DO WE NEED DOMAIN-SPECIFIC TIME-SERIES MODELS? INSIGHTS FROM EEG CLASSIFICATION BENCHMARKS

Liuyin Yang[*], Qiang Sun[*], & Marc M. Van Hulle
Laboratory for Neuro- and Psychophysiology, Department of Neurosciences
KU Leuven, Belgium
{liuyin.yang, qiang.sun, marc.vanhulle}@kuleuven.be

ABSTRACT

Generic time-series foundation models (TS FMs) are now trained on large-scale collections of heterogeneous time series, but whether this broad pre-training actually helps in specialized domains remains an open question. We take EEG classification as a case study, benchmarking two generic TS FMs—Mantis (∼8M parameters) and MOMENT (>300M)—on four public EEG datasets (TUEV, FACED, BCI-IV-2A, and Error) under linear probing and fine-tuning, and comparing them against EEG foundation models and classic neural baselines. Fine-tuning often matches EEG-specific models, while linear probing transfers poorly. Interestingly, randomly initialized Mantis performs comparably to its pre-trained version, suggesting that its architecture, rather than pre-training, may be driving much of its performance. These results illustrate both the promise and the limits of generic TS FMs for specialized domains.

Track: Research

1 INTRODUCTION

Electroencephalography (EEG) is a non-invasive brain recording technique widely used in brain–computer interfaces (BCIs). Yet EEG datasets are often small, differ in channel layouts, and exhibit strong subject/session variability. These factors limit the reliability of conventional deep learning models and frequently require task-specific training.

Motivated by the success of large-scale pre-training in vision and language, recent work has explored EEG FMs. Approaches such as BENDR (Kostas et al., 2021), BIOT (Yang et al., 2023), and LaBraM (Jiang et al., 2024) show that self-supervised pre-training can help, although gains over classical decoders are sometimes modest. Most EEG-specific models design pre-training tasks tailored to neural signals (e.g., reconstructing spectral structure in LaBraM), or with architectures adapted to multichannel inputs (e.g., spatial CNN in CBraMod (Wang et al., 2025)).

In parallel, a growing family of generic TS FMs, such as TimesFM (Das et al., 2024), Chronos (Ansari et al., 2024), UniTS (Gao et al., 2024), Mantis (Feofanov et al., 2025), MOMENT (Goswami et al., 2024), have demonstrated strong cross-domain results on forecasting and classification. Some of these models appear to transfer to EEG as well: Mantis, pre-trained mostly on non-neural time series, was shown to outperform EEGNet and CBraMod on motor imagery and sleep staging (Gnassounou et al., 2025). However, direct comparisons with EEG-specific foundation models on challenging BCI tasks are still rare. Recent EEG FM benchmarks (Yang et al., 2026; Xiong et al., 2026; Liu et al., 2026) report that fine-tuned foundation models work well in population-level settings but show diminishing returns under leave-one-out or low-data regimes, with linear probing consistently weak. An important question remains: when a generic TS FM does well on EEG, is it because of what it learned during pre-training, or because of its architecture? These open questions motivate a closer look at how well generalist models actually transfer to EEG.

[*]Equal contribution and shared correspondence.

In this work, we evaluate Mantis and MOMENT on four public EEG datasets: TUEV (6-class event classification) (Obeid & Picone, 2016), FACED (9-class emotion recognition) (Chen et al., 2023), BCI-IV-2A (4-class motor imagery) (Tangermann et al., 2012), and Error (error-related negativity) (Kueper et al., 2024). Following recent EEG FM benchmarks, we test cross-subject decoding on TUEV and FACED and compare population-level decoding with leave-one-out fine-tuning on BCI-IV-2A and Error. Complementing EEG-specific FM benchmarks (Yang et al., 2026), we focus on cross-domain transfer from generic TS FMs and include a random-initialization ablation to separate the effects of architecture and pre-training. We ask: (1) how well do generic TS FMs transfer to EEG, (2) is linear probing sufficient, (3) when does fine-tuning provide consistent gains, and (4) does Mantis's competitiveness stem from its pre-trained representations or its architecture?

2 METHODOLOGY

2.1 BENCHMARK MODELS

Classic EEG Neural Network Models We benchmarked two well-established convolutional neural network (CNN)-based EEG decoders: DeepConvNet (Schirrmeister et al., 2017) and EEGNet (Lawhern et al., 2018). Additionally, we included more recent, transformer-based architectures, EEG Conformer (Song et al., 2023) and CTNet (Zhao et al., 2024), which have demonstrated state-of-the-art performance yet remain computationally simpler compared to larger foundation models.

EEG Foundation Models For all benchmark datasets, we evaluated a range of EEG FMs, including BENDR (Kostas et al., 2021), BIOT (Yang et al., 2023), LaBraM (Jiang et al., 2024), EEGPT (Wang et al., 2024), CBraMod (Wang et al., 2025), and ST-EEGFormer (Yang et al., 2026). ST-EEGFormer is available in three capacity variants: small (ST-EEGFormer-s), base (ST-EEGFormer-b), and large (ST-EEGFormer-l), which share the same spatio-temporal transformer design but differ in width/depth and parameter count to trade off accuracy and compute. For all foundation models, we evaluated both linear probing and fine-tuning performance.

Time-Series Foundation Models We benchmarked Mantis (Feofanov et al., 2025), a lightweight contrastive model ($\sim$8M parameters) pre-trained on $\sim$7M time series with a small fraction of EEG data (SleepEEG and Epilepsy), and MOMENT (Goswami et al., 2024), a large masked-reconstruction transformer ($>$300M parameters). We additionally evaluated a randomly initialized Mantis (same architecture, no pre-trained weights) to disentangle the contributions of architecture and pre-training. A summary of model parameters can be found in Table 1.

2.2 EXPERIMENTAL SETUP

We consider three evaluation settings. **Population decoding** (BCI-IV-2A, Error) pools training data from all subjects to learn a single model, which is then evaluated separately on each subject. **Leave-one-subject-out** (LOO; BCI-IV-2A, Error) trains the population model on all but one held-out subject and reports its performance on the held-out subject (LOO zero-shot). We then fine-tune the same model on the held-out subject and re-evaluate it (LOO fine-tune); the change in performance on the non-held-out subjects after this adaptation measures the **generalization drop**. **Cross-subject** evaluation (TUEV, FACED) follows the standard zero-shot protocol used in prior work.

For all tasks, we report top-1 accuracy (Acc1), top-2 accuracy (Acc2), balanced

Table 1: Model parameter counts (BCI-IV-2A).

Model	Scratch	Fine-tune	Linear-prob
Classic neural networks			
EEGNet	5.8K	–	–
DeepConvNet	0.28M	–	–
Conformer	1.5M	–	–
CTNet	0.15M	–	–
EEG foundation models			
BENDR	–	3.9M	22.9K
BIOT	–	3.1M	1.4K
LaBraM	–	5.8M	0.8K
EEGPT	–	25.3M	34.3K
CBraMod	–	19.1M	14.2M
ST-EEGFormer-s	–	25.4M	2.0K
ST-EEGFormer-b	–	85.3M	3.0K
ST-EEGFormer-l	–	302.7M	4.1K
Time-series foundation models			
MOMENT	–	341.2M	4.1K
Mantis	–	8.1M	22.5K

