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

Gabriel Jason Lee [1 *] **Jathurshan Pradeepkumar** [1 *] **Jimeng Sun** [1]

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 multiple pretrained EEG foundation models across diverse downstream tasks and heterogeneous datasets spanning in-distribution, out-of-distribution, and extreme modality shifts (e.g., Ear-EEG). Our results show that standard TTA methods yield inconsistent gains and often degrade performance, with gradient-based approaches particularly prone to heavy degradation, while optimization-free methods demonstrate greater stability and more reliable improvements. Code is available at `https://github.com/leegabriel/NeuroAdapt-Bench`.

1. Introduction

Electroencephalography (EEG) offers high-resolution measurements of neuronal activity, capturing brain dynamics at the millisecond scale, making it 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.

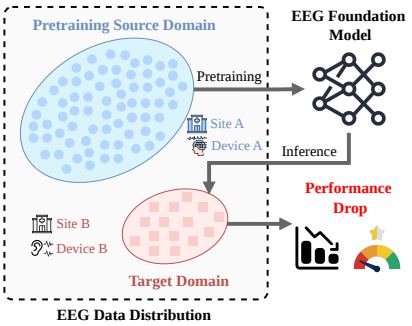

Figure 1. Distribution shift in EEG foundation model deployment. Pretrained EEG models often degrade when applied to new sites and devices, motivating the need for label and source-free test-time adaptation.

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-

[*]Equal contribution [1]University of Illinois Urbana-Champaign, Urbana, IL, USA. Correspondence to: Gabriel Jason Lee <gjlee4@illinois.edu>, Jathurshan Pradeepkumar <jp65@illinois.edu>, Jimeng Sun <jimeng@illinois.edu>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

training, prototype adjustment, and source-free pseudo-label refinement (Wang et al., 2021; 2022; Iwasawa & Matsuo, 2021; Liang et al., 2020; Liu et al., 2024; Wang et al., 2025b). Additional background on EEG foundation models and TTA is provided in Appendix A.

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 evaluate TTA across EEG tasks and deployment settings spanning (1) *in-distribution* subject-level variability, (2) *out-of-distribution* cross-dataset shifts in task, population, and acquisition protocol, and (3) *extreme* shifts from unseen modalities and recording configurations (e.g., Ear-EEG (Tabar et al., 2024)). We report performance relative to a No-TTA baseline throughout.
- **Key insights on TTA for EEG:** Standard TTA yields inconsistent gains and often degrades performance under distribution shift, while optimization-free methods are generally more stable than gradient-based alternatives. 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, with its largest gain reaching +18.9 points for REVE-Base on CHB-MIT (Shoeb & Guttag, 2010). Detailed per-dataset deltas are provided in Appendix D.5.
- **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.

2. NeuroAdapt-Bench

The benchmark follows a shared downstream pipeline: a lightweight classifier is fine-tuned on labeled source data while the pretrained backbone remains fixed, the held-out test split is treated as unlabeled target data for adaptation, and final performance is evaluated against a No-TTA baseline. We report balanced accuracy, ROC-AUC, and PR-AUC for binary tasks, and balanced accuracy, Cohen's κ, and weighted F_1 for multiclass tasks. To isolate adaptation effects, we measure per-seed relative improvement $\Delta_{\text{TTA}} = \text{metric}_{\text{TTA}} - \text{metric}_{\text{No-TTA}}$. The formal TTA setup, method definitions, and full benchmark design are detailed in Appendix B, Appendix B.1, and Appendix B.2.

Table 1. Evaluated TTA methods.

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

The benchmark spans five EEG datasets: TUEV and TUAB as in-distribution settings, CHB-MIT and SleepEDF-78 as cross-dataset out-of-distribution settings, and Ear-EEG as an extreme modality-shift setting (Harati et al., 2015; Lopez et al., 2015; Shoeb & Guttag, 2010; Goldberger et al., 2000; Eldele, 2022; Bjarke Mikkelsen et al., 2025; Tabar et al., 2024). We evaluate CBraMod, TFM-Tokenizer, REVE-Base, and REVE-Large (Wang et al., 2024b; Pradeepkumar et al., 2026b; Ouahidi et al., 2025). Full experiment-setup details are provided in Appendix B.3; preprocessing, fine-tuning, adaptation, evaluation, and full per-dataset tables are detailed in Appendix D.1, Appendix D.2, Appendix D.3, Appendix D.4, Appendix D.5, and Appendix D.6.

