A Multi-dimensional Framework for Evaluating Generalization in EEG Foundation Models

Aditya Kommineni AKOMMINE@USC.EDU
Emily Zhou EMILYZHO@USC.EDU
Kleanthis Avramidis AVRAMIDI@USC.EDU
Tiantian Feng TIANTIAF@USC.EDU
Shrikanth Narayanan SHRI@USC.EDU
Signal Analysis and Interpretation Laboratory, University of Southern California, USA

Abstract

Evaluating foundation models under appropriate adaptation settings is essential for understanding the quality and transferability of the learned representations. Recent EEG foundation models have demonstrated promising transfer capabilities across tasks and datasets, motivating their growing use in neurotechnology and clinical applications. However, these models are typically evaluated under full fine-tuning on well-curated downstream datasets, a setting that does not reflect biomedical domain constraints such as limited labeled data, reduced sensor coverage, or parameter-efficient adaptation. In this work, we propose a multi-dimensional evaluation framework for assessing EEG models under realistic low-resource conditions. Empirical analysis of both supervised EEG models and recent EEG foundation models, including LaBraM, CSBrain, and CBraMod, across 6 different datasets is performed under the proposed multi-dimensional evaluation framework. We find that EEG foundation models consistently provide performance gains on long-context tasks such as sleep stage prediction and mental health state classification. In contrast, for short-window Brain Computer Interface style tasks, supervised models achieve comparable despite having substantially fewer parameters. Additional analyses demonstrate that current foundation models provide limited robustness to short-window tasks and channel constrained settings. Together, these findings motivate the use of multi-dimensional evaluation protocols that characterize model behavior under realistic use constraints.

Data and Code Availability In this work, we use openly available datasets, Physionet MI (Schalk et al., 2004; Goldberger et al., 2000), BCI Competition IV-2A (Brunner et al., 2008), Kaggle ERN (Mattout et al., 2014), TUEV (Obeid and Picone, 2016), Depression Classification (MDD MAL) (Mumtaz, 2016) and Sleep EDF datasets (Kemp et al., 2000). All datasets are accessible on public platforms from respective dataset owners. Associated code is available in Github repository Link.

Institutional Review Board (IRB) Since all datasets are publicly available, this study did not require an IRB.

1. Introduction

Electroencephalography (EEG) plays a central role in clinical diagnostics and critical care. It is ubiquitously deployed in Epilepsy Monitoring Units (Maganti and Rutecki, 2013) for seizure monitoring and in polysomnography (Rundo and Downey III, 2019) for diagnosing sleep disorders. Beyond clinical settings, recent developments in sensor hardware such as dry electrodes and portable, in-ear EEG (Looney et al., 2012) yield promising directions for deployment of EEG wearable devices in ambulatory and real-world environments. Prior works have further demonstrated the utility of EEG in Brain-Computer Interfaces (BCIs) (Lotte et al., 2018) and neurorehabilitation (Daly and Wolpaw, 2008). However, the idiosyncratic characteristics of EEG datasets, such as heterogeneous recording settings, low signal-to-noise ratio (SNR), and high annotation costs, continue to limit generalizability and performance in these applications (Avramidis et al., 2025).

In response to these challenges, there has been growing interest in extending the foundation model paradigm to biosignals (Abbaspourazad et al., 2023),

particularly EEG (Zhou et al., 2025; Wang et al., 2024b; Kostas et al., 2021; Yang et al., 2023; Kommineni et al., 2024; Jiang et al., 2024; Avramidis et al., 2025). Motivated by the rapid development of foundation models in audio, vision, and text, researchers hypothesize that large-scale pre-training could enable develop generalized representations robust to intrinsic variability and data scarcity inherent to EEG analysis. In these other domains, evaluation paradigms for foundation models have evolved beyond full fine-tuning to include low-resource adaptation (Zhang et al., 2024; Dong et al., 2023), in-context learning (Brown et al., 2020), and zero-shot transfer (Elizalde et al., 2023). This shift reflects the intended use of foundation models, namely leveraging large-scale pre-training to reduce downstream supervision and adaptation costs.

Despite this potential, the majority of existing evaluations of EEG foundation models rely on fine-tuning of *all* model parameters using the entirety of downstream datasets (Zhou et al., 2025; Wang et al., 2024b; Yang et al., 2023; Jiang et al., 2024). Evaluating these models solely under full fine-tuning further obscures their deployment utility whenever data, channels, or computational resources are limited.

To address this methodological gap, we propose a multi-dimensional approach to systematically evaluate EEG foundation models across dimensions directly corresponding to real-world constraints. In realistic clinical or ambulatory scenarios, models must often operate with limited subject data, reduced channel montages, short recording durations, or strict parameter budgets. Consequently, our framework evaluates generalization capabilities across model scale, data availability, and channel layout. The contributions of this work are as follows:

- We introduce a multi-dimensional evaluation framework, grounded in real-world deployment constraints, to assess the generalization behavior of EEG foundation models along three key axes: *parameter*, *sample*, and *channel* efficiency.
- We implement parameter-efficient adaptation and low-resource training to probe the quality and utility of learned representations, quantifying the relative gains provided by large-scale pre-training over fully supervised baselines.
- We show that EEG foundation models exhibit improved sample efficiency on long-context temporal modeling tasks under severe data constraints, while achieving performance comparable to supervised models on short-window BCI tasks, highlighting task-dependent benefits and limitations of current pre-training strategies.

2. Related Work

2.1. Foundation models for EEG

Until recently, advances in representation learning for vision, audio, and text modeling were the primary motivators of foundation models for biosignals, and EEG in specific. Early work adapted self-supervised contrastive objectives (Kostas et al., 2021), demonstrating improved transferability across multiple disparate settings. The heterogeneous nature of neurophysiological recordings later shifted the research focus towards masked-reconstruction objectives (Cui et al., 2024; Wang et al., 2024b), mainly considering transformer architecture (Wang et al., 2024a). Still, scarcity of large-scale annotated data and high inter-subject variability of EEG acts as a performance bottleneck, such that more recent studies have sought to improve sample efficiency through methods, such as masked reconstruction (Wang et al., 2024b; Zhou et al., 2025; Wang et al., 2025; Ma et al., 2025), knowledge-driven objectives (Kommineni et al., 2024; Jiang et al., 2025), vector-quantized representation learning (Jiang et al., 2024; Avramidis et al., 2025) and autoregressive modelling Liu et al. (2025).

2.2. Foundation Model Evaluation

In adjacent domains, foundation models are primarily evaluated through transfer-oriented protocols, as full fine-tuning is often infeasible or costly. Common evaluation strategies include few-shot generalization, linear probing, and parameter-efficient adaptation, with comparisons against supervised models trained from scratch on each task. Among parameter-efficient methods, low-rank adaptation (LoRA) has emerged as a standard approach, updating only a small number of trainable parameters while keeping the model backbone frozen (Hu et al., 2022). Across vision, audio, and text, performance gains from foundation models are consistently most pronounced in low-resource regimes (Brown et al., 2020), where fully supervised models tend to underfit or overfit.

In contrast, evaluation of EEG foundation models remains less standardized and is often limited to fine-tuning on a diverse set of downstream tasks. Most existing studies report improvements over supervised

baselines via full end-to-end fine-tuning (Jiang et al., 2024; Wang et al., 2024b), with comparatively fewer works restricting adaptation to task-specific heads or linear probes (Lee et al., 2025). As a result, systematic evaluation protocols that isolate generalization and parameter efficiency across recording conditions remain limited (Bomatter and Gouk, 2025).

2.3. Heterogeneity in EEG Data

A prominent challenge in developing and evaluating EEG foundation models lies in the heterogeneity of EEG data. EEG signals vary substantially across subjects, sessions, tasks, and recording setups, which reflects differences in cognitive state, electrode layout, and acquisition hardware. Prior studies have shown that inter-subject variability is a major limiting factor for EEG-based classifications. As a result, models trained on limited-scale multi-subject data frequently struggle to generalize to unseen individuals. Moreover, topological heterogeneity is another fundamental and pervasive challenge in EEG modeling, where each public dataset typically uses its own electrode layout. Some recent studies, like LUNA (Döner et al., 2025) and MMM (Yi et al., 2023), have proposed a topological-invariant encoder to map arbitrary-sized channeled EEG data to a uniform and fixed-sized latent representation space. Finally, and importantly, task heterogeneity is also a frequently encountered barrier in developing EEG foundation models, as different cognitive states—such as sleep stages, motor imagery, or seizure events—are characterized by distinct spectral and temporal brain activities. As mentioned earlier, recent studies Zhou et al. (2025); Wang et al. (2024b); Jiang et al. (2024) aimed to improve cross-task generalization by scaling pre-training to larger volumes of diverse EEG data to learn robust and task-agnostic representations.

3. Evaluating Generalization in EEG Foundation Models

The core principle of foundation models is that large-scale unsupervised pre-training yields generalizable representations, enabling downstream adaptation under low-resource settings in a parameter efficient manner (Brown et al., 2020; Radford et al., 2023; Narayanswamy et al., 2024; Xu et al., 2025; Siméoni et al., 2025). The methodology for evaluating generalizability differs from one modality to other according to the idiosyncratic characteristics of the underlying signal. For speech and text, models are evaluated based on their low-resource cross-domain and cross-lingual transfer capabilities (Baevski et al., 2020; Wei et al., 2021), whereas generalization in the domain of vision is characterized by the ability to extract semantically robust features across heterogeneous settings (Radford et al., 2021; Siméoni et al., 2025).

