L-FAME: Longitudinal Focused Attention Meditation EEG Dataset and Benchmark

Angqi Li
Department of CMSE
Michigan State University
East Lansing, MI 48824
liangqi1@msu.edu

Ab Basit Rafi Syed
Department of CMSE
Michigan State University
East Lansing, MI 48824
syedab@msu.edu

Hamzeh Alzweri
Department of CSE
Michigan State University
East Lansing, MI 48824
alzwerih@msu.edu

Taosheng Liu
Department of Psychology
Michigan State University
East Lansing, MI 48824
tsliu@msu.edu

Barry H. Cohen
Department of Applied Psychology
New York University
New York, NY 10003
barry.cohen@nyu.edu

Saiprasad Ravishankar
Department of CMSE
Department of BME
Michigan State University
East Lansing, MI 48824
ravisha3@msu.edu

Abstract

We introduce a novel Longitudinal Focused Attention Meditation Electroencephalography (L-FAME) dataset and an accompanying benchmark, designed to foster research into the neural effects of various meditation practices and the evolution of these effects over a six-week training period. The dataset contains EEG recordings and psychological assessments from 74 healthy college participants, collected at two distinct time points: pre-intervention and post-intervention. Participants were randomly assigned to one of three distinct meditation groups: two mantra-based techniques (SA-TA-NA-MA and Hare Krishna) and one Breath Focus practice. Leveraging this unique longitudinal and comparative dataset, we propose a benchmark suite comprising three distinct classification tasks: (1) cognitive state decoding to distinguish between resting and meditation states, (2) fine-grained classification of the specific meditation techniques, and (3) cross-session adaptation to evaluate model generalization across the longitudinal time gap. We provide comprehensive baseline results for these tasks utilizing a range of classical machine learning algorithms and deep learning architectures. The complete dataset, preprocessing pipelines, and benchmark evaluation code will be publicly released, offering a valuable resource and a standardized framework for the development and comparison of new analytical methods in computational meditation research and EEG-based machine learning. https://huggingface.co/datasets/L-FAME-Dataset-Benchmark/L-FAME

1 Introduction

Brain-computer interfaces (BCIs) and neural decoding algorithms have received substantial attention as they transition into practical applications. Electroencephalography (EEG) remains the preferred modality for these tasks due to its non-invasive nature, high temporal resolution, cost-effectiveness, and portability [1]. These strengths, coupled with advancements in machine learning and deep learning architectures tailored for neural signals, have enabled breakthroughs in motor imagery and stimulus-evoked decoding [2, 3]. Beyond these traditional tasks, EEG is increasingly utilized to decode internally generated or spontaneous continuous cognitive states, such as sustained attention, mental fatigue, and mind-wandering. However, unlike stimulus-evoked paradigms, these spontaneous states

Preprint.

lack explicit external triggers, which makes precise objective labeling exceptionally challenging [4, 5, 6]. Consequently, critical deficiencies remain in the definition of tasks and the standardization of protocols.

Focused attention meditation (FAM) serves as both a continuous cognitive task and a clinical therapeutic tool for mental health. Research indicates that the neurobehavioral signatures of FAM are distinct from those of mind-wandering [7, 8, 9]. This distinction positions FAM as an ideal experimental proxy for evaluating whether deep learning models can capture subtle, non-stationary changes in neural topology. Furthermore, algorithmic developments in this area can support clinical meditation training and the quality control of interventions [10, 11].

The recent integration of deep learning with EEG analysis provides promising methods for objectively quantifying meditative states and facilitating personalized mental health monitoring [12, 13]. However, the practical deployment of these models requires robust generalization over time. Most existing datasets and studies on the decoding of meditation remain cross-sectional [14, 15]. While comparisons between meditators and non-meditators yield valuable insights, they cannot account for the dynamic changes within individuals during longitudinal training over weeks or months. Models trained on single-session data may achieve high accuracy in intra-subject evaluations, but they often experience catastrophic performance degradation when processing data from the same participant weeks later [16, 17, 18]. This degradation is primarily due to non-stationary temporal shifts. Furthermore, existing longitudinal datasets are often too small to support robust cross-subject generalization [19].