3. Results and Discussion

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

To establish the clearest baseline, we first examine whether TTA helps under the mildest deployment shift, where variability is primarily patient-level. 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 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, gradient-based methods (Tent

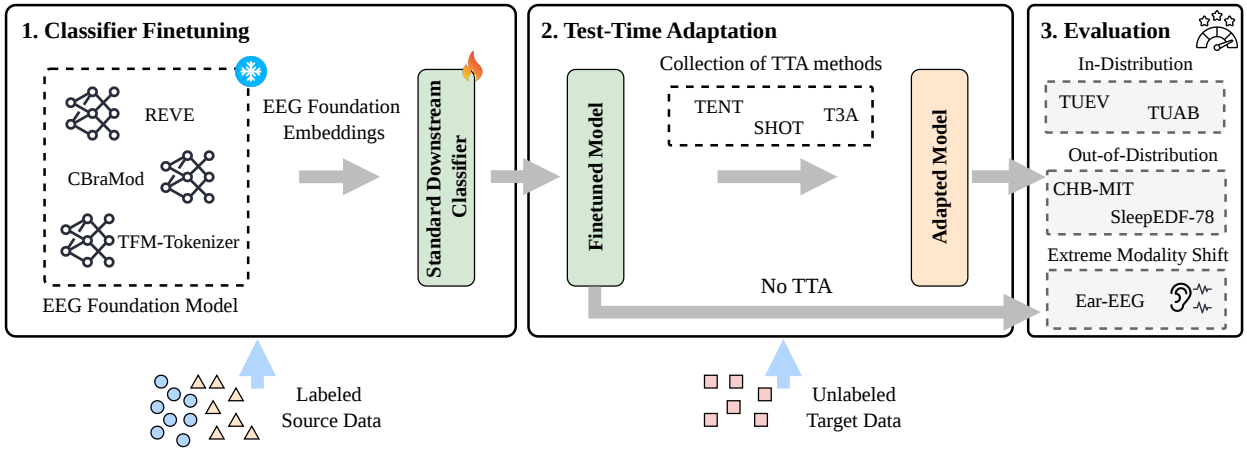

Figure 2. Overview of NeuroAdapt-Bench. The benchmark consists of three stages: (1) supervised fine-tuning of an EEG foundation model with a classification head on labeled source-domain data; (2) optional TTA on unlabeled target-domain data using Tent, SHOT, or T3A, alongside a No-TTA baseline; and (3) evaluation on target-domain samples to measure robustness under distribution shift.

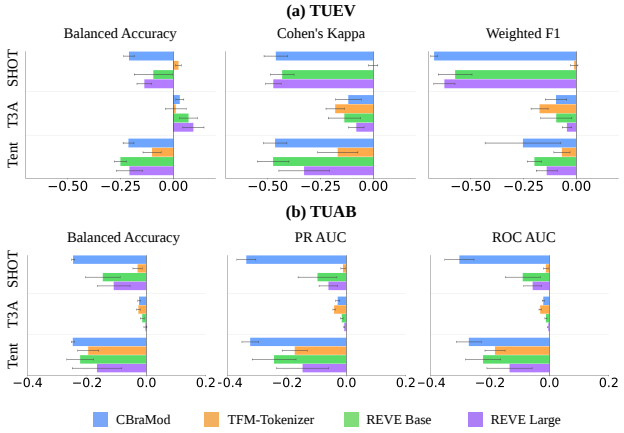

Figure 3. TTA relative performance on in-distribution datasets (TUEV and TUAB). (a) Δ_{TTA} on TUEV relative to the No-TTA baseline; (b) Δ_{TTA} on TUAB relative to the No-TTA baseline.