However, EEG presents a different set of challenges. Low signal to noise ratio, non-stationarity and large inter-subject and intra-subject variability often limit the ability of supervised models to generalize across settings. Moreover, large scale EEG data collections are constrained by participant fatigue, setup time and hardware limitations (Sugden et al., 2023). As a result, generalization for EEG foundation models cannot be adequately captured by a single notion of transfer performance or dataset scaling. We argue that generalization in EEG should be evaluated on a multidimensional axis, one that reflects realistic downstream constraints. Through a combination of the three dimensions of parameter, sample, and channel efficiency, as well as corresponding details to operationalize them, we propose a generalized evaluation framework to assess the capabilities of EEG foundation models.

Parameter Efficiency While parameter efficiency is not unique to EEG models, the majority of existing EEG evaluations rely on full fine-tuning, where all parameters of a pre-trained model are adapted to the downstream task. As a result, full fine-tuning performance alone provides limited insight into the quality of the learned representations, since improvements may stem primarily from parameter updates rather than transferable features.

We study parameter efficiency under two complementary settings: *linear probing* and *parameter-efficient fine-tuning (PEFT)*. Linear probing evaluates the quality and generalizability of representations learned by the foundation model through training only a lightweight classifier on frozen features, while PEFT assesses the model's ability to adapt to downstream tasks using a small number of additional or modified parameters. To operationalize parameter efficiency, we define a relative performance metric (PE_S) that normalizes downstream performance under a given setting with respect to full fine-tuning:

$$PE_S = \frac{P_S - P_{chance}}{P_{FT} - P_{chance}} \quad (1)$$

where $S \in \{\text{Linear Probe}, \text{PEFT}\}$, P_S denotes the performance under setting S, P_{FT} corresponds to per-

Dataset	Task	# Classes	# Subjects	Length (sec)	# Channels	Sampling Rate (Hz)
Physionet-MI	Motor Imagery	4	109	4	64	160
BCIC IV 2A	Motor Imagery	4	9	4	22	250
Kaggle ERN	P300	2	26	1	56	200
MDD MAL	Mental Health	2	64	10	19	256
Sleep EDF	Sleep Stage	5	78	30	2	100
TUEV	Events	6	370	5	21	200

Table 1: Summary of dataset used for downstream evaluation of models. Sampling rate is reported in Hz and length in seconds. Supervised models were trained at the sampling rate of the recording whereas for SSL models, signals were resampled to 200Hz. Refer to Appendix A for additional dataset details.

Model	Param Size	Training Data Size	Training Objective
LaBraM	5.8M	2.5+ kh	Masked Token Prediction
CBraMod	5M	27.1 kh	Masked Reconstruction
CSBrain	9M	9+ kh	Masked Reconstruction

Table 2: Overview of the parameter size, the pre-training data size, and the pre-training objective in EEG foundation models used for evaluation in this work.

formance under full fine-tuning, and P_{chance} is the random-chance baseline. The performance metric P is chosen according to the downstream task (e.g., balanced accuracy, macro-averaged F1). The parameter efficiency score PE_S measures the fraction of full fine-tuning performance achieved under a parameter-efficient regime, effectively disentangling representational quality from the benefits of extensive parameter adaptation. Models with high-quality and transferable representations are expected to exhibit values of both $PE_{\text{Linear Probe}}$ and PE_{PEFT} close to 1.

Sample Efficiency Beyond parameter efficiency, a core criterion for evaluating foundation models is their ability to adapt under low-label regimes. Such settings are pervasive in EEG due to high annotation costs, limited data availability, and privacy constraints that restrict large-scale annotated data sharing. To assess sample efficiency, we evaluate foundation models under varying total sample budgets, denoted by S_{total}, which corresponds to the total number of labeled training samples available for downstream adaptation. Experiments are conducted across representative values of S_{total} under both linear probing and PEFT settings.

In addition to the total number of samples, an important consideration in low-resource EEG settings is the trade-off between participant diversity and the number of samples per participant. To capture this effect, we vary the number of subjects while keeping S_{total} fixed. Hence, increasing the number of subjects leads to a proportionate decrease in number of samples per subject, enabling an explicit study of the balance between inter-subject diversity and per-subject data density under a fixed training budget. While absolute performance under a given budget indicates whether a model can operate effectively in low-resource conditions, it does not reveal whether pre-training provides benefits beyond those achievable by supervised learning alone. To isolate the contribution of pre-training, we define sample efficiency as a relative performance measured with respect to a supervised baseline trained under the same total sample budget:

$$SE^{D}_{S_{\text{Total}}} = \frac{P_D^{S_{\text{Total}}} - P_{\text{chance}}}{P_{\text{Sup}}^{S_{\text{Total}}} - P_{\text{chance}}} \tag{2}$$

where D denotes the adaptation setting (linear probe or PEFT), $P_D^{S_{\text{Total}}}$ is the downstream performance of the foundation model under budget S_{Total}, $P_{\text{Sup}}^{S_{\text{Total}}}$ is the performance of the supervised baseline trained with the same budget, and P_{chance} denotes chance-level performance for the chosen evaluation metric. Under this formulation, sample efficiency quantifies the relative advantage conferred by pre-training at a given data budget. Values of $SE^{D}_{S_{\text{Total}}} > 1$ indicate that the foundation model outperforms the supervised baseline under identical sampling constraints, while values near or below 1 suggest limited or no benefit from pre-training in that regime.

Channel Efficiency In most laboratory and clinical research settings, EEG is recorded using dense electrode montages, typically ranging from 64 to 256 channels. While such high-density recordings are essential for detailed spatial characterization and source localization of neural activity, they are impractical for many real-world and consumer-facing appli-

cations, where setup time, comfort, cost, and hardware constraints necessitate substantially fewer electrodes. To evaluate whether EEG foundation models exhibit robustness under reduced sensing configurations, we study model performance under systematically constrained channel settings. Rather than randomly subsampling electrodes, we define two structured channel selection criteria that reflect realistic deployment scenarios.

First, we consider a *sparse montage* setting, in which a fixed number of channels per cortical lobe are selected. This emulates low-density EEG caps while preserving coarse spatial coverage across the scalp. The number of channels per lobe is varied to progressively reduce the total channel count. Second, we consider a *lobe-restricted* setting, where electrodes are selected exclusively from a single cortical region (e.g., frontal, central, or midline). This setting reflects applications where electrodes are placed on localized brain regions due to task relevance or hardware limitations. By evaluating performance under these controlled channel constraints, we assess the extent to which foundation models rely on dense spatial information versus their ability to leverage robust, transferable representations across brain regions under severe channel reduction. Unlike parameter and sample efficiency, channel efficiency is not summarized by a single scalar metric; instead, performance trends across channel configurations provide insight into spatial robustness and inductive biases learned during pre-training.

4. Experimental Setup

4.1. Datasets

To enable a holistic evaluation of performance across the proposed generalization dimensions, we conduct experiments on six representative EEG datasets spanning both short and long-duration tasks. Short-window EEG tasks are often encountered in brain–computer interface (BCI) settings, where decisions are made from brief recording segments ($\leq$5s). To reflect this regime, PhysioNet Motor Imagery (PhysioNet-MI), BCI Competition IV-2A (BCIC IV-2A), Kaggle Error-Related Negativity (Kaggle-ERN) and TUEV are included in this evaluation. PhysioNet-MI (Schalk et al., 2004; Goldberger et al., 2000) and BCIC IV-2A (Brunner et al., 2008) are motor imagery datasets in which subjects perform real or imagined movements of specific body parts in response to visual cues. Kaggle-ERN (Margaux et al., 2012; Mattout et al., 2014) focuses on the detection of error-related responses, elicited when subjects attempt to spell a word using visually guided stimuli. TUEV (Obeid and Picone, 2016; Harati et al., 2015) is a neurological event detection dataset.

In contrast to short-window BCI tasks, an increasing number of clinically relevant EEG applications require modeling long temporal contexts. To capture this setting, we evaluate on MDD MAL (Mumtaz, 2016) and Sleep-EDF (Kemp et al., 2000) datasets which correspond to automated sleep staging and mental health assessment respectively.

Together, these datasets span diverse task structures, temporal scales, and clinical contexts, enabling a systematic assessment of generalization under varying resource and signal constraints. To ensure robustness in results, BCIC IV-2A was evaluated in a Leave-One-Subject-Out (LOSO) setting and all other datasets were tested in a 5-fold, between-subjects cross-validation setup. For additional dataset details please refer to Table 1 and Appendix A.

4.2. Data Preprocessing

We adopted a uniform preprocessing pipeline for all supervised experiments which included notch filtering for powerline noise removal, band-pass filtering, re-referencing to the common average, and subsequent z-score normalization. For TUEV dataset, inter-quantile normalization was chosen over z-score to better preserve high-amplitude signal characteristics associated with artifact and epileptic events. For foundation models, we follow the preprocessing procedures reported in the respective original publications to ensure reproducibility of results.