To address the gaps in temporal and cross-subject distribution shifts, we introduce the Longitudinal Focused Attention Meditation EEG (L-FAME) dataset. This dataset comprises high-resolution 64-channel EEG recordings from an initial cohort of 74 participants. It includes a subset of 44 individuals who completed a six-week pre- and post-intervention protocol. Rather than focusing on a single tradition, L-FAME systematically compares three distinct FAM paradigms, namely Breath Focus (BF), Hare Krishna (HK), and SA-TA-NA-MA (SA) [20]. We establish a benchmark suite that ranges from state classification to fine-grained technique classification and transfer learning. These tasks specifically assess cross-session generalization and cross-subject domain adaptation. They provide a standardized framework to evaluate how deep learning models manage non-stationary neural drifts over time.

The contributions of this work are summarized as follows:

- **A High-Quality Longitudinal EEG Dataset:** A well-documented, preprocessed, and publicly accessible dataset capturing 64-channel EEG across three FAM practices over a six-week intervention in 74 participants.
- **Novel Benchmark Suite:** We present a benchmark machine learning and deep learning (ML & DL) framework consisting of three evaluation tasks applied to a subset of our longitudinal meditation EEG dataset.
- **Psychometric Behavioral Labels:** We also provide psychological assessments collected at two time points to establish a behavioral ground truth for changes related to meditation.
- **Bridging Disciplines:** Beyond developing ML/DL for neural decoding, this dataset serves as a vital resource for the neuroscience community to investigate the neurophysiological mechanisms underlying meditation.

2 Related Work

Recent advances in EEG decoding have expanded beyond stimulus-locked tasks [21, 17] to continuous cognitive states, such as mental fatigue, affective states, and sustained attention [22, 23]. Evaluating these continuous states is crucial for the real-world deployment of brain-computer interfaces (BCIs), which require machine learning models to capture spontaneous, non-stationary neural dynamics over extended periods without explicit external triggers [24, 25]. Within this broader landscape, decoding meditative states, specifically differentiating FAM from spontaneous mind-wandering (MW), represents a unique and challenging frontier [19]. Unlike passive emotional or sensory responses, meditation involves active internal self-regulation. Consequently, this process serves as an

ideal paradigm for testing the ability of a model to track subtle, purely internally driven temporal shifts. Although recent studies have begun to explore deep learning for meditation decoding, progress remains heavily constrained by dataset limitations.

Related Datasets. Several publicly available EEG datasets incorporate meditation or related cognitive tasks. However, compared to L-FAME, they are often acquired using cross-sectional designs with smaller cohorts [26] or focus exclusively on specific sub-populations, like highly experienced monks [27], which limits the ability to evaluate training-induced temporal dynamics in the general population. A prominent example of a large-scale multimodal longitudinal dataset is PsiConnect, including EEG and fMRI data from a significant sample of participants, with a subset of 30 individuals completing an 8-week mindfulness program [15]. However, it primarily investigates the interaction between psilocybin and meditation, introducing complex pharmacological variables that may fall outside the scope of pure cognitive state decoding. While some recent pure longitudinal tracking efforts are severely bottlenecked by small sample sizes and single-technique restrictions [19], L-FAME provides a uniquely rigorous resource, by offering high-resolution recordings, robust longitudinal cohorts, and a systematic comparison of multiple FAM paradigms. Table 1 presents a comparative summary of L-FAME and other public EEG datasets in this domain.

Table 1: Comparison of the L-FAME dataset with existing publicly available EEG datasets for meditation.

Dataset	Meditation Type(s)	N	Age	Design
Delorme & Brandmeyer [26]	Focused Attention	24	31–78	Cross sectional
Wongupparaj et al. [27]	Mindfulness	60	19–70	Cross sectional
PsiConnect [15]	Mindfulness	62*	18–55	8-week intervention
Shang et al. [19]	MBSR	11	25-45	6-week intervention
L-FAME (Ours)	**Focused Attention†**	**74**	**18–32**	**6-week intervention**

* Only a subset of 30 participants underwent the meditation training. † Three different FAM techniques.