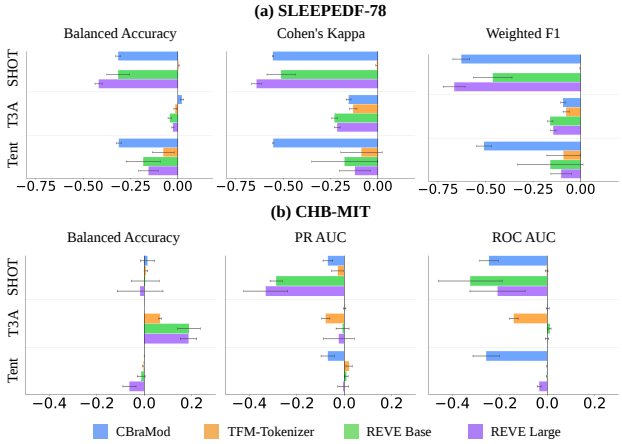

Figure 4. TTA relative performance on out-of-distribution datasets (SleepEDF-78 and CHB-MIT). (a) Δ_{TTA} on SleepEDF-78 relative to the No-TTA baseline; (b) Δ_{TTA} on CHB-MIT relative to the No-TTA baseline.

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 TTA methods degrade performance, with Tent showing the largest drop.

These results suggest that when the target data closely matches the pretraining distribution, the learned representations are already well-aligned, leaving limited room for improvement through adaptation. In such cases, TTA, particularly gradient-based approaches, can disrupt these representations and lead to negative transfer. The relative robustness of T3A likely stems from its optimization-free design, which avoids destabilizing updates and instead leverages confident predictions to refine class-level representations. Detailed deltas appear in Appendix D.5.

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

A more demanding question is whether TTA remains useful when deployment shifts across datasets, tasks, and acquisition protocols. Figure 4 summarizes the relative performance of TTA methods compared to the No-TTA baseline on SleepEDF-78 and CHB-MIT. 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. On CHB-MIT, T3A provides consistent improvements in balanced accuracy across most models. This behavior is likely due to the class imbalance in CHB-MIT, where T3A's prototype-based updates improve class-wise calibration. The REVE family, in particular, benefits from T3A

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

TTA Method	Foundation Model	Balanced Acc.Δ	Cohen's Kappa Δ	Weighted F1 Δ
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	**+0.048** ± 0.012	**+0.064** ± 0.016	**+0.018** ± 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

on balanced accuracy and shows little to no degradation in ROC-AUC and PR-AUC. However, REVE shows stronger degradation in PR-AUC and ROC-AUC under SHOT, indicating sensitivity of gradient-based adaptation to distribution shift.

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 and channel configuration, making adaptation particularly difficult. T3A offers only marginal gains there, while Tent and SHOT consistently degrade performance. TFM-Tokenizer shows relatively greater robustness, with smaller performance drops compared to other models.

3.3. Can TTA handle unseen EEG modalities such as Ear-EEG?

The strongest question is whether TTA remains useful under extreme modality shift, which we evaluate 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 (Bjarke Mikkelsen et al., 2025; Anandakumar et al., 2023). Table 2 summarizes the relative improvement in this setting. Overall, TTA methods are unstable under this modality shift. Gradient-based approaches consistently degrade performance across models and metrics. In contrast, T3A is more stable and yields improvements for some models, notably CBraMod across all metrics, with moderate balanced-accuracy gains for REVE.

3.4. Discussion and Implications for Future Works

In our experiments, T3A is by far the most stable method. It causes the least degradation relative to the No-TTA baseline and occasionally yields clear gains. As an optimization-free approach, T3A avoids updating model parameters, suggesting that methods that preserve pretrained representations are better suited for heterogeneous clinical EEG deployment settings. In contrast, gradient-based methods frequently lead to substantial performance degradation, often outweighing their occasional benefits on specific model architectures. This pattern also suggests that the update mechanism matters more than the online-versus-offline distinction in Table 1: although Tent and T3A are both online methods, their behavior differs sharply, while SHOT shares much of Tent's instability despite operating offline.

Batch size appears to play a more limited role. Larger adaptation batches sometimes mildly reduce degradation for gradient-based methods, but they do not consistently improve performance, and T3A is largely insensitive to batch size. These results suggest that simply increasing batch size is insufficient to stabilize or improve TTA performance in EEG. A dedicated batch-size ablation is provided in Appendix C.