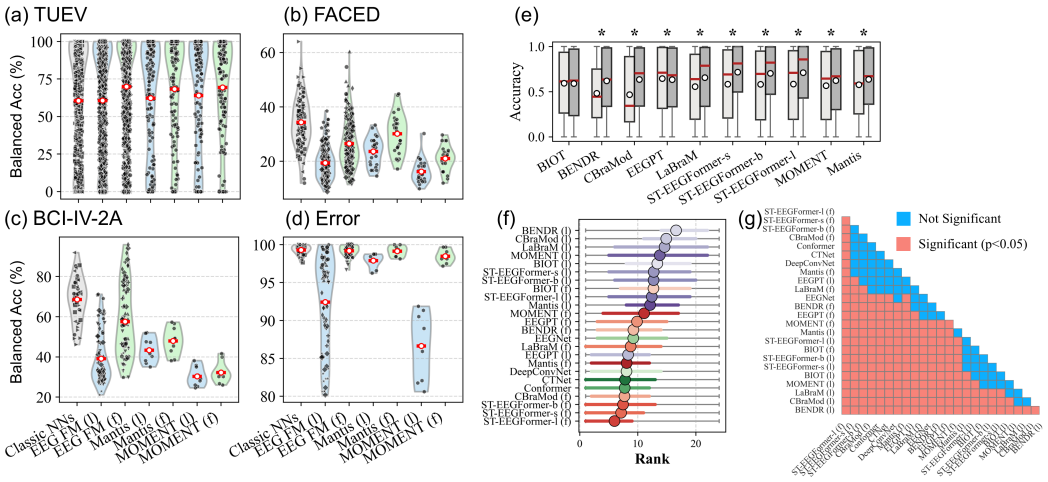

Figure 1: **Per-dataset balanced-accuracy breakdown and cross-model comparisons.** (a–d) Per-subject balanced accuracy on TUEV, FACED, BCI-IV-2A, and Error. TUEV and FACED use cross-subject evaluation, while BCI-IV-2A and Error report population decoding results. Models are grouped into classic neural-network decoders and EEG FMs, where *(f)* and *(l)* denote fine-tuning and linear probing, respectively; we compare against the generalist TS FMs Mantis and MOMENT under the same training regimes. Each dot corresponds to a subject and the red dot denotes the mean. (e) Aggregated accuracy across foundation models under linear probing vs. fine-tuning, where * indicates significance using the Wilcoxon signed-rank test with $p < 0.05$.

accuracy (BAcc), AUC, and Cohen's κ. A more detailed description of the training setup can be found in the Appendix.

3 Results and Discussion

We summarize the main findings in Figure 1 and Figure 2. Full metric tables are provided in the Appendix. We organize the discussion around four questions.

(1) How well do generic TS FMs transfer to EEG? Across datasets and evaluation protocols, the TS FMs are competitive, but they do not generally dominate. In particular, fine-tuned Mantis consistently performs in the upper tier, whereas fine-tuned MOMENT tends to lag behind most fine-tuned EEG FMs. The rank-based summary and significance testing in Figure 1(e) further support this: Mantis is among the stronger generalist baselines after fine-tuning, while MOMENT is often significantly worse than the best-performing EEG FMs.

(2) Is linear probing sufficient? Linear probing is consistently weaker than end-to-end fine-tuning for both EEG FMs and TS FMs. Figure 1(e) shows a clear aggregate advantage for fine-tuning over linear probing, indicating that the representations learned during pre-training are not directly applicable to downstream EEG tasks and typically require task-specific adaptation.

(3) When does fine-tuning provide consistent gains? Fine-tuning generally improves performance on the target subject/task, but the magnitude of the gain depends on the dataset and evaluation setting. In population or cross-subject decoding, fine-tuned foundation models can match or exceed classic neural network decoders trained from scratch on some datasets. For example, Mantis and MOMENT have higher average performance than classic neural network decoders on TUEV, but not on the others, especially on BCI-IV-2A, where both Mantis and MOMENT show much lower decoding performance than classic neural network decoders and the average of the EEG FMs. The LOO analysis in Figure 2 highlights an additional trade-off: adapting to the held-out subject can induce a generalization drop on the previously seen training subjects. Compared with scratch-trained

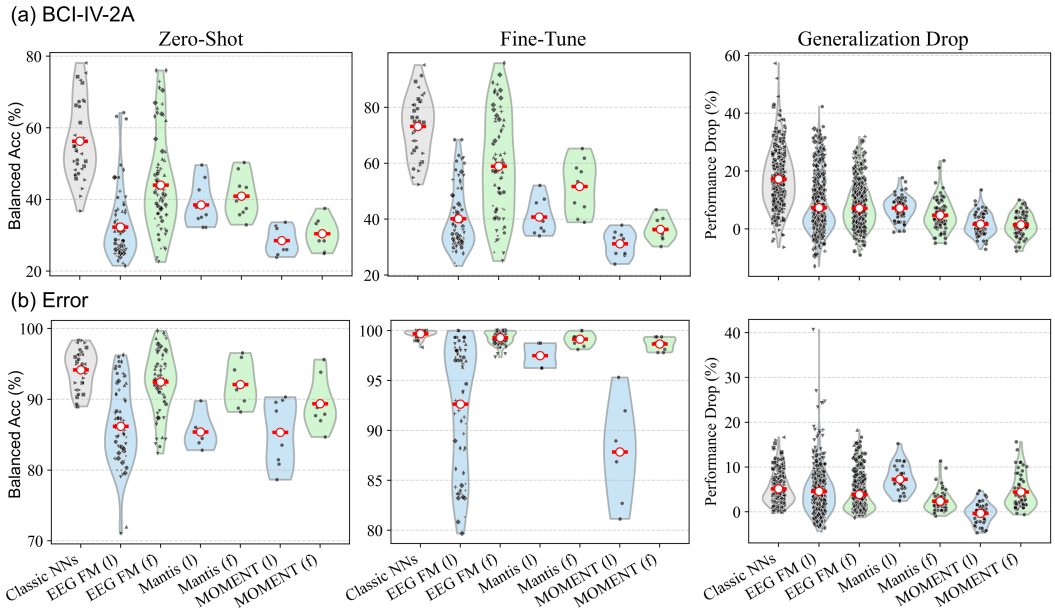

Figure 2: **Leave-one-subject-out fine-tuning analysis.** (a) BCI-IV-2A leave-one-subject-out (LOO) fine-tuning results. (b) Error dataset LOO fine-tuning results. From left to right, we show the zero-shot decoding results on the held-out (LOO) subject, the performance after fine-tuning on that held-out subject, and the generalization drop on the seen training subjects (before vs. after fine-tuning).

models, foundation models often show smaller drops, suggesting that pre-training yields representations that are more stable under subject-specific adaptation (i.e., less catastrophic forgetting).

(4) Comparing Mantis and MOMENT, and the role of pre-training. Despite having far fewer parameters, Mantis outperforms MOMENT under both linear probing and fine-tuning in our benchmarks. To examine the role of pre-training more directly, we fine-tuned a randomly initialized Mantis with the same architecture on all four datasets. The pattern is informative. On BCI-IV-2A, the randomly initialized model outperforms the pre-trained model in population decoding (balanced accuracy $p<0.01$, AUC $p<0.05$, paired Wilcoxon test), whereas on TUEV and FACED the two perform similarly ($p>0.5$). By contrast, on Error, pre-training provides a modest but significant advantage in the leave-one-out fine-tuning setting (Cohen's κ 0.971 vs. 0.936, $p<0.01$). Overall, these findings suggest that Mantis's EEG performance is driven more by its architecture than by its pre-trained representations. Consistent with this interpretation, CTNet, a much smaller model with only 0.15M parameters trained from scratch, still outperforms both TS FMs on BCI-IV-2A. This observation underscores that model scale alone does not guarantee better transfer when the architecture is not well matched to the task. Full results are reported in Table 12 in the Appendix.

4 Conclusion

This work evaluated whether generic TS FMs can transfer to EEG classification across four public datasets and multiple evaluation protocols. Overall, our results indicate that using TS FMs for EEG decoding is feasible: after end-to-end adaptation, models such as Mantis can reach competitive performance and sometimes match or surpass EEG-specific baselines.

However, linear probing is consistently weak across datasets, suggesting that off-the-shelf TS FM representations are not directly applicable to downstream EEG tasks and typically require task-specific adaptation. Moreover, even after fine-tuning, TS FMs are often not the top-performing approach and can lag behind small, task-trained neural networks, highlighting a remaining gap between current generic TS FMs and the requirements of robust EEG decoding.