4.3. Model Architectures

We evaluate 3 representative EEG foundation models as described in Table 2: LaBraM (Jiang et al., 2024), CBraMod (Wang et al., 2024b), and CSBrain (Zhou et al., 2025). For comparison against fully supervised approaches, the following EEG models are included: EEGNet (Lawhern et al., 2018), EEGNeX (Chen et al., 2024), and SparcNet (Jing et al., 2023). Refer to Appendix C for further implementation details and model training. To assess the parameter efficiency of foundation models during downstream adaptation, three fine-tuning strategies were considered: (1) *Full fine-tuning*, in which all model parameters are updated; (2) *Linear probing (LP)*, where

Model	K-ERN	P-MI	IV-2A	TUEV	MDD MAL	Sleep EDF
`Supervised`						
EEGNet (4K)	62.74 ± 2.09	61.34 ± 1.91	54.96 ± 7.05	51.67 ± 2.10	86.89 ± 3.88	70.20 ± 1.29
EEGNet Large (20K)	64.94 ± 2.23	61.85 ± 1.95	58.14 ± 7.07	53.27 ± 3.88	84.98 ± 6.82	71.61 ± 1.20
EEGNet Huge (110K)	$\underline{64.54 \pm 1.48}$	61.74 ± 1.73	$\underline{58.76 \pm 6.60}$	52.58 ± 1.82	83.61 ± 7.21	71.14 ± 1.36
EEGNex (50k)	$\mathbf{65.00 \pm 1.38}$	$\mathbf{65.58 \pm 1.73}$	$\mathbf{59.10 \pm 11.25}$	47.50 ± 3.01	86.21 ± 8.36	70.31 ± 1.22
SparcNet (1M)	61.70 ± 1.30	62.02 ± 1.54	56.96 ± 10.43	52.10 ± 3.51	79.32 ± 6.22	71.01 ± 2.64
`Full Finetune`						
LaBraM (5.8M)	58.51 ± 1.79	57.25 ± 1.55	50.58 ± 9.20	$\mathbf{57.58 \pm 1.91}$	$\underline{88.72 \pm 3.12}$	$\underline{72.86 \pm 1.22}$
CBraMod (5M)	61.47 ± 1.25	$\underline{64.01 \pm 2.52}$	56.15 ± 8.77	$\underline{54.70 \pm 2.34}$	83.08 ± 6.69	$\mathbf{74.58 \pm 0.85}$
CSBrain (9M)	63.19 ± 1.60	62.34 ± 2.33	55.27 ± 8.39	51.88 ± 2.16	$\mathbf{88.81 \pm 5.41}$	71.17 ± 1.17

Table 3: Downstream classification results for end-to-end training of supervised models and full-finetuning setting for foundation models. K-ERN, P-MI and IV-2A refer to Kaggle ERN, Physionet-MI and BCI Competition IV-2A dataset respectively. Balanced Accuracy (BAC) metric is reported. For additional evaluation metrics, refer to Appendix F. Parameter count for each model is reported within parentheses. Best and second-best performing metrics are in bold and underline, respectively.

the pre-trained backbone is frozen and only a linear head is trained; and (3) *Parameter Efficient Fine-tuning (PEFT)*, which introduces a small number of trainable parameters to enable efficient adaptation. These adaptation strategies allow us to disentangle performance gains arising from representation quality versus those attributable to increased parameter capacity.

For all linear probe evaluations, embeddings from the foundation models were flattened and a fully connected block with a single multi-head projection layer was used for fine-tuning. For parameter-efficient fine-tuning, low-rank adaptation (LoRA, by Hu et al. (2022)) modules were used for linear layers.

4.4. Evaluation Metrics

We evaluated all models using performance metrics that are well-established in EEG evaluation. We report balanced accuracy (BAC), F1-macro, and area under the receiver operating characteristic curve (AUROC)–for multiclass classification we instead report Cohen's kappa. All the utilized metrics are described in detail in Appendix D.

5. Results

To establish an upper bound on downstream performance and contextualize subsequent resource-constrained analyses, we first evaluate supervised baseline models and fully fine-tuned foundation models (LaBraM, CBraMod and CSBrain) across the six evaluation datasets. During foundation model fine-tuning, weights were initialized from their respective pre-trained checkpoints. Results in Table 3 reveal complementary strengths across model classes. The benefits of foundation models are most pronounced in long-window clinical tasks. For example, in depression classification the top-performing foundation model achieves an absolute improvement of 4.2% in Cohen's Kappa, whereas on TUEV it achieves 4.0% Macro-F1 improvement compared to the top-performing fully supervised model. However, foundation models are on par or slightly underperform supervised models in short-window BCI/ERN tasks, indicating limited effectiveness in modeling short temporal contexts. Ablation experiments comparing the performance of supervised models at native sampling rate of corresponding dataset against the sampling rate of foundation models (200 Hz) indicate no noticeable difference, indicating that the performance discrepancy on short window tasks arises from model architecture and not sampling discrepancies (Appendix G). Additional evaluation metrics for all experiments are reported in Appendix F.

5.1. Parameter Efficiency

Evaluating foundation models under parameter constraints offers insights into the intrinsic quality of the pre-trained representations. As outlined in Section 4.3, we evaluate efficiency under two distinct regimes: a linear probe setting, which measures representation quality of frozen pre-trained features, and a PEFT setting, which tests the model's ability to adapt using only a fraction of the trainable parameters required

Model	K-ERN	P-MI	IV-2A	TUEV	MDD MAL	Sleep EDF
`Linear Probe`						
LaBraM	$0.74 \pm .16$	$0.74 \pm .01$	$0.82 \pm .11$	$0.64 \pm .07$	$0.95 \pm .06$	$0.84 \pm .02$
CBraMod	$0.68 \pm .08$	$0.69 \pm .01$	$0.62 \pm .12$	$0.55 \pm .06$	$0.96 \pm .25$	$0.69 \pm .02$
CSBrain	$0.83 \pm .05$	$0.72 \pm .03$	$0.71 \pm .10$	$0.87 \pm .08$	$0.98 \pm .11$	$0.85 \pm .03$
`PEFT`						
LaBraM	$1.02 \pm .25$	$0.75 \pm .01$	$0.70 \pm .07$	$0.67 \pm .07$	$0.97 \pm .06$	$0.87 \pm .05$
CBraMod	$0.93 \pm .05$	$0.95 \pm .02$	$0.85 \pm .18$	$0.95 \pm .08$	$1.01 \pm .15$	$0.91 \pm .03$
CSBrain	$0.89 \pm .15$	$0.91 \pm .04$	$1.05 \pm .13$	$1.01 \pm .08$	$1.01 \pm .07$	$1.03 \pm .02$

Table 4: Performance Efficiency (PE_S) values computed on Balanced Accuracy using Eq 1 for LaBraM, CBraMod and CSBrain foundation models for the 6 evaluation datasets. For long-window tasks such as depression classification (MDD MAL) and sleep stage detection (Sleep EDF), foundation models provide higher parameter efficiency values compared to short-window tasks, indicating their representations being better aligned to long-context modeling. K-ERN, P-MI and IV-2A refer to Kaggle ERN, Physionet-MI and BCI Competition IV-2A dataset respectively.

for full fine-tuning. In Table 4 we report parameter efficiency values (PE_D) computed using balanced accuracy for linear probe and PEFT settings across evaluation datasets.

Linear Probe CSBrain provides better representation quality over LaBraM and CbraMod models, as indicated by the higher linear probe parameter efficiency $PE_{\text{Linear probe}}$ values. This could be owing to better spatiotemporal information encapsulated in CSBrain, facilitated by inclusion of cortical lobe based embeddings (Zhou et al., 2025). When comparing linear probe PE across tasks, a distinct phenomenon is observed; across the three foundation models, short-window tasks (BCI IV-2A, Physionet-MI, Kaggle ERN, and TUEV) on average have lower values compared to long-window tasks (Sleep EDF, MDD MAL). Plot of average $PE_{\text{Linear probe}}$ between short-window and long-window tasks for the EEG foundation models in Figure 1 reveals the average $PE_{\text{Linear probe}}$ for short-window tasks to be at least 12% lower (absolute) than for long-window tasks. This performance gap suggests a fundamental difference in representation quality wherein foundation models provide better quality representations for longer window tasks.

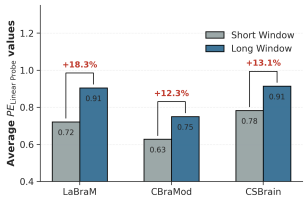

Figure 1: $PE_{\text{Linear probe}}$ values computed on BAC for LaBraM, CBraMod and CSBrain, aggregated based on task window length. Long-window tasks (Blue) parameter efficiency values are higher than short-window tasks (Grey) across all models.

PEFT Performance: For all parameter-efficient fine-tuning (PEFT) experiments, we employ LoRA with rank $r = 4$ and scaling factor $\alpha = 8$, resulting in only approximately 2–4% of the full model parameters being trainable. Despite this substantial reduction in trainable parameters, PEFT-adapted models achieve performance comparable to full fine-tuning on depression classification (MDD-MAL) and EEG event classification (TUEV), as shown in Table 4. In BCI tasks, PEFT substantially outperforms linear probing, with CBraMod and CSBrain attaining parameter efficiency values of up to 0.95. This indicates that updating the full parameter set (such as full fine-tuning setting) yields limited additional gains relative to the increase in model capacity, highlighting diminishing returns from full fine-tuning in these settings. These results suggest that PEFT provides a more effective trade-off between performance and trainable parameter count for EEG foundation models, particularly in resource-constrained adaptation scenarios.