Finally, representation type appears to influence TTA behavior in a separate way. Across models with continuous and discrete tokenization approaches, responses to adaptation methods vary substantially. TFM-Tokenizer shows more resistance to degradation, particularly under SHOT, while continuous embedding-based models such as REVE benefit most from T3A on CHB-MIT.

4. Conclusion

Our results highlight several important considerations for deploying TTA with EEG foundation models. First, although current models perform well on in-distribution datasets (e.g., TUAB and TUEV), they are not plug-and-play in real-world clinical settings: performance degrades substantially under out-of-distribution conditions, particularly under extreme shifts such as Ear-EEG. This underscores distribution shift as a primary barrier to reliable deployment. Second, we observe that optimization-free TTA methods exhibit greater stability than gradient-based approaches, which are prone to performance degradation. This suggests that future work should prioritize robustness and explore alternative adaptation strategies that minimize disruptive updates. Finally, differences in representation, particularly between continuous embeddings and discrete tokenization, appear to influence adaptation behavior. This points to an important and underexplored research direction: designing TTA methods tailored to the underlying representations of EEG foundation models.

References

Anandakumar, M., Pradeepkumar, J., Kappel, S. L., Edussooriya, C. U., and De Silva, A. C. 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)*, pp. 514–519. IEEE, 2023.

Bjarke Mikkelsen, K., Rezai Tabar, Y., Rævsbæk Birch, L., Lind Kappel, S., Bech Christensen, C., Dalskov Mosgaard, L., Otto, M., Christian Hemmsen, M., Lind Rank, M., and Kidmose, P. Ear-eeg sleep monitoring data sets. *Scientific Data*, 12(1):301, 2025.

Dong, J., Jia, H., Chatterjee, S., Ghosh, A., Bailey, J., and Dang, T. 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.

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

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

Guo, H., Guo, H., Li, C., Peng, H., Han, Z., and Chen, X. 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.

Harati, A., Golmohammadi, M., Lopez, S., Obeid, I., and Picone, J. Improved eeg event classification using differential energy. In *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–4. IEEE, 2015.

Huang, P.-K., Chen, T.-H., Chan, Y.-T., Chen, K.-W., and Hsu, C.-T. 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.

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

Jang, G.-D., Han, D.-K., and Lee, S.-W. Test-time adaptation for eeg-based driver drowsiness classification. In *Pattern Recognition and Artificial Intelligence*, volume 14892 of *Lecture Notes in Computer Science*, pp. 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.

Jayaram, V., Alamgir, M., Altun, Y., Scholkopf, B., and Grosse-Wentrup, M. Transfer learning in brain-computer interfaces. *IEEE Computational Intelligence Magazine*, 11(1):20–31, 2016.

Jia, H., Chen, Z., Zhu, L., Kotoge, R., Pradeepkumar, J., Matsubara, Y., Sun, J., Sakurai, Y., and Matsubara, T. 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.

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

Jiang, W.-B., Zhao, L.-M., and Lu, B.-L. Large brain model for learning generic representations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*, 2024.

Jo, Y.-Y., Lee, B. T., Hong, J.-H., Lee, H. S., Kwon, J.-m., and Kim, B. J. 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*, pp. 381–394. PMLR, 2025. URL https://proceedings.mlr.press/v287/jo25a.html.

Kastrati, A., Bürki, J., Lauer, J., Xuan, C., Iaquinto, R., and Wattenhofer, R. Eeg-bench: A benchmark for eeg foundation models in clinical applications. *arXiv preprint arXiv:2512.08959*, 2025.

Kuruppu, G., Wagh, N., Kremen, V., and Varatharajah, Y. 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.

Liang, J., Hu, D., and Feng, J. 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)*, pp. 6028–6039, 2020.

Lin, G.-T., Huang, W. P., and Lee, H.-y. 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*, pp. 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/.

Liu, H., Huang, H., and Wang, Y. Advancing test-time adaptation in wild acoustic test settings. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 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/.