Taken together, these findings suggest that domain-specific EEG FMs remain important in the near term. From a representation-learning perspective, this is plausible: generic TS FMs are trained to capture structure that transfers across many time-series domains, whereas EEG decoding may additionally require representations that are more closely aligned with domain-specific signal characteristics. The random-initialization results are consistent with this interpretation: in our benchmarks, architectural choices can matter as much as, or more than, generic pre-training, and the benefits of pre-training are not uniform across datasets. At the same time, the strong performance of Mantis—even without pre-training—suggests that lightweight architectures provide a promising starting point for EEG decoding.

Limitations and future work. This study has several limitations. First, we evaluate only two generic TS FMs, so it remains unclear how broadly our conclusions extend to other general-purpose time-series models such as TimesFM, Chronos, Timer, or UniTS. Second, our analysis is primarily performance-based and does not directly compare the representations learned by generic TS FMs and EEG-specific FMs. A more detailed representational analysis could help clarify whether the observed performance gap reflects differences in the features encoded before adaptation or differences in how readily those features can be adapted to EEG tasks.

Several directions follow from this. An immediate next step is to expand the benchmark to a broader set of generic TS FMs and EEG datasets, including settings with stronger domain shift and lower data availability. It would also be valuable to characterize learned representations more directly, for example through probing analyses, cross-model similarity measures, or frequency-aware evaluations that test sensitivity to EEG-relevant signal structure. Another promising direction is to investigate intermediate adaptation strategies, such as continued pre-training on unlabeled EEG, which may help bridge the gap between generic pre-training and domain-specific decoding. More broadly, an important goal is to develop stronger EEG representations that can consistently outperform compact task-specific models such as CTNet, rather than merely approaching their performance with substantially larger models.

ACKNOWLEDGMENTS

*L.Y. is supported by the Research Foundation – Flanders (FWO) grant 1S65622N.
*Q.S. is supported by the China Scholarship Council (no. 202206050022).
*M.M.V.H. is supported by research grants received from Horizon Europe's Marie Sklodowska-Curie Action (grant agreement No. 101118964), Horizon 2020 research and innovation programme under grant agreement No. 857375, the special research fund of the KU Leuven (C24/18/098), the Belgian Fund for Scientific Research – Flanders (G0A4321N, G0C1522N, G031426N), and the Hercules Foundation (AKUL 043).
*The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. In *Advances in Neural Information Processing Systems*, 2024.

Jingjing Chen, Xiaobin Wang, Chen Huang, Xin Hu, Xinke Shen, and Dan Zhang. A large finer-grained affective computing EEG dataset. *Sci. Data*, 10(1):740, October 2023.

Abhimanyu Das, Weihao Kong, Andrew Leber, Rajat Mathews, and Ravikumar Sen. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Vasilii Feofanov, Songkang Wen, Marius Alonso, Romain Ilbert, Hongbo Guo, Malik Tiomoko, Lujia Pan, Jianfeng Zhang, and Ievgen Redko. Mantis: Lightweight calibrated foundation model for user-friendly time series classification, 2025. URL https://arxiv.org/abs/2502.15637.

Shanghua Gao, Teddy Koker, Owen Queen, Thomas Hartvigsen, Theodoros Tsiligkaridis, and Marinka Zitnik. Units: A unified multi-task time series model. In *Advances in Neural Information Processing Systems*, 2024.

Théo Gnassounou, Yessin Moakher, Shifeng Xie, Vasilii Feofanov, and Ievgen Redko. Leveraging generic time series foundation models for eeg classification, 2025. URL https://arxiv.org/abs/2510.27522.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models, 2024. URL https://arxiv.org/abs/2402.03885.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci, 2024. URL https://arxiv.org/abs/2405.18765.

Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. Bendr: using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data, 2021. URL https://arxiv.org/abs/2101.12037.

Niklas Kueper, Kartik Chari, Judith Bütefür, Julia Habenicht, Tobias Rossol, Su Kyoung Kim, Marc Tabie, Frank Kirchner, and Elsa Andrea Kirchner. EEG and EMG dataset for the detection of errors introduced by an active orthosis device. *Front. Hum. Neurosci.*, 18:1304311, January 2024.

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

Dingkun Liu, Yuheng Chen, Zhu Chen, Zhenyao Cui, Yaozhi Wen, Jiayu An, Jingwei Luo, and Dongrui Wu. Eeg foundation models: Progresses, benchmarking, and open problems, 2026. URL https://arxiv.org/abs/2601.17883.

Iyad Obeid and Joseph Picone. The temple university hospital EEG data corpus. *Front. Neurosci.*, 10:196, May 2016.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp.*, 38(11):5391–5420, November 2017.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 31:710–719, February 2023.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J Miller, Gernot R Müller-Putz, Guido Nolte, Gert Pfurtscheller, Hubert Preissl, Gerwin Schalk, Alois Schlögl, Carmen Vidaurre, Stephan Waldert, and Benjamin Blankertz. Review of the BCI competition IV. *Front. Neurosci.*, 6:55, July 2012.

Guagnyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. Eegpt: Pre-trained transformer for universal and reliable representation of eeg signals. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 39249–39280. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4540d267eeec4e5dbd9dae9448f0b739-Paper-Conference.pdf.

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, 2025. URL https://arxiv.org/abs/2412.07236.

Wei Xiong, Jiangtong Li, Jie Li, Kun Zhu, and Changjun Jiang. Eeg-fm-bench: A comprehensive benchmark for the systematic evaluation of eeg foundation models, 2026. URL https://arxiv.org/abs/2508.17742.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. Biot: Cross-data biosignal learning in the wild, 2023. URL https://arxiv.org/abs/2305.10351.

Liuyin Yang, Qiang Sun, Ang Li, and Marc M. Van Hulle. Are EEG foundation models worth it? comparative evaluation with traditional decoders in diverse BCI tasks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=5Xwm8e6vbh.

Wei Zhao, Xiaolu Jiang, Baocan Zhang, Shixiao Xiao, and Sujun Weng. CTNet: a convolutional transformer network for EEG-based motor imagery classification. *Sci. Rep.*, 14(1):20237, August 2024.

A Appendix

A.1 Data Pre-processing

We apply minimal, standardized pre-processing across all datasets to improve comparability while preserving the raw signal characteristics. We band-pass filter EEG signals between 0.1 and 128 Hz and apply a notch filter at the local power-line frequency using the `mne.filter` module. We then downsample all datasets to a common baseline of 256 Hz to match EEGPT (which operates natively at 256 Hz). For foundation models that require lower sampling rates, we additionally resample the data on the fly using `mne.resample` and then apply each model's default normalization. For the TS FMs, we follow ST-EEGFormer and use a sampling rate of 128 Hz.

A.2 Model training

All experiments were run on an HPC cluster with NVIDIA H100 GPUs. We consider two standard adaptation strategies for foundation models: linear probing and end-to-end fine-tuning. For linear probing, we freeze the pre-trained backbone and train only a lightweight classification head, which assesses the quality of the learned representations. For fine-tuning, we jointly update all parameters (backbone and head) to adapt the representation to the downstream dataset.

Unless otherwise stated, we follow the training protocol of the EEG benchmark study (Yang et al., 2026), using the AdamW optimizer, a cosine decay learning-rate schedule, and 10 warmup epochs. For Mantis, we adopt the fine-tuning strategy of (Gnassounou et al., 2025): we concatenate tokens across EEG channels and fine-tune with a small backbone learning rate (3
$times10^{-4}$) for up to 50 epochs. For MOMENT, we use a multichannel adapter as recommended for multivariate time series and train with the same overall procedure as Mantis.

A.3 Benchmark Results

Table 2: FACED dataset results (all metrics as fractions; mean ± standard deviation). Best results are shown in bold.