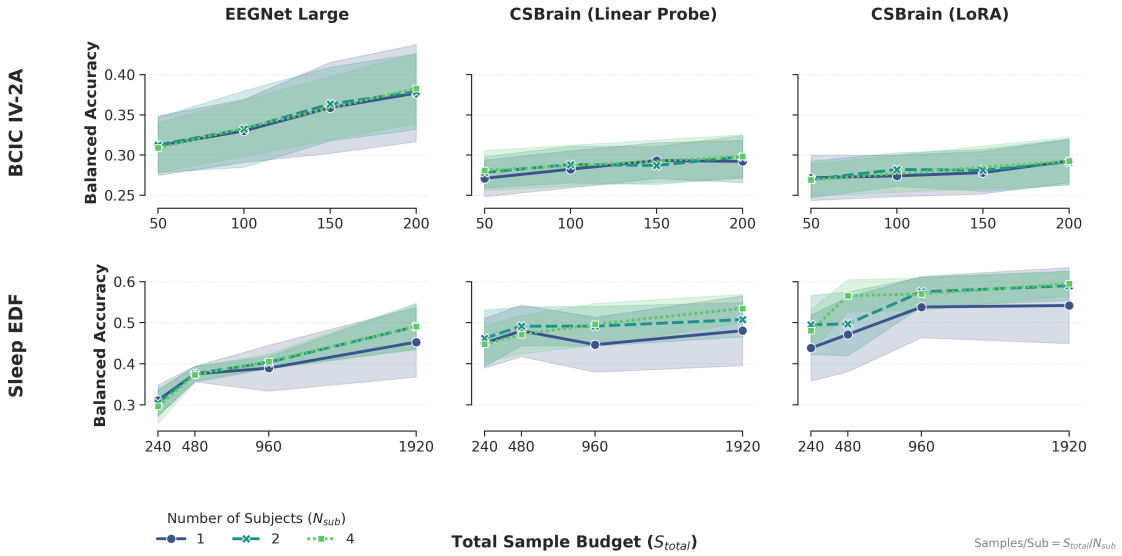

Figure 2: *Sampling Efficiency*: Classification results for BCIC IV-2A and Sleep EDF datasets under fixed Total Sample Budget (S_{total}) for EEGNet Large and CSBrain under Linear Probe and LoRA. To ensure fixed S_{total} under $N_{subjects} \in \{1, 2, 4\}$, number of samples per subject are adjusted accordingly. Each model run is repeated for 3 random seeds for 5 folds in Sleep EDF and Leave-one-subject-out setting for BCIC IV-2A.

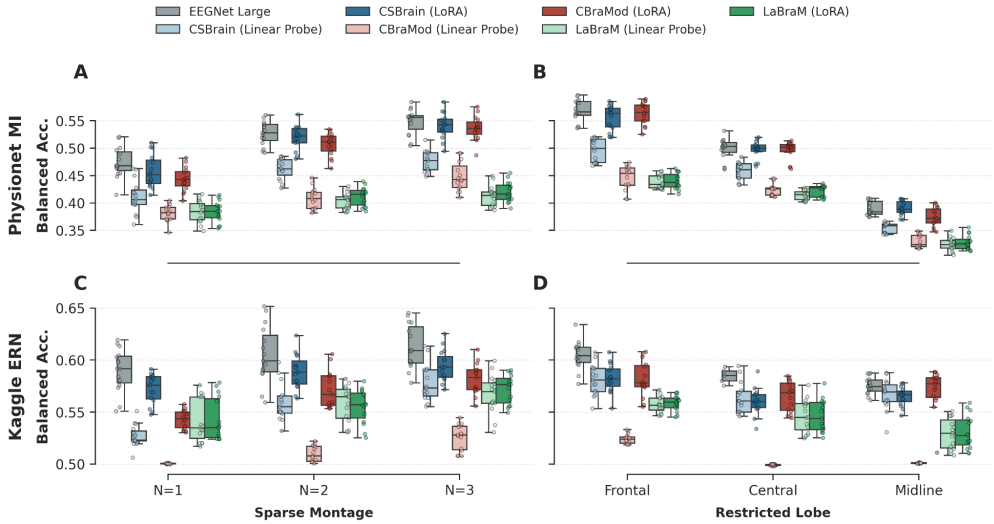

Figure 3: *Channel Efficiency*: Boxplots of balanced accuracy of supervised (EEGNet Large) and foundation models (CSBrain, CBraMod, LaBraM) performance for Physionet MI and Kaggle ERN datasets under sparse montage and restricted spatial lobe conditions. (A,C) correspond to classification performance under uniformly reducing channels per corticol lobe (central, frontal, temporal, parietal and occipital) to $N \in \{1, 2, 3\}$. (B,D) correspond to classification performance of models under selecting frontal, central and midline lobes.

Dataset	Model	Total Sampling Budget (S_{Total})			
		50	**100**	**150**	**200**
BCIC IV-2A	Linear	0.76 ± 1.45	0.40 ± 0.44	0.48 ± 0.61	0.69 ± 0.93
	LoRA	0.56 ± 0.89	0.41 ± 0.45	0.16 ± 0.81	0.75 ± 1.59
		240	**480**	**960**	**1920**
Sleep EDF	Linear	$2.91 \pm 1.60^{**}$	$1.63 \pm 0.37^{**}$	$1.37 \pm 0.29^{**}$	$1.12 \pm 0.30^{*}$
	LoRA	$3.12 \pm 2.04^{**}$	$1.79 \pm 0.51^{**}$	$1.78 \pm 0.28^{**}$	$1.31 \pm 0.36^{**}$

Table 5: Average Sampling Efficiency values at a given sampling budget across different test folds (using performance metric as Balanced accuracy in Eq. 2) for CSBrain (Foundation Model) Linear and LoRA settings at fixed total sampling budget compared to EEGNet Large (Supervised). Column headers denote budget for BCIC IV-2A and Sleep EDF respectively. Significance test for CSBrain performance being better than EEGNet Large are reported $^{*}p < 0.05$, $^{**}p < 0.001$.

5.2. Sampling Efficiency

To study model behavior under low-resource settings, we evaluate performance under a fixed total sample budget (S_{Total}), representing the size of the training set used for downstream classification. Experiments are conducted on two representative datasets: a short-window motor imagery task (BCIC IV-2A) and a long-window sleep staging task (Sleep EDF). Sampling efficiency values ($SE^{D}_{S_{\text{Total}}}$) are computed using EEGNet-Large as the supervised baseline, while CSBrain is evaluated under linear probing and LoRA fine-tuning ($r = 1$) settings. For BCIC IV-2A, total sample budgets of 50 ($\sim$5 minutes of data), 100, 150, and 200 samples are considered. For Sleep EDF, budgets are set to 240 ($\sim$2 hours), 480, 960, and 1920 samples. In all cases, samples are class-stratified within each subject to control for class imbalance.

Sampling efficiency results in Table 5 reveal distinct behaviors across the two tasks. For Sleep EDF, CSBrain under both linear probing and LoRA adaptation consistently outperforms EEGNet-Large. Moreover, the relative performance gains increase as the total sample budget decreases, with average $SE^{D}_{S_{\text{Total}}}$ values of 2.91 for linear probing and 3.12 for LoRA at 240 samples ($\sim$2 hours of data). As more training data becomes available, the magnitude of $SE^{D}_{S_{\text{Total}}}$ decreases, yet remains statistically significantly greater than one, indicating sustained advantages for foundation model representations. In contrast, for the BCIC IV-2A dataset, $SE^{D}_{S_{\text{Total}}}$ values remain below one across all sampling budgets, highlighting the limited effectiveness of current EEG foundation models under low-sample regimes for short-window BCI tasks.

Low-resource settings in the context of EEG arise along two distinct axes, limited number of subjects and a limited number of samples per subject. As shown in Fig. 2, under severe data constraints, increasing subject diversity does not compensate for a limited total sample budget. Instead, performance is primarily governed by the total number of training samples (S_{total}), largely independent of how those samples are distributed across subjects. While prior works (Bomatter and Gouk, 2025) indicate some improvement in downstream performance with increase in subject diversity, their evaluations were conducted under much larger training sample set, which could indicate different behavior depending on scale of training data. Additional Sample Efficiency results for MDD-MAL dataset are provided in Appendix H.

5.3. Channel Efficiency

To study the robustness of EEG foundation models under reduced channel availability, we consider two complementary channel-reduction strategies, as described in Section 4: a sparse-montage setting and a lobe-restricted setting. In the sparse-montage setting, the number of channels per cortical lobe (central, frontal, temporal, parietal, and occipital) is uniformly reduced to $N \in \{1, 2, 3\}$, resulting in total channel counts of $\{5, 10, 15\}$, respectively. This setup emulates scenarios where sparse yet spatially distributed electrode coverage is required. In the lobe-restricted setting, channels are confined to specific regions of the montage—namely frontal, central, and

midline—reflecting practical constraints where only partial scalp coverage is available. Both channel-reduction settings are evaluated on two BCI tasks: motor imagery classification using the Physionet MI dataset and error-related negativity detection using the Kaggle ERN dataset. We compare the supervised baseline EEGNet-Large with the foundation models CBraMod, CSBrain and LaBraM under parameter-efficient adaptation via linear probing and LoRA.

Results in Fig. 3 show that foundation models perform on par with, or slightly worse than, EEGNet-Large across all channel configurations. This suggests that current EEG foundation models do not yet exhibit enhanced robustness to reduced or regionally restricted channel inputs, indicating the need for improved pre-training strategies that more explicitly capture channel-specific and spatial inductive biases. Comparing the two channel selection strategies, uniform per-lobe sampling yields more stable performance across tasks despite introducing sparsity. In contrast, restricting channels to a single region can lead to pronounced performance degradation when task-relevant information is distributed across lobes. This effect is particularly evident for Physionet MI when using only midline electrodes, as motor imagery signals are known to be strongly lateralized as seen in Fig. 3B.