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

Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., and Tan, M. 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*, pp. 16888–16905. PMLR, 2022. URL https://proceedings.mlr.press/v162/niu22a.html.

Ouahidi, Y. E., Lys, J., Thölke, P., Farrugia, N., Pasdeloup, B., Gripon, V., Jerbi, K., and Lioi, G. Reve: A foundation model for eeg–adapting to any setup with large-scale pretraining on 25,000 subjects. *arXiv preprint arXiv:2510.21585*, 2025.

Phan, H., Mikkelsen, K., Chén, O. Y., Koch, P., Mertins, A., and De Vos, M. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering*, 69 (8):2456–2467, 2022.

Pradeepkumar, J., Anandakumar, M., Kugathasan, V., Suntharalingham, D., Kappel, S. L., De Silva, A. C., and Edussooriya, C. U. Toward interpretable sleep stage classification using cross-modal transformers. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:2893–2904, 2024.

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

Pradeepkumar, J., Piao, X., Chen, Z., and Sun, J. 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.

Raj, V. A., Parupudi, T., Thalengala, A., and Nayak, S. G. 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.

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

Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A. A., and Hardt, M. Test-time training with self-supervision for generalization under distribution shifts. In Daume III, H. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9229–9248. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/sun20b.html.

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

Tabar, Y. R., Mikkelsen, K., Birch, L., Shenton, N., Kappel, S. L., Bertelsen, A. R., Nikbakht, R., Toft, H. O., Henriksen, C. H., Hemmsen, M. C., Rank, M. L., Otto, M., and Kidmose, P. "ear-eeg sleep monitoring 2023 (eesm23)", 2024.

Wang, D., Shelhamer, E., Liu, S., Olshausen, B., and Darrell, T. Tent: Fully test-time adaptation by entropy minimization. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

Wang, G., Liu, W., He, Y., Xu, C., Ma, L., and Li, H. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. *Advances in Neural Information Processing Systems*, 37:39249–39280, 2024a.

Wang, J., Zhao, S., Luo, Z., Zhou, Y., Jiang, H., Li, S., Li, T., and Pan, G. Cbramod: A criss-cross brain foundation model for eeg decoding. *arXiv preprint arXiv:2412.07236*, 2024b.

Wang, J., Zhao, S., Luo, Z., Zhou, Y., Jiang, H., Li, S., Li, T., and Pan, G. 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.

Wang, Q., Fink, O., Van Gool, L., and Dai, D. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Continual_

Test-Time_Domain_Adaptation_CVPR_2022_paper.html.

Wang, Y., Zhou, Y., Lin, Y., Chen, H., Zhang, J., Zhu, W., Hong, J., and Li, X. Dynamic model-bank test-time adaptation for automatic speech recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 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/.

Wang, Z., Luo, Y., Zheng, L., Chen, Z., Wang, S., and Huang, Z. 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.

Wu, X., Rezki, Z., Balusu, B., Zhou, J., and Bai, J. 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.

Xu, Z., Shuai, Z., Mozaffari, E., Aysola, R. S., Kumar, R., and Yang, Y. Sleeplm: Natural-language intelligence for human sleep. *arXiv preprint arXiv:2602.23605*, 2026.

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

Yang, C., Westover, M. B., and Sun, J. Manydg: Many-domain generalization for healthcare applications. *arXiv preprint arXiv:2301.08834*, 2023b.

Yao, Y., Wang, H., Chen, L., Peng, Y., and Luo, J. 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.

A. Related Work

A.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.

A.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 & 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.

B. NeuroAdapt-Bench

This section introduces *NeuroAdapt*, our benchmark for systematically evaluating representative TTA methods on EEG foundation models across diverse tasks. Section B.1 presents the problem formulation and describes the TTA methods considered. Sections B.2 and B.3 detail the benchmark design and experiment setups.

B.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 & 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.

T3A (Iwasawa & 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.

B.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 D.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 the offline method 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.