Model	Training	Kappa	Acc1	Balanced Acc	AUC
Classic neural networks					
CTNet	Scratch	0.206 ± 0.075	0.313 ± 0.075	0.294 ± 0.067	0.732 ± 0.051
DeepConvNet	Scratch	0.239 ± 0.100	0.340 ± 0.093	0.325 ± 0.089	0.769 ± 0.063
EEGNet	Scratch	0.216 ± 0.062	0.305 ± 0.058	0.301 ± 0.053	0.732 ± 0.052
Conformer	Scratch	$\mathbf{0.328 \pm 0.094}$	$\mathbf{0.416 \pm 0.090}$	$\mathbf{0.405 \pm 0.085}$	$\mathbf{0.771 \pm 0.061}$
EEG foundation models					
BENDR	Fine-tune	0.242 ± 0.083	0.333 ± 0.082	0.326 ± 0.073	0.728 ± 0.056
BENDR	Linear Probe	0.079 ± 0.038	0.187 ± 0.034	0.181 ± 0.032	0.596 ± 0.028
BIOT	Fine-tune	0.078 ± 0.045	0.186 ± 0.042	0.181 ± 0.041	0.580 ± 0.029
BIOT	Linear Probe	0.075 ± 0.038	0.193 ± 0.046	0.178 ± 0.032	0.612 ± 0.044
CBraMod	Fine-tune	0.282 ± 0.094	0.361 ± 0.087	0.363 ± 0.084	0.751 ± 0.060
CBraMod	Linear Probe	0.110 ± 0.038	0.205 ± 0.040	0.212 ± 0.035	0.654 ± 0.040
EEGPT	Fine-tune	0.100 ± 0.046	0.206 ± 0.042	0.200 ± 0.040	0.627 ± 0.039
EEGPT	Linear Probe	0.185 ± 0.061	0.279 ± 0.056	0.275 ± 0.055	0.707 ± 0.047
LaBraM	Fine-tune	0.143 ± 0.051	0.233 ± 0.045	0.239 ± 0.045	0.645 ± 0.044
LaBraM	Linear Probe	0.031 ± 0.034	0.153 ± 0.043	0.137 ± 0.029	0.582 ± 0.034
ST-EEGFormer-s	Fine-tune	0.170 ± 0.075	0.269 ± 0.073	0.264 ± 0.067	0.675 ± 0.058
ST-EEGFormer-s	Linear Probe	0.063 ± 0.035	0.172 ± 0.044	0.163 ± 0.029	0.641 ± 0.044
ST-EEGFormer-b	Fine-tune	0.132 ± 0.073	0.227 ± 0.066	0.230 ± 0.065	0.639 ± 0.057
ST-EEGFormer-b	Linear Probe	0.062 ± 0.052	0.170 ± 0.059	0.163 ± 0.044	0.636 ± 0.063
ST-EEGFormer-l	Fine-tune	0.219 ± 0.088	0.303 ± 0.075	0.306 ± 0.079	0.691 ± 0.057
ST-EEGFormer-l	Linear Probe	0.078 ± 0.055	0.192 ± 0.054	0.178 ± 0.048	0.636 ± 0.054
Time-series foundation models					
Mantis	Fine-tune	0.223 ± 0.084	0.301 ± 0.074	0.311 ± 0.076	0.725 ± 0.063
Mantis	Linear Probe	0.135 ± 0.052	0.236 ± 0.052	0.233 ± 0.047	0.659 ± 0.055
MOMENT	Fine-tune	0.128 ± 0.059	0.210 ± 0.046	0.225 ± 0.053	0.654 ± 0.052
MOMENT	Linear Probe	0.050 ± 0.037	0.162 ± 0.043	0.153 ± 0.032	0.591 ± 0.045

Table 3: TUEV dataset results (all metrics as fractions; mean ± standard deviation). Best results are shown in bold.

Model	Training	Kappa	Acc1	Balanced Acc	AUC
Classic neural networks					
CTNet	Scratch	0.351 ± 0.075	0.632 ± 0.345	0.635 ± 0.233	0.776 ± 0.197
DeepConvNet	Scratch	0.333 ± 0.297	0.608 ± 0.393	0.568 ± 0.194	0.756 ± 0.215
EEGNet	Scratch	0.317 ± 0.239	0.560 ± 0.390	0.544 ± 0.195	0.734 ± 0.222
Conformer	Scratch	0.371 ± 0.191	0.616 ± 0.388	0.627 ± 0.217	0.749 ± 0.225
EEG foundation models					
BENDR	Fine-tune	0.279 ± 0.364	0.645 ± 0.374	0.611 ± 0.217	0.741 ± 0.222
BENDR	Linear Probe	0.033 ± 0.116	0.498 ± 0.300	0.400 ± 0.142	0.613 ± 0.190
BIOT	Fine-tune	0.265 ± 0.354	0.636 ± 0.360	0.553 ± 0.222	0.715 ± 0.240
BIOT	Linear Probe	0.259 ± 0.346	0.638 ± 0.344	0.560 ± 0.215	0.739 ± 0.226
CBraMod	Fine-tune	0.302 ± 0.370	0.643 ± 0.377	0.624 ± 0.188	0.751 ± 0.241
CBraMod	Linear Probe	0.092 ± 0.238	0.473 ± 0.386	0.435 ± 0.150	0.630 ± 0.254
EEGPT	Fine-tune	0.357 ± 0.374	0.696 ± 0.331	0.641 ± 0.203	0.767 ± 0.243
EEGPT	Linear Probe	0.363 ± 0.375	0.676 ± 0.373	0.658 ± 0.182	$\mathbf{0.773 \pm 0.239}$
LaBraM	Fine-tune	0.381 ± 0.389	0.725 ± 0.333	0.670 ± 0.203	0.759 ± 0.219
LaBraM	Linear Probe	0.125 ± 0.256	0.627 ± 0.380	0.461 ± 0.171	0.725 ± 0.221
ST-EEGFormer-s	Fine-tune	0.418 ± 0.449	$\mathbf{0.765 \pm 0.304}$	$\mathbf{0.700 \pm 0.248}$	0.756 ± 0.237
ST-EEGFormer-s	Linear Probe	0.107 ± 0.258	0.645 ± 0.404	0.456 ± 0.166	0.749 ± 0.207
ST-EEGFormer-b	Fine-tune	0.413 ± 0.414	0.745 ± 0.324	0.687 ± 0.250	0.750 ± 0.235
ST-EEGFormer-b	Linear Probe	0.284 ± 0.284	0.645 ± 0.396	0.466 ± 0.181	0.751 ± 0.216
ST-EEGFormer-l	Fine-tune	0.398 ± 0.303	0.738 ± 0.331	0.690 ± 0.240	0.750 ± 0.243
ST-EEGFormer-l	Linear Probe	0.273 ± 0.273	0.643 ± 0.398	0.473 ± 0.170	0.741 ± 0.217
Time-series foundation models					
Mantis	Fine-tune	$\mathbf{0.424 \pm 0.423}$	0.682 ± 0.356	0.667 ± 0.240	0.732 ± 0.256
Mantis	Linear Probe	0.331 ± 0.331	0.622 ± 0.377	0.561 ± 0.225	0.722 ± 0.242
MOMENT	Fine-tune	0.286 ± 0.347	0.693 ± 0.323	0.589 ± 0.227	0.745 ± 0.229
MOMENT	Linear Probe	0.079 ± 0.225	0.637 ± 0.385	0.436 ± 0.172	0.670 ± 0.265

Table 4: BCI-IV-2A population results (all metrics as fractions; mean ± standard deviation). Best results are shown in bold.