6. Discussion

Tokenization Granularity could limit short window generalizability in Foundation Models. Despite strong performance on long-sequence inputs (> 5s), foundation models underperform supervised baselines on short-window tasks such as Physionet-MI, BCIC IV-2A, and Kaggle-ERN, even with substantially higher parameter counts. This discrepancy could be attributed to the tokenization strategies employed by foundation models. Specifically, EEG signals are typically segmented into coarse one second time window patches and processed using transformer-based backbones in order to enable modeling of long temporal contexts during pre-training (Zhou et al., 2025; Wang et al., 2024b). While such tokenization is necessary to model long-context temporal dependencies, as reflected in the observed performance gains in sleep stage prediction and depression classification, it might limit the model's ability to capture fine-grained, transient neural dynamics, such as ERP-like responses, that are critical for short-window classification tasks.

EEGNeX Performance in Short-Window BCI Tasks. EEGNeX consistently outperforms other models on short-window BCI tasks. Prior work attributes this advantage to the use of strided convolutions (Chen et al., 2024), which may enable more effective aggregation of local temporal context within brief signal segments compared to standard CNN designs and transformer-based foundation models.

Limitations This work primarily focuses on masked reconstruction based EEG foundation models (LaBram, CBraMod and CSBrain) trained under varying pre-training data sizes, parameter counts and compute budgets. Future work should consider evaluating EEG foundation models trained under identical settings, that enables isolating impact of individual pre-training axes on downstream performance. Additionally, while we analyze sample efficiency under constrained data budgets on representative datasets, we do not provide a comprehensive evaluation across a broader range of tasks and recording conditions. In particular, the interaction between inter-subject and intra-subject variability and different data sampling strategies in low-resource EEG settings remains an open problem that warrants further study.

7. Conclusion

In this work, we propose a multi-dimensional framework to evaluate EEG foundation models. By systematically assessing performance across parameter, sample, and channel constraints, we uncovered distinct trade-offs across these dimensions in current EEG foundation models. Our empirical analysis across six diverse datasets reveals that the utility of current EEG foundation models is highly task-dependent. We observed that these models excel in long-context tasks, such as sleep staging and mental health assessment, where they demonstrate performance gains over supervised baselines under full-finetuning and resource constrained settings. However, this advantage does not yet extend to short-window BCI tasks or scenarios with channel constraints. In these settings, supervised models (like EEGNet, EEGNeX) remain competitive or superior, indicating that current tokenization and pre-training objectives may not adequately capture the fine-grained spatiotemporal features necessary for tasks like motor imagery. Future research could focus on developing pre-training objectives that explicitly encode short-window dynamics and channel invari-

ance to bridge the gap between high-resource research benchmarks and realistic, low-resource deployment.

Generative AI Use Disclosure

Large Language Models (LLMs) were used for refining text during preparation of manuscript. Additionally, LLMs were employed to generate parts of the code implementation. For LLM-generated content in the manuscript and code, the content was verified by the authors prior to inclusion.

Acknowledgments

This study was sponsored by the Defense Advanced Research Projects Agency (DARPA) under cooperative agreement No. N660012324006. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Salar Abbaspourazad, Oussama Elachqar, Andrew C Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. *arXiv preprint arXiv:2312.05409*, 2023.

Kleanthis Avramidis, Tiantian Feng, Woojae Jeong, Jihwan Lee, Wenhui Cui, Richard M Leahy, and Shrikanth Narayanan. Neural codecs as biosignal tokenizers. *arXiv preprint arXiv:2510.09095*, 2025.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Philipp Bomatter and Henry Gouk. Is limited participant diversity impeding EEG-based machine learning? *arXiv preprint arXiv:2503.13497*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Clemens Brunner, Robert Leeb, Gernot Müller-Putz, Alois Schlögl, and Gert Pfurtscheller. Bci competition 2008–graz data set a. *Institute for knowledge discovery (laboratory of brain-computer interfaces), Graz University of Technology*, 16(1-6):34, 2008.

Xia Chen, Xiangbin Teng, Han Chen, Yafeng Pan, and Philipp Geyer. Toward reliable signals decoding for electroencephalogram: A benchmark study to EEGNeX. *Biomedical Signal Processing and Control*, 87:105475, 2024.

Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A Joshi, and Richard M Leahy. Neuro-gpt: Towards a foundation model for eeg. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.

Janis J Daly and Jonathan R Wolpaw. Brain–computer interfaces in neurological rehabilitation. *The Lancet Neurology*, 7(11):1032–1043, 2008.

Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. Luna: Efficient and topology-agnostic foundation model for EEG signal analysis. *arXiv preprint arXiv:2510.22257*, 2025.

Wei Dong, Dawei Yan, Zhijun Lin, and Peng Wang. Efficient adaptation of large vision transformer via adapter re-composing. *Advances in Neural Information Processing Systems*, 36:52548–52567, 2023.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti S. Hämäläinen. MEG

and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7(267):1–13, 2013. doi: 10.3389/fnins.2013.00267.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

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

Weibang Jiang, Yansen Wang, Bao liang Lu, and Dongsheng Li. NeuroLM: A universal multi-task foundation model for bridging the gap between language and EEG signals. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Io9yFt7XH7.

Jin Jing, Wendong Ge, Shenda Hong, Marta Bento Fernandes, Zhen Lin, Chaoqi Yang, Sungtae An, Aaron F Struck, Aline Herlopian, Ioannis Karakis, et al. Development of expert-level classification of seizures and rhythmic and periodic patterns during EEG interpretation. *Neurology*, 100(17):e1750–e1762, 2023.

Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47 (9):1185–1194, 2000.

Aditya Kommineni, Kleanthis Avramidis, Richard Leahy, and Shrikanth Narayanan. Knowledge-guided eeg representation learning. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–6. IEEE, 2024.

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. *Frontiers in Human Neuroscience*, 15:653659, 2021.

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

Na Lee, Konstantinos Barmpas, Yannis Panagakis, Dimitrios Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. Are large brainwave foundation models capable yet? insights from fine-tuning. *arXiv preprint arXiv:2507.01196*, 2025.

Chenyu Liu, Yuqiu Deng, Tianyu Liu, Jinan Zhou, Xinliang Zhou, Ziyu Jia, and Yi Ding. Echo: Toward contextual seq2seq paradigms in large eeg models. *arXiv preprint arXiv:2509.22556*, 2025.

David Looney, Preben Kidmose, Cheolsoo Park, Michael Ungstrup, Mike Lind Rank, Karin Rosenkranz, and Danilo P Mandic. The in-the-ear recording concept: User-centered and wearable brain monitoring. *IEEE pulse*, 3(6):32–42, 2012.

Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. *Journal of neural engineering*, 15(3):031005, 2018.

Jingying Ma, Feng Wu, Qika Lin, Yucheng Xing, Chenyu Liu, Ziyu Jia, and Mengling Feng. Codebrain: Towards decoupled interpretability and multi-scale architecture for eeg foundation model. *arXiv preprint arXiv:2506.09110*, 2025.

Rama K Maganti and Paul Rutecki. EEG and epilepsy monitoring. *CONTINUUM: Lifelong Learning in Neurology*, 19(3):598–622, 2013.

Perrin Margaux, Maby Emmanuel, Daligault Sébastien, Bertrand Olivier, and Mattout Jérémie. Objective and subjective evaluation of online error correction during p300-based spelling. *Advances in Human-Computer Interaction*, 2012(1):578295, 2012.

Jérémie Mattout, Manu, maucle, and Wendy Kan. Bci challenge @ ner 2015. https://kaggle.com/competitions/inria-bci-challenge, 2014. Kaggle.

Wajid Mumtaz. Mdd patients and healthy controls EEG data (new). *figshare, Dataset*, 2016.

Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam Tailor, Jake Sunshine, Yun Liu, et al. Scaling wearable foundation models. *arXiv preprint arXiv:2410.13638*, 2024.

Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in neuroscience*, 10:196, 2016.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

Jessica Vensel Rundo and Ralph Downey III. Polysomnography. *Handbook of clinical neurology*, 160:381–392, 2019.

Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Transactions on biomedical engineering*, 51(6):1034–1043, 2004.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Richard James Sugden, Viet-Linh Luke Pham-Kim-Nghiem-Phu, Ingrid Campbell, Alberto Leon, and Phedias Diamandis. Remote collection of electrophysiological data with brain wearables: opportunities and challenges. *Bioelectronic Medicine*, 9(1): 12, 2023.

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

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

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Shijian Li, and Gang Pan. Eegmamba: An eeg foundation model with mamba. *Neural Networks*, page 107816, 2025.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Maxwell A Xu, Girish Narayanswamy, Kumar Ayush, Dimitris Spathis, Shun Liao, Shyam A Tailor, Ahmed Metwally, A Ali Heydari, Yuwei Zhang, Jake Garrison, et al. Lsm-2: Learning from incomplete wearable sensor data. *arXiv preprint arXiv:2506.05321*, 2025.

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

Ke Yi, Yansen Wang, Kan Ren, and Dongsheng Li. Learning topology-agnostic EEG representations with geometry-aware modeling. *Advances in Neural Information Processing Systems*, 36:53875–53891, 2023.

Yunhua Zhang, Hazel Doughty, and Cees GM Snoek. Low-resource vision challenges for foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21956–21966, 2024.

Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. Csbrain: A cross-scale spatiotemporal brain foundation model for EEG decoding. *arXiv preprint arXiv:2506.23075*, 2025.

Appendix A. Referenced Datasets

A.1. Motor Imagery

Motor Imagery classification tasks include identifying a user's intention to move specific parts of the body. Robust motor imagery classification could allow for development of neuro-assistive technologies.