3 L-FAME Dataset

3.1 Data Acquisition

Participants. A total of 74 healthy college students (mean age: 22 years; 46 females, 28 males) were recruited from Michigan State University to participate in the study. The cohort consisted primarily of right-handed individuals (68 right-handed, 5 left-handed, 1 missing). Following recruitment, participants were randomly assigned to one of three meditation training groups: Breath Focus (BF), Hare Krishna (HK), or SA-TA-NA-MA (SA) . Collected demographic data included age, sex, and handedness. A comprehensive summary of the demographics of the participants is provided in Table 2 and the detailed participant demographics is in Appendix A.1. Detailed criteria for screening are in Appendix B.1 and descriptions of the specific meditation techniques appear in Appendix B.2.

Table 2: Participant demographics by meditation paradigm ($N = 74$). All paradigms follow the Focused Attention (FA) framework with distinct cognitive focus objects.

Category	Paradigm	Focus Object	N	Female/Male	Age (mean ± SD)
Focused Attention Meditation (FAM)	Breath Focus (BF)	Respiration	16	11 / 5	22.2 ± 3.9
	Hare Krishna (HK)	Long mantra	31	18 / 13	22.2 ± 4.2
	SA-TA-NA-MA (SA)	Simpler mantra	27	17 / 10	21.7 ± 2.7
Total			**74**	**46 / 28**	**22.0 ± 3.6**

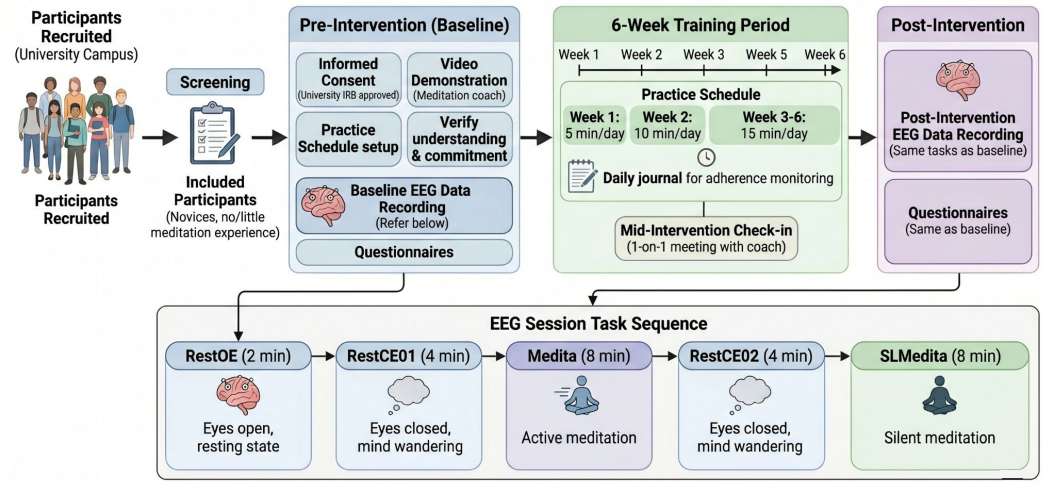

Figure 1: Overview of Longitudinal Meditation Study Design and EEG Task Sequence

Interventions. Participants were trained in specific techniques categorized under FAM. These variants included focusing on sensations created by breathing, referred to here as Breath Focus, and Japa (mantra-based meditation), which utilized the mantras SA-TA-NA-MA and Hare Krishna.

EEG Recording setup. EEG data were recorded in a controlled environment using a *mBrainTrain Smarting Pro X* amplifier (mBrainTrain, Belgrade, Serbia) and a 64-channel cap manufactured by *EASYCAP* (EASYCAP, Hersching, Germany). The cap featured Ag/AgCl electrodes arranged according to the international 10-10 system layout. To ensure high signal fidelity, the FCz electrode served as the reference, while FPz was utilized as the ground. High-chloride abrasive electrolyte gel (abralyt HiCl gel) was applied to achieve stable conductivity, with electrode impedances strictly maintained below 20 $k\Omega$ throughout the session.