B.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 & Guttag, 2010) from PhysioNet (Goldberger et al., 2000), Ear-EEG (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. This setting focuses on sleep staging using Ear-EEG data and represents

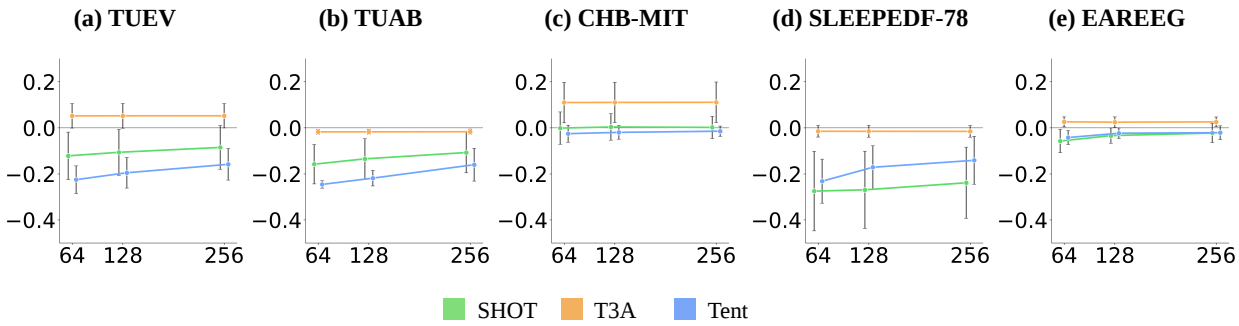

Figure 5. Balanced-accuracy improvements relative to the No-TTA baseline across adaptation batch sizes (64, 128, 256). Results are shown for (a) TUEV, (b) TUAB, (c) CHB-MIT, (d) SleepEDF-78, and (e) Ear-EEG.

a challenging out-of-domain evaluation scenario. All evaluations use patient-disjoint splits to avoid subject leakage. 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. Additional preprocessing and split details are provided in Appendix D.1.

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. Additional classifier and fine-tuning details are provided in Appendix D.2.

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. Additional test-time adaptation and evaluation details are provided in Appendix D.3 and Appendix D.4.

C. Batch Size Ablation

Figure 5 shows the effect of adaptation batch size on balanced accuracy across all models and datasets. Overall, increasing batch size does not provide consistent performance gains. For gradient-based methods, scaling the batch size from 64 to 256 mildly reduces degradation in some settings. In contrast, T3A is largely insensitive to batch size because it updates class prototypes without gradient-based optimization. These results suggest that simply increasing batch size is insufficient to stabilize or improve TTA performance in EEG.

D. Experiment Details

D.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.

D.2. Shared Downstream Classifier and Fine-Tuning

Table 4 summarizes the shared downstream classifier and fine-tuning configuration used across all foundation models. 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 Ear-EEG, 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.

D.3. Test-Time Adaptation Configuration

Table 5 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.

D.4. Evaluation and Aggregation

Table 6 summarizes the evaluation settings used throughout the benchmark.

D.5. Per-Dataset Delta Tables

Our quantitative results are summarized in Tables 7, 8, 9, 10, and 11, which report the per-dataset deltas for TUEV, TUAB, CHB-MIT, SleepEDF-78, and Ear-EEG, respectively. 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.

D.6. Absolute Performance by Dataset

Values are mean $\pm$ standard deviation aggregated across adaptation batch sizes (64, 128, 256) and reported study seeds. Tables 12 and 13 report the full absolute results.

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
Ear-EEG	Multiclass	6	4	30 s	250 Hz	Per-channel 95th percentile	Precomputed train/val/test split; final stored channel dropped before preprocessing

Table 4. 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

Table 5. 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

Table 6. 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	$\Delta_{\mathrm{TTA}} = \mathrm{metric}_{\mathrm{TTA}} - \mathrm{metric}_{\mathrm{No\text{-}TTA}}$, computed per seed before averaging
TTA evaluation batch sizes	64, 128, 256

Table 7. 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 8. 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 9. 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 10. SleepEDF-78 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 11. 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

Table 12. CHB-MIT and TUAB absolute performance.

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 13. Ear-EEG, SleepEDF-78, and TUEV absolute performance.

Dataset	Method	Foundation Model	Performance Metrics		
			Bal. Acc.	Cohen's κ	Weighted F1
Ear-EEG	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