Physionet MI (Schalk et al., 2004; Goldberger et al., 2000): consists of 1500 1-2 minute EEG recordings from 109 participants, for four real and imagined motor tasks (open and close left fist, right fist, both fists and both feet). EEG recordings consist of 64 channels and were sampled at 160Hz. Each trial was segmented into 4-second windows and all experiments were performed in a 5-fold between subjects cross validation setting, such that each subject was in the test fold once. Balanced Accuracy, Cohen's Kappa and F1-Macro are reported.

BCI IV Competition 2A (Brunner et al., 2008): Motor Imagery Classification dataset consists of data recorded from 9 participants across two sessions recorded on different days. Data were collected using 22 EEG channels at 250Hz for four imagined motor movements (left arm, right arm, both feet and tongue). Trials are segmented into 4 second chunks and Leave-One-Subject-Out cross validation is performed for all classification experiments using this dataset. This is done to evaluate models ability to generalize across subjects.

A.2. Error Related Negativity

Kaggle ERN (Mattout et al., 2014): The dataset includes EEG recordings from 26 participants who perform tasks using an online P300 speller interface, and is primarily used to study event-related potentials related to erroneous responses. The EEG data were collected using 56 EEG electrodes and were downsampled to 200 Hz. The classification task is to detect when the selected item is not the intended.

A.3. Mental Health

MDD MAL (Mumtaz, 2016): Depression classification task from EEG recorded at resting state, composed of eyes closed and eyes open conditions. Each condition is recorded for 5 minutes with 19 electrodes at 200Hz sampling rate. The dataset consists of 34 patients diagnosed with major depressive disorder and 30 healthy controls. For classification, non-overlapping 10s windows are considered as input samples to the models. Evaluation is performed in a between subject 5-fold cross validation setup.

A.4. Sleep Stages

Sleep EDFx (Kemp et al., 2000): contains 197 whole-night PolySomnoGraphic sleep recordings with EEG, EOG and chin EMG recorded at 100Hz and annotated for sleep stages every 30s. In this work, EEG electrodes Fpz-Cz and Pz-Oz are considered for analysis. The sleep stages are annotated for Wake, REM, Movement, NREM-1, NREM-2, NREM-3 and NREM-4. For comparability to prior studies, NREM-3 and NREM-4 have been combined to a single class yielding 5 classes. 5-fold between-subjects cross validation is performed for evaluation.

A.5. Events

TUEV (Harati et al., 2015): contains EEG dataset derived from Temple University EEG Corpus (Obeid and Picone, 2016) with annotations for categorizing EEG segments into six classes: (1) spike and sharp wave (SPSW), (2) generalized periodic epileptiform discharges (GEPD), (3) periodic latealized epileptiform discharges (PLED), (4) eye movements (EYEM), (5) artifact and (6) background (BCKG). TUEV is recorded at 200Hz sampling rate and each annotation is segmented into 5s long EEG timeseries.

Appendix B. Referenced Models

EEGNet (Lawhern et al., 2018) is a lightweight convolution-based model that combines temporal and spatial convolution layers followed by a classification head. To accommodate for the varying sampling rates of the datasets, the kernel size of the temporal convolution layer and separable convolution layer had been set to half the sampling rate and one-eigth the sampling rate of the respective downstream dataset. To test the effects of scaling model sizes within EEGNet, three versionswere defined: (1) `EEGNet`: F1=8, D=2; (2) `EEGNet large`: F1=16, D=4; (3) `EEGNet Huge`: F1=32, D=8, where F! is the number of temporal filters and D is the depth multiplier.

EEGNeX (Chen et al., 2024) is an improved version of EEGNet architecture that employs strided convolutions to increase the effective temporal window that the model is able to capture, thereby providing performance improvements on BCI tasks.

SparcNet (Jing et al., 2023): 1D CNN model that leverages dense residual connections to capture spatiotemporal relations in EEG signals effectively.

LaBraM (Jiang et al., 2024) is an EEG foundation model with a pre-training objective based on masked token prediction. The underlying neural tokenizer is trained with large-scale EEG data through patching the EEG time series into tokens.

CBraMod (Wang et al., 2024b) is a foundation model trained on 25,000 hours pre-training data that aims to model temporal and spatial characteristics through distinct (criss-cross) attention mechanisms. This model employs a patch-based masked reconstruction scheme for pre-training.

CSBrain (Zhou et al., 2025) is an attention-based foundation model for EEG decoding with novel cross-scale spatiotemporal tokenization and structured sparse attention. It is pre-trained using masked reconstruction objectives.

Appendix C. Implementation Details

All experiments were conducted in Python 3.13. Preprocessing utilized the MNE-Python (Gramfort et al., 2013). PyTorch (Paszke et al., 2019) was used to build and train all deep learning models. All experiments were conducted on a cluster of 4 RTX A6000 GPUs.

All PEFT experiments utilized LoRa with rank $r = 4$. Gradient clipping with value 1.0 was used for full fine-tuning experiments. AdamW optimizer with learning rates between [1e-2, 1e-3] were used for linear probe, [1e-3, 1e-4] for LoRA and [2e-4, 1e-5] for full-finetuning experiments. Cosine Anneal Learning rate scheduler with 10% epochs as warmup steps was used across all experiments. For Fully supervised models, a learning rate plateau scheduler was used. All models were trained for up to 30 epochs for Physionet MI, BCIC IV-2A, Kaggle ERN and MDD MAL datasets, and for 50 epochs for Sleep EDF and TUEV datasets. The model checkpoint with the lowest overall validation loss was used for evaluation of downstream performance. Further details on hyperparameters are provided with the accompanying Github repository.

Parameter counts for Linear probe and LoRA corresponding to each dataset and model are as reported in Table 6

Appendix D. Metrics

Consistent with prior EEG self-supervised modeling works (Zhou et al., 2025; Kommineni et al., 2024; Avramidis et al., 2025; Wang et al., 2024b; Jiang et al., 2024), we use the following evaluation metrics:

- **Balanced Accuracy (BAC)**: Computes the unweighted average recall across classes which is able to better account for class imbalances in datasets compared to raw accuracy scores.

$$\text{BAC} = \frac{1}{K}\Sigma_{j=1}^{K}\frac{TP_j}{TP_j + FN_j}$$

  where K is the number of classes, TP_j is the number of True Positives and FN_j is the number of False Negatives for class j.

- **Cohen's Kappa (κ)**: Measures the mutual agreement between model predictions and true labels while accounting for agreement by chance. A value of 0 indicates no agreement between the model predictions and true labels, whereas 1.0 indicates perfect alignment.

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

  where p_o is the observed agreement and p_e is the hypothetical agreement by chance.

- **Area Under the Receiver Operating Curve (AUROC)**: Computes the area under the True Positive Rate (TPR) versus False Positive Rate (FPR) curve, computed at different thresholds for classification. An AUROC of 0.5 thus indicates random-chance performance, whereas a value of 1.0 indicates perfect classification.

- **F1-Macro**: Computes unweighted average of F1 scores across all classes in dataset. Since the average is unweighted, this metric is more robust to class imbalances.

$$\text{F1-Macro} = \frac{1}{K}\Sigma_{j=1}^{K}\frac{2.\text{Pr}_j.\text{Re}_j}{\text{Pr}_j + \text{Re}_j}$$

  where K is the number of classes, Pr_j and Re_j refer to precision and recall of class j.

Appendix E. CbraMod & LaBraM Sample Efficiency

Sampling Efficiency plot for CBraMod (Fig 4) and LaBraM (Fig 5) in comparison with EEGNet.

Model	Kaggle ERN	Physionet MI	BCIC IV 2A	TUEV	MDD MAL	Sleep EDF
LaBraM(LP)	23k	205k	71k	103k	76k	61k
LaBraM(LoRA)	118k	301k	167k	200k	172k	157k
CBraMod(LP)	22k	204k	70k	102k	76k	60k
CBraMod(LoRA)	137k	320k	185k	217k	191k	175k
CSBrain(LP)	22k	205k	70k	102k	76k	60k
CSBrain(LoRA)	157k	340k	205k	236k	210k	195k

Table 6: Model parameters for LoRA and Linear probe (LP) settings across foundation models for evaluation datasets.

