 0.024
		TFM	0.601 ± 0.011	0.253 ± 0.011	0.723 ± 0.020
		REVE-Base	0.741 ± 0.019	0.310 ± 0.027	0.853 ± 0.012
		REVE-Large	0.795 ± 0.010	0.380 ± 0.054	0.889 ± 0.010
TUAB	No-TTA	CBraMod	0.749 ± 0.004	0.823 ± 0.002	0.822 ± 0.001
		TFM	0.761 ± 0.007	0.830 ± 0.003	0.844 ± 0.004
		REVE-Base	0.801 ± 0.005	0.889 ± 0.004	0.879 ± 0.004
		REVE-Large	0.810 ± 0.005	0.899 ± 0.003	0.889 ± 0.004
	SHOT	CBraMod	0.502 ± 0.004	0.486 ± 0.032	0.519 ± 0.049
		TFM	0.732 ± 0.014	0.818 ± 0.007	0.832 ± 0.009
		REVE-Base	0.654 ± 0.058	0.791 ± 0.065	0.790 ± 0.059
		REVE-Large	0.700 ± 0.057	0.839 ± 0.033	0.833 ± 0.032
	Tent	CBraMod	0.501 ± 0.003	0.499 ± 0.028	0.551 ± 0.041
		TFM	0.565 ± 0.035	0.656 ± 0.041	0.661 ± 0.032
		REVE-Base	0.578 ± 0.044	0.645 ± 0.074	0.656 ± 0.059
		REVE-Large	0.644 ± 0.082	0.751 ± 0.088	0.755 ± 0.076
	T3A	CBraMod	0.724 ± 0.006	0.794 ± 0.006	0.801 ± 0.004
		TFM	0.734 ± 0.003	0.788 ± 0.004	0.813 ± 0.004
		REVE-Base	0.787 ± 0.005	0.873 ± 0.002	0.867 ± 0.002
		REVE-Large	0.807 ± 0.005	0.892 ± 0.004	0.885 ± 0.007

Table 25: EarEEG, SleepEDF-78, and TUEV

Dataset	Method	Foundation Model	Performance Metrics		
			Bal. Acc.	Cohen's κ	Weighted F1
EarEEG	No-TTA	CBraMod	0.238 ± 0.019	0.093 ± 0.026	0.299 ± 0.012
		TFM	0.371 ± 0.007	0.285 ± 0.007	0.426 ± 0.005
		REVE-Base	0.360 ± 0.008	0.307 ± 0.010	0.468 ± 0.008
		REVE-Large	0.400 ± 0.018	0.373 ± 0.021	0.510 ± 0.013
	SHOT	CBraMod	0.173 ± 0.012	0.007 ± 0.013	0.075 ± 0.015
		TFM	0.372 ± 0.007	0.285 ± 0.007	0.426 ± 0.005
		REVE-Base	0.342 ± 0.034	0.222 ± 0.065	0.349 ± 0.076
		REVE-Large	0.332 ± 0.064	0.218 ± 0.091	0.360 ± 0.091
	Tent	CBraMod	0.174 ± 0.018	0.009 ± 0.021	0.109 ± 0.057
		TFM	0.371 ± 0.007	0.286 ± 0.007	0.425 ± 0.005
		REVE-Base	0.328 ± 0.022	0.260 ± 0.035	0.422 ± 0.029
		REVE-Large	0.382 ± 0.023	0.348 ± 0.030	0.491 ± 0.022
	T3A	CBraMod	0.286 ± 0.012	0.157 ± 0.016	0.316 ± 0.006
		TFM	0.367 ± 0.009	0.243 ± 0.007	0.417 ± 0.003
		REVE-Base	0.397 ± 0.014	0.308 ± 0.020	0.469 ± 0.017
		REVE-Large	0.422 ± 0.014	0.364 ± 0.020	0.501 ± 0.013
SleepEDF-78	No-TTA	CBraMod	0.514 ± 0.013	0.554 ± 0.002	0.661 ± 0.004
		TFM	0.564 ± 0.003	0.577 ± 0.004	0.686 ± 0.003
		REVE-Base	0.622 ± 0.006	0.637 ± 0.006	0.735 ± 0.003
		REVE-Large	0.651 ± 0.003	0.678 ± 0.004	0.766 ± 0.003
	SHOT	CBraMod	0.199 ± 0.002	-0.000 ± 0.001	0.034 ± 0.043
		TFM	0.568 ± 0.004	0.572 ± 0.005	0.684 ± 0.003
		REVE-Base	0.307 ± 0.061	0.126 ± 0.075	0.273 ± 0.101
		REVE-Large	0.233 ± 0.021	0.038 ± 0.026	0.102 ± 0.059
	Tent	CBraMod	0.202 ± 0.002	0.002 ± 0.002	0.153 ± 0.039
		TFM	0.488 ± 0.058	0.491 ± 0.109	0.596 ± 0.087
		REVE-Base	0.439 ± 0.091	0.461 ± 0.175	0.575 ± 0.173
		REVE-Large	0.496 ± 0.052	0.557 ± 0.082	0.664 ± 0.055
	T3A	CBraMod	0.535 ± 0.008	0.400 ± 0.011	0.568 ± 0.010
		TFM	0.550 ± 0.006	0.447 ± 0.018	0.610 ± 0.014
		REVE-Base	0.580 ± 0.005	0.408 ± 0.012	0.575 ± 0.013
		REVE-Large	0.625 ± 0.006	0.463 ± 0.013	0.622 ± 0.013
TUEV	No-TTA	CBraMod	0.378 ± 0.026	0.464 ± 0.056	0.720 ± 0.028
		TFM	0.369 ± 0.011	0.399 ± 0.030	0.684 ± 0.016
		REVE-Base	0.454 ± 0.016	0.559 ± 0.038	0.771 ± 0.018
		REVE-Large	0.494 ± 0.007	0.585 ± 0.018	0.786 ± 0.009
	SHOT	CBraMod	0.168 ± 0.001	0.003 ± 0.002	0.047 ± 0.016
		TFM	0.394 ± 0.017	0.398 ± 0.025	0.674 ± 0.019
		REVE-Base	0.359 ± 0.099	0.128 ± 0.064	0.197 ± 0.086
		REVE-Large	0.357 ± 0.038	0.112 ± 0.031	0.161 ± 0.049
	Tent	CBraMod	0.166 ± 0.003	-0.001 ± 0.002	0.467 ± 0.177
		TFM	0.268 ± 0.039	0.231 ± 0.080	0.616 ± 0.029
		REVE-Base	0.204 ± 0.025	0.085 ± 0.067	0.572 ± 0.029
		REVE-Large	0.286 ± 0.061	0.258 ± 0.121	0.646 ± 0.050
	T3A	CBraMod	0.408 ± 0.032	0.346 ± 0.040	0.624 ± 0.037
		TFM	0.381 ± 0.045	0.219 ± 0.026	0.510 ± 0.039
		REVE-Base	0.526 ± 0.050	0.422 ± 0.106	0.676 ± 0.089
		REVE-Large	0.588 ± 0.048	0.505 ± 0.052	0.742 ± 0.031