Model	Training	Kappa	Acc1	Balanced Acc	AUC
Classic neural networks					
DeepConvNet	Scratch	0.620 ± 0.118	0.715 ± 0.089	0.715 ± 0.089	0.909 ± 0.051
EEGNet	Scratch	0.614 ± 0.104	0.711 ± 0.078	0.711 ± 0.078	0.911 ± 0.045
Conformer	Scratch	0.423 ± 0.098	0.567 ± 0.073	0.567 ± 0.073	0.806 ± 0.054
CTNet	Scratch	0.665 ± 0.127	0.749 ± 0.096	0.749 ± 0.096	$\mathbf{0.929 \pm 0.045}$
EEG foundation models					
BIOT	Fine-tune	0.255 ± 0.128	0.441 ± 0.096	0.441 ± 0.096	0.687 ± 0.085
BIOT	Linear Probe	0.240 ± 0.171	0.430 ± 0.128	0.430 ± 0.128	0.692 ± 0.116
BENDR	Fine-tune	0.287 ± 0.076	0.465 ± 0.057	0.465 ± 0.057	0.725 ± 0.053
BENDR	Linear Probe	0.095 ± 0.035	0.321 ± 0.026	0.321 ± 0.026	0.573 ± 0.019
CBraMod	Fine-tune	0.490 ± 0.102	0.618 ± 0.077	0.617 ± 0.077	0.851 ± 0.051
CBraMod	Linear Probe	0.092 ± 0.044	0.319 ± 0.033	0.319 ± 0.033	0.634 ± 0.049
EEGPT	Fine-tune	0.203 ± 0.061	0.402 ± 0.046	0.402 ± 0.046	0.674 ± 0.042
EEGPT	Linear Probe	0.475 ± 0.104	0.606 ± 0.078	0.606 ± 0.078	0.841 ± 0.054
LaBraM	Fine-tune	0.226 ± 0.085	0.419 ± 0.064	0.419 ± 0.064	0.677 ± 0.050
LaBraM	Linear Probe	0.048 ± 0.056	0.286 ± 0.041	0.286 ± 0.042	0.572 ± 0.056
ST-EEGFormer-s	Fine-tune	0.586 ± 0.154	0.689 ± 0.116	0.689 ± 0.116	0.863 ± 0.069
ST-EEGFormer-s	Linear Probe	0.162 ± 0.108	0.371 ± 0.081	0.371 ± 0.081	0.690 ± 0.098
ST-EEGFormer-b	Fine-tune	0.700 ± 0.131	0.775 ± 0.098	0.775 ± 0.098	0.908 ± 0.058
ST-EEGFormer-b	Linear Probe	0.215 ± 0.128	0.412 ± 0.096	0.411 ± 0.096	0.713 ± 0.092
ST-EEGFormer-l	Fine-tune	$\mathbf{0.728 \pm 0.147}$	$\mathbf{0.796 \pm 0.110}$	$\mathbf{0.796 \pm 0.110}$	0.928 ± 0.052
ST-EEGFormer-l	Linear Probe	0.188 ± 0.146	0.390 ± 0.110	0.391 ± 0.109	0.716 ± 0.103
Time-series foundation models					
MOMENT	Fine-tune	0.096 ± 0.063	0.322 ± 0.048	0.322 ± 0.048	0.588 ± 0.047
MOMENT	Linear Probe	0.070 ± 0.059	0.304 ± 0.044	0.303 ± 0.044	0.586 ± 0.054
Mantis	Fine-tune	0.307 ± 0.087	0.480 ± 0.065	0.480 ± 0.065	0.749 ± 0.063
Mantis	Linear Probe	0.246 ± 0.077	0.434 ± 0.057	0.434 ± 0.058	0.695 ± 0.054

Table 5: BCI-IV-2A leave-one-out zero-shot results (all metrics as fractions; mean ± standard deviation). Best results are shown in bold.

Model	Training	Kappa	Acc1	Balanced Acc	AUC
Classic neural networks					
DeepConvNet	Scratch	0.418 ± 0.108	0.563 ± 0.081	0.563 ± 0.081	0.838 ± 0.069
EEGNet	Scratch	0.454 ± 0.129	0.591 ± 0.097	0.591 ± 0.097	0.858 ± 0.063
Conformer	Scratch	0.308 ± 0.092	0.481 ± 0.069	0.481 ± 0.069	0.738 ± 0.062
CTNet	Scratch	$\mathbf{0.489 \pm 0.150}$	$\mathbf{0.617 \pm 0.112}$	$\mathbf{0.617 \pm 0.112}$	$\mathbf{0.876 \pm 0.068}$
EEG foundation models					
BIOT	Fine-tune	0.084 ± 0.094	0.313 ± 0.071	0.313 ± 0.071	0.578 ± 0.064
BIOT	Linear Probe	0.086 ± 0.071	0.315 ± 0.053	0.315 ± 0.053	0.597 ± 0.071
BENDR	Fine-tune	0.217 ± 0.063	0.412 ± 0.048	0.412 ± 0.047	0.676 ± 0.046
BENDR	Linear Probe	0.065 ± 0.026	0.299 ± 0.021	0.299 ± 0.020	0.555 ± 0.024
CBraMod	Fine-tune	0.359 ± 0.097	0.519 ± 0.073	0.519 ± 0.073	0.775 ± 0.062
CBraMod	Linear Probe	0.114 ± 0.039	0.335 ± 0.030	0.335 ± 0.029	0.623 ± 0.050
EEGPT	Fine-tune	0.129 ± 0.049	0.347 ± 0.037	0.347 ± 0.037	0.618 ± 0.035
EEGPT	Linear Probe	0.330 ± 0.137	0.498 ± 0.103	0.498 ± 0.103	0.774 ± 0.064
LaBraM	Fine-tune	0.081 ± 0.056	0.311 ± 0.042	0.311 ± 0.042	0.573 ± 0.040
LaBraM	Linear Probe	0.040 ± 0.044	0.279 ± 0.031	0.280 ± 0.033	0.557 ± 0.050
ST-EEGFormer-s	Fine-tune	0.349 ± 0.166	0.512 ± 0.125	0.512 ± 0.125	0.746 ± 0.083
ST-EEGFormer-s	Linear Probe	0.031 ± 0.074	0.273 ± 0.055	0.273 ± 0.055	0.642 ± 0.087
ST-EEGFormer-b	Fine-tune	0.377 ± 0.151	0.533 ± 0.114	0.533 ± 0.114	0.780 ± 0.085
ST-EEGFormer-b	Linear Probe	0.043 ± 0.089	0.282 ± 0.066	0.282 ± 0.066	0.660 ± 0.095
ST-EEGFormer-l	Fine-tune	0.434 ± 0.197	0.575 ± 0.148	0.575 ± 0.148	0.800 ± 0.103
ST-EEGFormer-l	Linear Probe	0.074 ± 0.104	0.305 ± 0.078	0.306 ± 0.078	0.672 ± 0.105
Time-series foundation models					
MOMENT	Fine-tune	0.073 ± 0.052	0.304 ± 0.039	0.304 ± 0.039	0.557 ± 0.034
MOMENT	Linear Probe	0.047 ± 0.045	0.285 ± 0.034	0.285 ± 0.034	0.571 ± 0.055
Mantis	Fine-tune	0.212 ± 0.075	0.409 ± 0.056	0.409 ± 0.056	0.688 ± 0.048
Mantis	Linear Probe	0.181 ± 0.078	0.385 ± 0.059	0.386 ± 0.058	0.653 ± 0.044

Table 6: BCI-IV-2A leave-one-out fine-tuning results (all metrics as fractions; mean ± standard deviation). Best results are shown in bold.