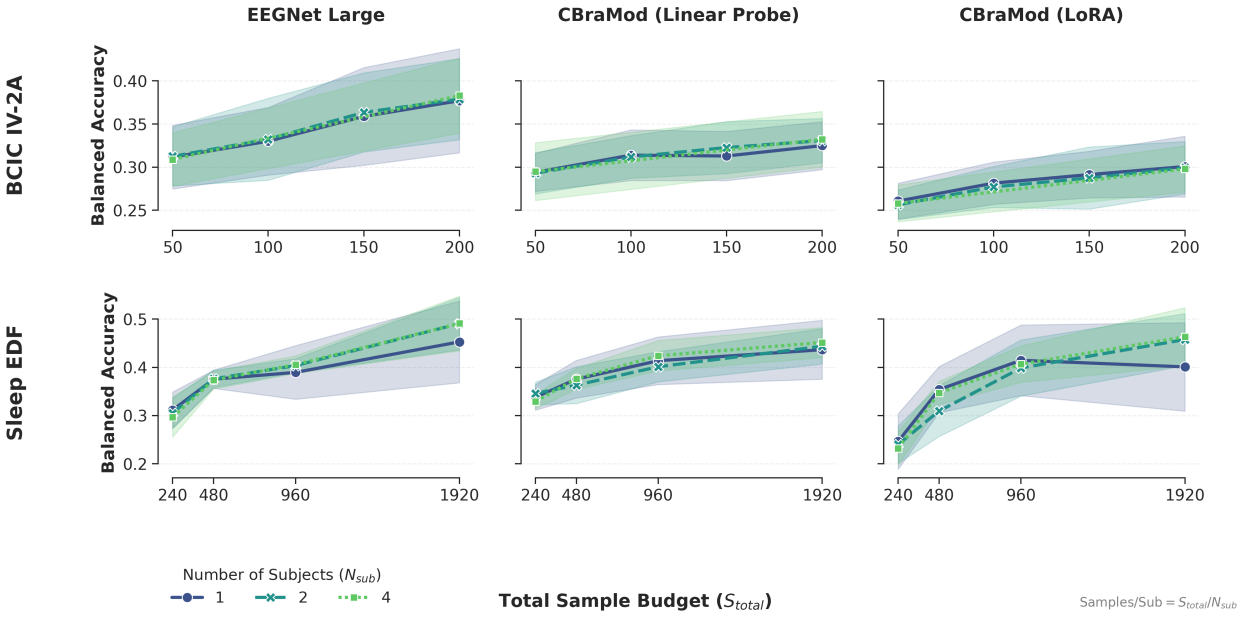

Figure 4: Total Budget sampling plot for CBramod under linear probe and LoRA settings compared to EEGNet Large

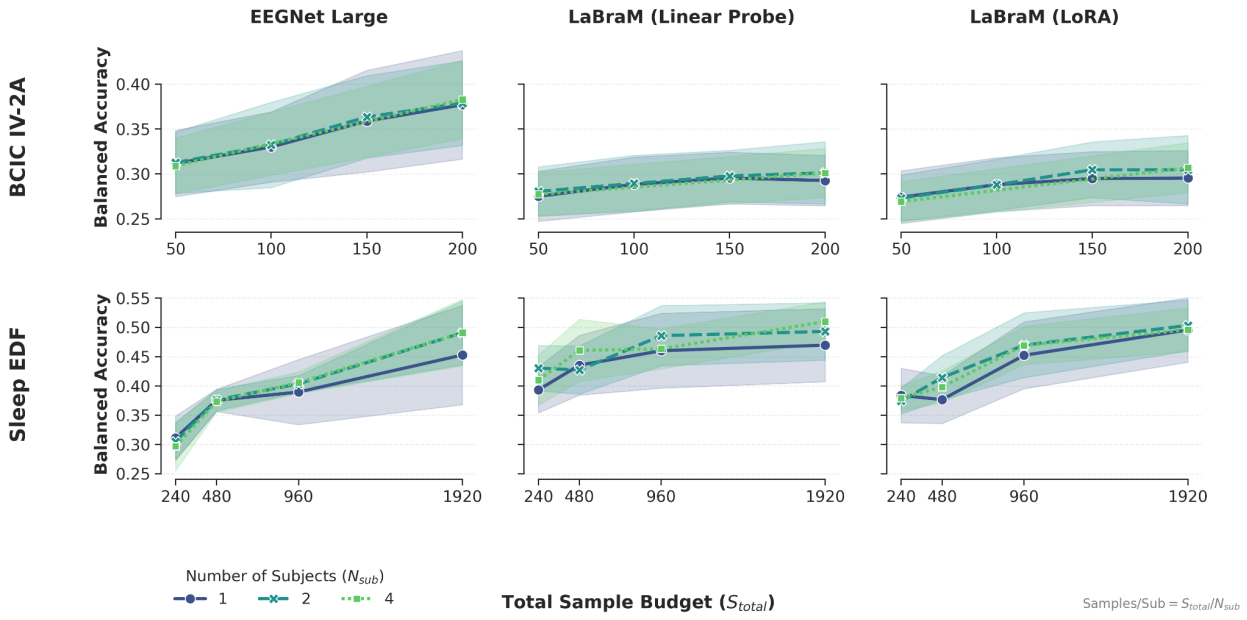

Figure 5: Total Budget sampling plot for LaBraM under linear probe and LoRA settings compared to EEGNet Large.

Appendix F. Full Data Results

Classification results for each evaluation dataset on all the listed models can be found below:

- Physionet MI: Table 7
- BCIC IV-2A: Table 8
- Kaggle ERN: Table 9
- MDD MAL: Table 10
- Sleep EDF: Table 11
- TUEV: Table 12

Appendix G. Sampling Rate Ablations

Sampling rate ablations for supervised model performance between native dataset and the resampled data to match preprocessing steps of foundation models. Results are as reported in Table 13.

Appendix H. MDD MAL Sample Efficiency Results

Sample efficiency results for MDD MAL dataset are reported in Table 14

Appendix I. MDD MAL Channel Efficiency

Additional channel efficiency results for MDD MAL dataset are reported in Table 15.

Model	BAC	F1-Macro	κ
`Supervised Models`			
EEGNet	61.34 ± 1.91	61.40 ± 1.96	48.45 ± 2.54
EEGNet Large	61.85 ± 1.95	61.83 ± 2.02	49.14 ± 2.59
EEGNet Huge	61.74 ± 1.75	61.70 ± 1.81	49.00 ± 2.34
EEGNeX	$\mathbf{65.58 \pm 1.73}$	$\mathbf{65.65 \pm 1.82}$	$\mathbf{54.13 \pm 2.30}$
SparcNet	62.02 ± 1.54	61.85 ± 1.57	49.37 ± 2.05
`LaBraM`			
Full-finetune	57.25 ± 1.55	57.22 ± 1.62	43.02 ± 2.05
Linear Probe	48.82 ± 1.31	48.74 ± 1.30	31.77 ± 1.77
LoRA	49.10 ± 1.32	49.06 ± 1.30	32.16 ± 1.74
`CBraMod`			
Full-finetune	64.02 ± 2.52	63.96 ± 2.47	52.06 ± 3.36
Linear Probe	51.97 ± 1.63	51.84 ± 1.53	35.98 ± 2.17
LoRA	62.05 ± 2.52	61.94 ± 2.51	49.41 ± 3.36
`CSBrain`			
Full-finetune	62.34 ± 2.33	62.22 ± 2.38	49.80 ± 3.10
Linear Probe	51.81 ± 1.66	51.59 ± 1.52	35.77 ± 2.21
LoRA	58.95 ± 1.78	58.90 ± 1.72	45.28 ± 2.37

Table 7: Classification results for Physionet MI dataset for Supervised and Foundation Models. Balanced Accuracy (BAC), F1-Macro and Cohens Kappa (κ) metrics are reported. Best value under each metric is reported in bold.

Model	BAC	F1-Macro	κ
`Supervised Models`			
EEGNet	54.96 ± 7.05	53.64 ± 7.32	39.94 ± 9.40
EEGNet Large	58.14 ± 7.07	56.84 ± 7.78	44.19 ± 9.43
EEGNet Huge	58.76 ± 6.60	$\mathbf{57.86 \pm 7.20}$	45.01 ± 8.80
EEGNeX	$\mathbf{59.10 \pm 11.25}$	56.93 ± 13.38	$\mathbf{45.47 \pm 15.00}$
SparcNet	56.96 ± 10.43	55.09 ± 12.31	42.62 ± 13.91
`LaBraM`			
Full-finetune	50.58 ± 9.20	49.01 ± 10.22	34.10 ± 12.26
Linear Probe	45.25 ± 5.47	43.62 ± 6.15	27.01 ± 7.29
LoRA	42.57 ± 5.47	41.09 ± 6.06	23.43 ± 7.29
`CBraMod`			
Full-finetune	56.15 ± 8.77	54.91 ± 9.51	41.54 ± 11.69
Linear Probe	43.81 ± 5.18	42.89 ± 5.35	25.08 ± 6.91
LoRA	51.37 ± 11.95	48.08 ± 14.32	35.16 ± 15.93
`CSBrain`			
Full-finetune	55.27 ± 8.39	54.18 ± 8.82	40.35 ± 11.19
Linear Probe	46.51 ± 6.82	44.88 ± 8.19	28.68 ± 9.09
LoRA	55.83 ± 6.46	54.61 ± 6.91	41.10 ± 8.61

Table 8: Classification results for BCI Competition IV-2A dataset for Supervised and Foundation Models. Balanced Accuracy (BAC), F1-Macro and Cohens Kappa (κ) metrics are reported. Best value under each metric is reported in bold.

Model	BAC	AUROC	κ
`Supervised Models`			
EEGNet	62.74 ± 2.09	65.06 ± 1.83	27.76 ± 4.39
EEGNet Large	64.94 ± 2.23	68.41 ± 1.80	32.43 ± 4.70
EEGNet Huge	64.54 ± 1.48	68.12 ± 1.02	31.29 ± 3.06
EEGNeX	$\mathbf{65.00 \pm 1.38}$	$\mathbf{70.69 \pm 1.19}$	$\mathbf{33.49 \pm 2.55}$
SparcNet	61.70 ± 1.30	66.49 ± 2.30	24.47 ± 3.14
`LaBraM`			
Full-finetune	58.51 ± 1.79	64.39 ± 1.92	18.66 ± 2.81
Linear Probe	56.11 ± 0.97	60.89 ± 2.01	13.15 ± 2.15
LoRA	58.25 ± 1.04	63.06 ± 1.59	17.20 ± 2.58
`CBraMod`			
Full-finetune	61.47 ± 1.25	68.21 ± 1.43	24.49 ± 2.44
Linear Probe	57.72 ± 0.57	64.34 ± 1.06	17.33 ± 1.32
LoRA	60.71 ± 1.02	66.36 ± 1.20	22.24 ± 1.91
`CSBrain`			
Full-finetune	63.19 ± 1.60	70.42 ± 1.99	28.16 ± 2.30
Linear Probe	60.97 ± 1.19	65.48 ± 0.81	21.76 ± 2.84
LoRA	61.55 ± 1.20	67.49 ± 1.53	24.53 ± 1.89

Table 9: Classification results for Kaggle ERN dataset for Supervised and Foundation Models. Balanced Accuracy (BAC), AUROC and Cohens Kappa (κ) metrics are reported. Best value under each metric is reported in bold.

Model	BAC	AUROC	κ
`Supervised Models`			
EEGNet	86.89 ± 3.88	93.17 ± 3.33	72.39 ± 7.97
EEGNet Large	84.98 ± 6.82	90.98 ± 6.65	68.83 ± 13.14
EEGNet Huge	83.61 ± 7.21	90.64 ± 4.28	66.03 ± 13.78
EEGNeX	86.21 ± 8.36	93.36 ± 5.08	71.44 ± 16.67
SparcNet	79.32 ± 6.22	91.10 ± 5.92	57.74 ± 12.03
`LaBraM`			
Full-finetune	88.72 ± 3.12	95.61 ± 2.75	76.61 ± 6.71
Linear Probe	87.24 ± 4.68	93.78 ± 4.35	73.81 ± 9.08
LoRA	87.88 ± 4.30	93.67 ± 4.15	75.12 ± 9.16
`CBraMod`			
Full-finetune	83.08 ± 6.69	91.72 ± 6.25	64.94 ± 12.66
Linear Probe	80.71 ± 5.43	87.73 ± 7.42	60.72 ± 11.29
LoRA	83.95 ± 10.10	92.16 ± 7.74	67.59 ± 19.89
`CSBrain`			
Full-finetune	88.81 ± 5.41	95.19 ± 5.30	76.69 ± 10.76
Linear Probe	88.03 ± 5.46	95.40 ± 4.00	75.48 ± 11.11
LoRA	$\mathbf{89.34 \pm 5.54}$	$\mathbf{96.08 \pm 4.03}$	$\mathbf{77.99 \pm 11.05}$

Table 10: Classification results for supervised and foundation models trained on depression classification task using MDD MAL dataset. Balanced Accuracy (BAC), AUROC and Cohens Kappa (κ) metrics are reported. Best value under each metric is reported in bold.

Model	BAC	F1-Macro	κ
`Supervised Models`			
EEGNet	70.20 ± 1.29	64.58 ± 2.22	71.94 ± 2.43
EEGNet Large	71.61 ± 1.20	65.87 ± 2.26	73.51 ± 1.71
EEGNet Huge	71.14 ± 1.36	66.53 ± 1.02	74.14 ± 1.64
EEGNeX	70.31 ± 1.22	69.36 ± 1.03	77.55 ± 2.21
SparcNet	71.01 ± 2.64	68.52 ± 2.73	75.41 ± 2.87
`LaBraM`			
Full-finetune	72.86 ± 1.22	69.88 ± 2.24	75.09 ± 3.22
Linear Probe	64.85 ± 0.50	63.91 ± 0.83	69.88 ± 1.43
LoRA	65.96 ± 1.86	64.56 ± 2.36	71.16 ± 2.16
`CBraMod`			
Full-finetune	$\mathbf{74.58 \pm 0.85}$	$\mathbf{73.34 \pm 0.39}$	$\mathbf{79.63 \pm 1.36}$
Linear Probe	57.71 ± 1.19	59.50 ± 1.28	69.48 ± 0.80
LoRA	69.61 ± 2.22	67.06 ± 1.72	75.42 ± 0.86
`CSBrain`			
Full-finetune	71.17 ± 1.17	70.16 ± 0.63	76.25 ± 2.30
Linear Probe	63.44 ± 2.17	63.11 ± 1.20	67.39 ± 1.91
LoRA	72.73 ± 0.98	71.38 ± 1.69	77.80 ± 2.05

Table 11: Classification results for Sleep EDF dataset for Supervised and Foundation Models. Balanced Accuracy (BAC), F1-Macro and Cohens Kappa (κ) metrics are reported. Best value under each metric is reported in bold.

Model	BAC	F1-Macro	κ
`Supervised Models`			
EEGNet	51.67 ± 2.10	53.62 ± 1.22	$\mathbf{54.29 \pm 4.42}$
EEGNet Large	53.27 ± 3.88	$\mathbf{54.33 \pm 3.01}$	50.58 ± 3.19
EEGNet Huge	52.58 ± 1.82	51.62 ± 2.00	49.31 ± 2.01
EEGNeX	47.50 ± 3.01	48.55 ± 2.87	52.07 ± 2.93
SparcNet	52.10 ± 3.51	49.66 ± 2.15	50.63 ± 4.06
`LaBraM`			
Full-finetune	$\mathbf{57.58 \pm 1.92}$	53.01 ± 3.43	50.83 ± 4.89
Linear Probe	42.81 ± 2.35	38.39 ± 1.64	34.71 ± 2.42
LoRA	43.94 ± 2.53	39.50 ± 1.73	35.17 ± 3.63
`CBraMod`			
Full-finetune	54.70 ± 2.34	50.46 ± 1.74	50.85 ± 1.65
Linear Probe	37.40 ± 1.60	36.66 ± 2.58	36.91 ± 3.53
LoRA	52.77 ± 1.57	50.55 ± 1.32	47.49 ± 2.38
`CSBrain`			
Full-finetune	51.88 ± 2.16	45.38 ± 3.55	44.38 ± 4.16
Linear Probe	47.35 ± 3.15	42.54 ± 2.79	41.76 ± 4.94
LoRA	52.41 ± 2.61	47.71 ± 2.82	50.47 ± 2.10

Table 12: Classification results for TUEV dataset for Supervised and Foundation Models. Balanced Accuracy (BAC), F1-Macro and Cohens Kappa (κ) metrics are reported. Best value under each metric is reported in bold.

Model	Sampling Rate	Physionet MI	BCIC IV-2A	MDD MAL	Sleep-EDF
EEGNet Large	200 Hz	62.8	57.3	84.5	69.7
	Native	61.9	58.1	85.0	71.6
EEGNet Small	200 Hz	62.5	54.2	83.1	68.8
	Native	61.3	54.9	86.8	70.2
EEGNet Huge	200 Hz	61.5	60.2	84.5	68.9
	Native	61.7	58.7	83.6	71.1

Table 13: Sampling rate ablations for model performance between native sampling rate of datasets compared to sampling rate of foundation models (200Hz). Results indicate no noticeable impact of sampling rate on the model performance.

Setting	Budget	Supervised	Foundation Models		
		EEGNet	**LaBraM**	**CBraMod**	**CSBrain**
Linear Probe (LP)	40	49.1 (8.8)	63.1 (16.5)	60.3 (11.8)	69.1 (12.7)
	80	56.2 (10.1)	72.7 (13.7)	68.3 (8.5)	67.1 (19.3)
	160	66.2 (12.6)	80.6 (11.6)	73.1 (10.5)	77.0 (11.9)
	320	78.1 (4.9)	86.7 (5.1)	75.6 (7.4)	82.8 (9.7)
	640	84.6 (6.0)	88.0 (4.8)	77.9 (5.7)	86.4 (6.3)
LoRA	40	49.1 (8.8)	62.7 (13.7)	60.4 (13.6)	61.1 (17.2)
	80	56.2 (10.1)	69.6 (18.0)	62.5 (13.1)	75.1 (10.0)
	160	66.2 (12.6)	79.5 (11.1)	67.4 (11.7)	76.5 (12.5)
	320	78.1 (4.9)	83.7 (9.0)	76.9 (14.0)	82.5 (8.4)
	640	84.6 (6.0)	88.6 (4.3)	81.7 (9.6)	88.4 (5.9)

Table 14: Sample efficiency results of Linear Probe (LP) and LoRA setting for MDD MAL dataset. Results indicate that EEG foundation models provide noticeable performance gains compared to supervised models under low samples for long window tasks. EEGNet does not have Linear Probe or LoRA it is fully supervised and the results are copied to make comparison easier for linear probe and LoRA experiments.

Model	PEFT	Central	Frontal	Midline	1 channel per lobe	2 channels per lobe
EEGNet Large	-	84.5 (7.3)	83.8 (2.9)	84.5 (6.8)	85.8 (7.0)	84.1 (6.5)
LaBraM	LP	83.5 (3.2)	86.0 (6.7)	85.7 (2.7)	89.7 (2.5)	88.7 (3.2)
	LoRA	84.2 (2.0)	86.0 (8.5)	85.2 (2.1)	89.0 (3.4)	89.4 (2.9)
CBraMod	LP	72.2 (10.8)	76.4 (6.4)	73.8 (12.2)	73.6 (11.2)	75.0 (7.7)
	LoRA	84.2 (7.4)	85.8 (9.3)	87.1 (3.6)	85.0 (7.1)	83.3 (9.5)
CSBrain	LP	77.2 (9.2)	82.2 (8.9)	79.2 (9.0)	78.3 (9.8)	85.4 (4.7)
	LoRA	81.8 (4.1)	82.5 (9.9)	84.1 (7.6)	85.5 (7.5)	87.5 (4.8)

Table 15: Channel efficiency experiments for MDD MAL dataset for supervised model EEGNet and foundation models (LaBraM, CBraMod and CSBrain) under linear probe and LoRA settings.