Model	Training	Kappa	Acc1	Balanced Acc	AUC
Classic neural networks					
DeepConvNet	Scratch	0.665 ± 0.118	0.748 ± 0.089	0.748 ± 0.089	0.920 ± 0.049
EEGNet	Scratch	0.681 ± 0.113	0.760 ± 0.085	0.760 ± 0.085	0.929 ± 0.043
Conformer	Scratch	0.501 ± 0.091	0.626 ± 0.068	0.626 ± 0.068	0.846 ± 0.052
CTNet	Scratch	0.725 ± 0.136	0.794 ± 0.102	0.794 ± 0.102	$\mathbf{0.947 \pm 0.040}$
EEG foundation models					
BIOT	Fine-tune	0.254 ± 0.165	0.440 ± 0.124	0.440 ± 0.124	0.689 ± 0.098
BIOT	Linear Probe	0.306 ± 0.149	0.480 ± 0.111	0.480 ± 0.111	0.736 ± 0.100
BENDR	Fine-tune	0.347 ± 0.119	0.510 ± 0.089	0.510 ± 0.089	0.769 ± 0.064
BENDR	Linear Probe	0.101 ± 0.032	0.326 ± 0.024	0.326 ± 0.024	0.589 ± 0.020
CBraMod	Fine-tune	0.528 ± 0.105	0.646 ± 0.079	0.646 ± 0.079	0.871 ± 0.052
CBraMod	Linear Probe	0.131 ± 0.047	0.348 ± 0.035	0.348 ± 0.035	0.634 ± 0.048
EEGPT	Fine-tune	0.245 ± 0.095	0.434 ± 0.071	0.434 ± 0.071	0.698 ± 0.054
EEGPT	Linear Probe	0.428 ± 0.121	0.571 ± 0.091	0.571 ± 0.091	0.815 ± 0.067
LaBraM	Fine-tune	0.164 ± 0.075	0.373 ± 0.056	0.373 ± 0.056	0.617 ± 0.061
LaBraM	Linear Probe	0.085 ± 0.071	0.313 ± 0.053	0.314 ± 0.053	0.581 ± 0.061
ST-EEGFormer-s	Fine-tune	0.634 ± 0.130	0.725 ± 0.098	0.725 ± 0.098	0.896 ± 0.060
ST-EEGFormer-s	Linear Probe	0.186 ± 0.117	0.390 ± 0.088	0.390 ± 0.088	0.713 ± 0.092
ST-EEGFormer-b	Fine-tune	0.707 ± 0.119	0.780 ± 0.089	0.780 ± 0.089	0.929 ± 0.049
ST-EEGFormer-b	Linear Probe	0.218 ± 0.102	0.413 ± 0.077	0.414 ± 0.077	0.714 ± 0.089
ST-EEGFormer-l	Fine-tune	$\mathbf{0.744 \pm 0.143}$	$\mathbf{0.808 \pm 0.108}$	$\mathbf{0.808 \pm 0.108}$	0.937 ± 0.057
ST-EEGFormer-l	Linear Probe	0.161 ± 0.117	0.371 ± 0.088	0.371 ± 0.088	0.713 ± 0.099
Time-series foundation models					
MOMENT	Fine-tune	0.152 ± 0.051	0.363 ± 0.039	0.364 ± 0.038	0.627 ± 0.038
MOMENT	Linear Probe	0.083 ± 0.059	0.312 ± 0.044	0.312 ± 0.044	0.592 ± 0.057
Mantis	Fine-tune	0.355 ± 0.122	0.517 ± 0.091	0.517 ± 0.091	0.768 ± 0.071
Mantis	Linear Probe	0.210 ± 0.079	0.408 ± 0.059	0.408 ± 0.059	0.675 ± 0.051

Table 7: BCI-IV-2A leave-one-out generalization drop after fine-tuning on seen subjects (all metrics as fractions; mean ± standard deviation). Smaller is better; best (smallest) results are shown in bold.

Model	Training	Kappa	Acc1	Balanced Acc	AUC
Classic neural networks					
DeepConvNet	Scratch	0.256 ± 0.048	0.192 ± 0.036	0.192 ± 0.036	0.098 ± 0.020
EEGNet	Scratch	0.245 ± 0.034	0.183 ± 0.025	0.183 ± 0.025	0.086 ± 0.011
Conformer	Scratch	0.097 ± 0.021	0.073 ± 0.016	0.073 ± 0.016	0.047 ± 0.013
CTNet	Scratch	0.323 ± 0.071	0.242 ± 0.054	0.242 ± 0.054	0.113 ± 0.019
EEG foundation models					
BIOT	Fine-tune	0.089 ± 0.050	0.067 ± 0.037	0.067 ± 0.037	0.050 ± 0.022
BIOT	Linear Probe	0.129 ± 0.063	0.097 ± 0.047	0.097 ± 0.047	0.068 ± 0.034
BENDR	Fine-tune	0.068 ± 0.029	0.051 ± 0.022	0.051 ± 0.022	0.047 ± 0.014
BENDR	Linear Probe	$\mathbf{0.007 \pm 0.013}$	$\mathbf{0.005 \pm 0.009}$	$\mathbf{0.005 \pm 0.009}$	0.005 ± 0.007
CBraMod	Fine-tune	0.103 ± 0.029	0.077 ± 0.022	0.077 ± 0.022	0.054 ± 0.011
CBraMod	Linear Probe	0.043 ± 0.029	0.033 ± 0.022	0.032 ± 0.022	0.043 ± 0.013
EEGPT	Fine-tune	0.031 ± 0.029	0.023 ± 0.022	0.023 ± 0.022	0.018 ± 0.017
EEGPT	Linear Probe	0.267 ± 0.048	0.200 ± 0.036	0.200 ± 0.036	0.146 ± 0.038
LaBraM	Fine-tune	0.012 ± 0.016	0.009 ± 0.012	0.009 ± 0.012	0.003 ± 0.008
LaBraM	Linear Probe	0.017 ± 0.019	0.013 ± 0.014	0.013 ± 0.015	0.021 ± 0.013
ST-EEGFormer-s	Fine-tune	0.176 ± 0.050	0.132 ± 0.038	0.132 ± 0.037	0.063 ± 0.009
ST-EEGFormer-s	Linear Probe	0.094 ± 0.042	0.070 ± 0.032	0.070 ± 0.032	0.021 ± 0.016
ST-EEGFormer-b	Fine-tune	0.145 ± 0.034	0.108 ± 0.025	0.108 ± 0.025	0.048 ± 0.012
ST-EEGFormer-b	Linear Probe	0.118 ± 0.053	0.088 ± 0.039	0.088 ± 0.039	0.019 ± 0.015
ST-EEGFormer-l	Fine-tune	0.146 ± 0.035	0.109 ± 0.026	0.109 ± 0.026	0.044 ± 0.017
ST-EEGFormer-l	Linear Probe	0.115 ± 0.051	0.086 ± 0.038	0.086 ± 0.038	0.020 ± 0.015
Time-series foundation models					
MOMENT	Fine-tune	0.020 ± 0.024	0.015 ± 0.018	0.015 ± 0.018	0.008 ± 0.016
MOMENT	Linear Probe	0.022 ± 0.023	0.017 ± 0.018	0.017 ± 0.017	$\mathbf{0.000 \pm 0.003}$
Mantis	Fine-tune	0.064 ± 0.046	0.048 ± 0.034	0.048 ± 0.034	0.027 ± 0.011
Mantis	Linear Probe	0.097 ± 0.020	0.073 ± 0.015	0.073 ± 0.015	0.062 ± 0.015

Table 8: Error population results (all metrics as fractions; mean ± standard deviation). Best results are shown in bold.

Model	Training	Kappa	Acc1	Balanced Acc	AUC
Classic neural networks					
DeepConvNet	Scratch	0.979 ± 0.013	0.993 ± 0.004	0.990 ± 0.009	0.998 ± 0.004
EEGNet	Scratch	0.976 ± 0.021	0.993 ± 0.007	0.986 ± 0.015	$\mathbf{0.999 \pm 0.003}$
Conformer	Scratch	0.974 ± 0.021	0.992 ± 0.006	0.985 ± 0.014	0.997 ± 0.007
CTNet	Scratch	0.981 ± 0.020	0.994 ± 0.006	0.988 ± 0.014	0.998 ± 0.002
EEG foundation models					
BIOT	Fine-tune	0.944 ± 0.028	0.983 ± 0.009	0.964 ± 0.016	0.992 ± 0.009
BIOT	Linear Probe	0.880 ± 0.036	0.962 ± 0.011	0.937 ± 0.021	0.979 ± 0.012
BENDR	Fine-tune	0.976 ± 0.012	0.992 ± 0.004	0.985 ± 0.010	0.998 ± 0.002
BENDR	Linear Probe	0.939 ± 0.032	0.981 ± 0.010	0.963 ± 0.019	0.988 ± 0.009
CBraMod	Fine-tune	0.976 ± 0.021	0.992 ± 0.007	0.988 ± 0.012	0.996 ± 0.006
CBraMod	Linear Probe	0.898 ± 0.033	0.967 ± 0.011	0.951 ± 0.015	0.987 ± 0.008
EEGPT	Fine-tune	0.967 ± 0.016	0.990 ± 0.005	0.975 ± 0.012	$\mathbf{0.999 \pm 0.001}$
EEGPT	Linear Probe	0.982 ± 0.014	0.994 ± 0.004	0.987 ± 0.009	$\mathbf{0.999 \pm 0.002}$
LaBraM	Fine-tune	0.979 ± 0.015	0.993 ± 0.005	0.986 ± 0.011	$\mathbf{0.999 \pm 0.003}$
LaBraM	Linear Probe	0.197 ± 0.101	0.823 ± 0.016	0.570 ± 0.039	0.859 ± 0.058
ST-EEGFormer-s	Fine-tune	$\mathbf{0.988 \pm 0.015}$	$\mathbf{0.996 \pm 0.005}$	0.991 ± 0.011	0.998 ± 0.005
ST-EEGFormer-s	Linear Probe	0.625 ± 0.134	0.898 ± 0.030	0.778 ± 0.079	0.951 ± 0.024
ST-EEGFormer-b	Fine-tune	0.979 ± 0.017	0.994 ± 0.005	0.985 ± 0.012	0.993 ± 0.011
ST-EEGFormer-b	Linear Probe	0.410 ± 0.159	0.852 ± 0.035	0.667 ± 0.077	0.929 ± 0.037
ST-EEGFormer-l	Fine-tune	0.987 ± 0.015	$\mathbf{0.996 \pm 0.004}$	$\mathbf{0.992 \pm 0.010}$	0.998 ± 0.005
ST-EEGFormer-l	Linear Probe	0.703 ± 0.120	0.918 ± 0.027	0.817 ± 0.071	0.965 ± 0.020
Time-series foundation models					
MOMENT	Fine-tune	0.949 ± 0.028	0.985 ± 0.008	0.971 ± 0.017	0.997 ± 0.002
MOMENT	Linear Probe	0.488 ± 0.191	0.866 ± 0.041	0.710 ± 0.093	0.931 ± 0.040
Mantis	Fine-tune	0.971 ± 0.022	0.991 ± 0.007	0.979 ± 0.016	0.997 ± 0.005
Mantis	Linear Probe	0.932 ± 0.030	0.979 ± 0.009	0.966 ± 0.014	0.992 ± 0.007

Table 9: Error leave-one-out zero-shot results (all metrics as fractions; mean ± standard deviation). Best results are shown in bold.

Model	Training	Kappa	Acc1	Balanced Acc	AUC
Classic neural networks					
DeepConvNet	Scratch	0.824 ± 0.104	0.946 ± 0.029	0.905 ± 0.068	0.964 ± 0.040
EEGNet	Scratch	0.764 ± 0.114	0.932 ± 0.028	0.863 ± 0.082	0.961 ± 0.037
Conformer	Scratch	0.791 ± 0.107	0.938 ± 0.030	0.873 ± 0.066	0.961 ± 0.035
CTNet	Scratch	0.836 ± 0.077	0.950 ± 0.021	0.905 ± 0.056	$\mathbf{0.973 \pm 0.031}$
EEG foundation models					
BIOT	Fine-tune	0.539 ± 0.112	0.862 ± 0.027	0.761 ± 0.079	0.857 ± 0.066
BIOT	Linear Probe	0.400 ± 0.122	0.817 ± 0.049	0.693 ± 0.066	0.820 ± 0.062
BENDR	Fine-tune	0.737 ± 0.117	0.922 ± 0.032	0.851 ± 0.071	0.934 ± 0.046
BENDR	Linear Probe	0.536 ± 0.164	0.868 ± 0.042	0.747 ± 0.090	0.855 ± 0.068
CBraMod	Fine-tune	0.733 ± 0.171	0.925 ± 0.037	0.848 ± 0.101	0.943 ± 0.060
CBraMod	Linear Probe	0.533 ± 0.162	0.855 ± 0.061	0.753 ± 0.077	0.877 ± 0.069
EEGPT	Fine-tune	0.709 ± 0.152	0.918 ± 0.037	0.826 ± 0.080	0.935 ± 0.067
EEGPT	Linear Probe	0.767 ± 0.114	0.933 ± 0.028	0.860 ± 0.069	0.950 ± 0.051
LaBraM	Fine-tune	0.767 ± 0.141	0.933 ± 0.035	0.861 ± 0.086	0.959 ± 0.036
LaBraM	Linear Probe	0.160 ± 0.094	0.817 ± 0.016	0.556 ± 0.035	0.831 ± 0.070
ST-EEGFormer-s	Fine-tune	$\mathbf{0.837 \pm 0.104}$	$\mathbf{0.952 \pm 0.028}$	$\mathbf{0.907 \pm 0.073}$	0.967 ± 0.023
ST-EEGFormer-s	Linear Probe	0.550 ± 0.165	0.877 ± 0.037	0.750 ± 0.099	0.926 ± 0.046
ST-EEGFormer-b	Fine-tune	0.774 ± 0.135	0.937 ± 0.033	0.862 ± 0.085	0.958 ± 0.033
ST-EEGFormer-b	Linear Probe	0.323 ± 0.192	0.837 ± 0.038	0.633 ± 0.097	0.909 ± 0.045
ST-EEGFormer-l	Fine-tune	0.811 ± 0.156	0.948 ± 0.040	0.886 ± 0.101	0.964 ± 0.039
ST-EEGFormer-l	Linear Probe	0.581 ± 0.193	0.892 ± 0.044	0.756 ± 0.109	0.946 ± 0.033
Time-series foundation models					
MOMENT	Fine-tune	0.647 ± 0.122	0.894 ± 0.034	0.814 ± 0.072	0.926 ± 0.042
MOMENT	Linear Probe	0.427 ± 0.213	0.853 ± 0.044	0.682 ± 0.100	0.920 ± 0.050
Mantis	Fine-tune	0.717 ± 0.131	0.921 ± 0.030	0.834 ± 0.086	0.952 ± 0.034
Mantis	Linear Probe	0.503 ± 0.103	0.854 ± 0.024	0.743 ± 0.075	0.839 ± 0.063

Table 10: Error leave-one-out fine-tuning results (all metrics as fractions; mean ± standard deviation). Best results are shown in bold.

Model	Training	Kappa	Acc1	Balanced Acc	AUC
Classic neural networks					
DeepConvNet	Scratch	0.990 ± 0.013	0.997 ± 0.004	$\mathbf{0.996 \pm 0.006}$	$\mathbf{1.000 \pm 0.000}$
EEGNet	Scratch	$\mathbf{0.991 \pm 0.006}$	$\mathbf{0.997 \pm 0.002}$	0.994 ± 0.005	1.000 ± 0.001
Conformer	Scratch	0.989 ± 0.011	0.997 ± 0.003	0.993 ± 0.009	0.999 ± 0.004
CTNet	Scratch	0.989 ± 0.017	0.997 ± 0.005	0.994 ± 0.011	0.999 ± 0.002
EEG foundation models					
BIOT	Fine-tune	0.955 ± 0.029	0.986 ± 0.009	0.972 ± 0.021	0.991 ± 0.010
BIOT	Linear Probe	0.925 ± 0.021	0.976 ± 0.007	0.956 ± 0.014	0.993 ± 0.004
BENDR	Fine-tune	0.979 ± 0.014	0.993 ± 0.004	0.987 ± 0.009	$\mathbf{1.000 \pm 0.000}$
BENDR	Linear Probe	0.916 ± 0.037	0.973 ± 0.012	0.957 ± 0.015	0.995 ± 0.005
CBraMod	Fine-tune	0.984 ± 0.015	0.995 ± 0.005	0.989 ± 0.012	0.998 ± 0.004
CBraMod	Linear Probe	0.933 ± 0.026	0.978 ± 0.008	0.965 ± 0.011	0.992 ± 0.009
EEGPT	Fine-tune	0.968 ± 0.018	0.990 ± 0.006	0.979 ± 0.011	0.999 ± 0.002
EEGPT	Linear Probe	0.986 ± 0.012	0.995 ± 0.004	0.989 ± 0.008	$\mathbf{1.000 \pm 0.000}$
LaBraM	Fine-tune	0.982 ± 0.013	0.994 ± 0.004	0.987 ± 0.011	0.998 ± 0.004
LaBraM	Linear Probe	0.368 ± 0.158	0.851 ± 0.027	0.643 ± 0.073	0.905 ± 0.044
ST-EEGFormer-s	Fine-tune	0.983 ± 0.019	0.995 ± 0.006	0.990 ± 0.013	0.997 ± 0.005
ST-EEGFormer-s	Linear Probe	0.693 ± 0.164	0.916 ± 0.034	0.821 ± 0.097	0.967 ± 0.016
ST-EEGFormer-b	Fine-tune	0.980 ± 0.016	0.994 ± 0.005	0.985 ± 0.013	0.996 ± 0.006
ST-EEGFormer-b	Linear Probe	0.259 ± 0.168	0.834 ± 0.026	0.598 ± 0.071	0.930 ± 0.035
ST-EEGFormer-l	Fine-tune	0.985 ± 0.020	0.996 ± 0.006	0.990 ± 0.015	0.998 ± 0.003
ST-EEGFormer-l	Linear Probe	0.554 ± 0.163	0.887 ± 0.033	0.732 ± 0.087	0.963 ± 0.022
Time-series foundation models					
MOMENT	Fine-tune	0.955 ± 0.021	0.987 ± 0.006	0.974 ± 0.009	0.998 ± 0.001
MOMENT	Linear Probe	0.533 ± 0.202	0.879 ± 0.046	0.732 ± 0.102	0.938 ± 0.040
Mantis	Fine-tune	0.971 ± 0.022	0.991 ± 0.006	0.979 ± 0.015	0.998 ± 0.005
Mantis	Linear Probe	0.918 ± 0.041	0.975 ± 0.013	0.963 ± 0.016	0.980 ± 0.015

Table 11: Generalization drop on the Error dataset after fine-tuning on seen subjects (all metrics as fractions; mean ± standard deviation). Smaller is better; best (smallest) results are shown in bold.

Model	Training	Kappa	Acc1	Balanced Acc	AUC
Classic neural networks					
DeepConvNet	Scratch	0.168 ± 0.061	0.048 ± 0.015	0.101 ± 0.040	0.022 ± 0.010
EEGNet	Scratch	0.207 ± 0.103	0.055 ± 0.024	0.130 ± 0.063	0.018 ± 0.007
Conformer	Scratch	0.195 ± 0.049	0.055 ± 0.010	0.116 ± 0.040	0.027 ± 0.012
CTNet	Scratch	0.169 ± 0.095	0.047 ± 0.022	0.107 ± 0.061	0.019 ± 0.011
EEG foundation models					
BIOT	Fine-tune	0.184 ± 0.048	0.053 ± 0.010	0.101 ± 0.041	0.039 ± 0.024
BIOT	Linear Probe	0.410 ± 0.085	0.132 ± 0.048	0.196 ± 0.065	0.087 ± 0.021
BENDR	Fine-tune	0.307 ± 0.046	0.082 ± 0.010	0.187 ± 0.029	0.059 ± 0.020
BENDR	Linear Probe	0.224 ± 0.065	0.063 ± 0.017	0.135 ± 0.036	0.045 ± 0.016
CBraMod	Fine-tune	0.265 ± 0.146	0.069 ± 0.031	0.155 ± 0.093	0.028 ± 0.021
CBraMod	Linear Probe	0.189 ± 0.054	0.053 ± 0.018	0.123 ± 0.039	0.038 ± 0.015
EEGPT	Fine-tune	0.156 ± 0.091	0.043 ± 0.023	0.096 ± 0.056	0.015 ± 0.011
EEGPT	Linear Probe	0.165 ± 0.051	0.046 ± 0.013	0.109 ± 0.032	0.015 ± 0.004
LaBraM	Fine-tune	0.090 ± 0.105	0.025 ± 0.025	0.055 ± 0.067	0.004 ± 0.004
LaBraM	Linear Probe	-0.002 ± 0.099	0.006 ± 0.009	-0.006 ± 0.041	0.049 ± 0.030
ST-EEGFormer-s	Fine-tune	0.050 ± 0.080	0.014 ± 0.021	0.032 ± 0.053	0.002 ± 0.002
ST-EEGFormer-s	Linear Probe	0.086 ± 0.135	0.030 ± 0.024	0.017 ± 0.085	0.018 ± 0.004
ST-EEGFormer-b	Fine-tune	0.039 ± 0.037	0.012 ± 0.011	0.027 ± 0.025	0.005 ± 0.012
ST-EEGFormer-b	Linear Probe	0.085 ± 0.133	0.012 ± 0.017	0.032 ± 0.058	0.007 ± 0.004
ST-EEGFormer-l	Fine-tune	0.043 ± 0.042	0.013 ± 0.012	0.027 ± 0.031	0.004 ± 0.005
ST-EEGFormer-l	Linear Probe	0.124 ± 0.157	0.024 ± 0.029	0.059 ± 0.087	0.003 ± 0.004
Time-series foundation models					
MOMENT	Fine-tune	0.169 ± 0.152	0.044 ± 0.034	0.098 ± 0.096	0.019 ± 0.019
MOMENT	Linear Probe	$\mathbf{-0.022 \pm 0.117}$	$\mathbf{-0.003 \pm 0.018}$	$\mathbf{-0.020 \pm 0.069}$	$\mathbf{0.001 \pm 0.001}$
Mantis	Fine-tune	0.081 ± 0.060	0.024 ± 0.016	0.051 ± 0.044	0.003 ± 0.003
Mantis	Linear Probe	0.265 ± 0.059	0.073 ± 0.010	0.158 ± 0.048	0.052 ± 0.009

Table 12: Pre-trained vs. randomly initialized Mantis (fine-tuning only). We report balanced accuracy and Cohen's κ (mean $\pm$ std). $^{*}p<0.05$, $^{**}p<0.01$ (paired Wilcoxon signed-rank test). Bold indicates the significantly better result.

		Balanced Acc			Kappa		
Dataset	Setting	Pre-trained	Random Init	Sig.	Pre-trained	Random Init	Sig.
TUEV	Cross-sub (n=61)	0.667 ± 0.239	0.649 ± 0.231	n.s.	0.400 ± 0.420	0.382 ± 0.407	n.s.
FACED	Cross-sub (n=25)	0.311 ± 0.075	0.308 ± 0.074	n.s.	0.223 ± 0.083	0.220 ± 0.082	n.s.
BCI-IV-2A	Pop (n=9)	0.480 ± 0.069	$\mathbf{0.535 \pm 0.060}$	$**$	0.307 ± 0.091	$\mathbf{0.381 \pm 0.081}$	$**$
BCI-IV-2A	LOO ZS (n=9)	0.409 ± 0.056	0.462 ± 0.060	n.s.	0.212 ± 0.075	$\mathbf{0.282 \pm 0.081}$	$*$
BCI-IV-2A	LOO FT (n=9)	0.517 ± 0.091	0.510 ± 0.064	n.s.	0.355 ± 0.122	0.346 ± 0.086	n.s.
Error	Pop (n=8)	0.979 ± 0.017	0.964 ± 0.037	n.s.	$\mathbf{0.971 \pm 0.024}$	0.934 ± 0.060	$*$
Error	LOO ZS (n=8)	0.834 ± 0.086	$\mathbf{0.881 \pm 0.062}$	$*$	0.717 ± 0.131	0.778 ± 0.094	n.s.
Error	LOO FT (n=8)	0.979 ± 0.015	0.968 ± 0.021	n.s.	$\mathbf{0.971 \pm 0.022}$	0.936 ± 0.034	$**$