3.2 Experimental Paradigms

In this section, we provide an overview of the experimental paradigms used in our study, which consists of a structured longitudinal design and a standardized EEG session protocol. Participants were screened, as detailed in Appendix B.1, and were guided through a consistent daily practice schedule, detailed in Appendix B.2. EEG sessions conducted before and after the 6-week intervention included a sequence of resting and meditation tasks, allowing for evaluation of both state- and trait-level neural changes associated with different meditation practices. The full design is shown in Figure 1.

3.2.1 Longitudinal Design

Screening. Potential adult novice participants were recruited primarily from Michigan State University (MSU) campus and underwent an initial screening process based on specific inclusion and exclusion criteria (see Section B.1). This ensured participants had little to no prior meditation experience.

Pre-Intervention (Baseline). Following successful screening and before the start of the 6-week training period, participants attended a baseline session. Upon arrival, they first read and signed the MSU Institutional Review Board (IRB) approved consent form. Next, participants were shown a pre-recorded video by the study's meditation coach demonstrating the meditation practice corresponding to their assigned group. Researchers then verified that the participants understood the practice and could commit to the daily schedule. Finally, the participants completed all baseline data collection, including an EEG recording session detailed in Section 3.2.2 and questionnaires (see Appendix B.4). After all tests are done, participants were instructed to practice daily following a gradually increasing schedule (see Appendix B.2).

Mid-Intervention (Check-in). After three weeks of practice, participants scheduled a one-on-one meeting with the meditation coach to discuss their progress and address any questions.

Post-Intervention. Within one week following the completion of the 6-week training period, participants returned for a post-intervention session. During this session, they performed the same tasks as in the baseline, including the EEG data recording detailed in Section 3.2.2.

3.2.2 EEG Session Tasks

During both the baseline and post-intervention EEG sessions, participants completed a standardized sequence of five tasks. The protocol comprised a two-minute resting state with eyes open (restOE), followed by a four-minute resting state with eyes closed (restCE01) during which participants were instructed to allow their mind to wander. Subsequently, participants engaged in an eight-minute active meditation (Medita). Because meditative practices frequently utilize overt physical or vocal engagement to facilitate initial focus, this active phase was incorporated despite the introduction of artifacts. Following a second four-minute resting state (restCE02) utilizing identical instructions, the session concluded with an eight-minute silent meditation (slMedita). During this final phase, participants executed the assigned techniques entirely internally without movement. This silent execution demands greater cognitive focus while providing the artifact-free neural data required for the primary analyses. Instructions were provided orally by the research assistant; comprehensive details regarding the specific task execution and the synchronization of event markers are available in Appendix B.3.

3.3 Data Structure and Format

The dataset is organized and formatted in accordance with the Brain Imaging Data Structure (BIDS) standard, specifically the BIDS-EEG extension [28]. Adherence to BIDS ensures standardization, facilitates data sharing and reuse, and promotes interoperability with a diverse range of analysis software, e.g., MNE-Python, EEGLAB, and FieldTrip. Following the BIDS guidelines, each data file is accompanied by a corresponding metadata file in JSON format.

Table 3: Overview of dataset structure and processing derivatives.

Data type	Folder	Participant ID	Session	Modality
Defaced raw data	/L-FAME	/sub-xx	/ses-premedita /ses-posmedita	BrainVision (.eeg, .vhdr, .vmrk)
Training data	./derivatives/ml_preproc_data	/sub-xxx	N/A	Tensors (.npy)
Cleaned EEG	./derivatives/eeglab_preproc	/sub-xxx	/ses-premedita /ses-posmedita	EEGlab (.set)

The de-identified raw EEG data are provided in BrainVision format, with all supplementary information, such as electrode locations and event markers, complying with the BIDS guidelines. Furthermore, the continuous data are pre-segmented into five tasks (and saved in the EEGLAB format. This structure allows researchers to apply any preferred preprocessing methods for subsequent analysis. In addition, a cleaned EEG dataset is provided, which underwent the preprocessing pipeline described below for immediate further analysis. Detailed preprocessing steps are included in Appendix C.1.

Cleaned EEG. Meditation recordings were partitioned into five segments defined in Section 3.2.2. To ensure signal quality for machine learning, we implemented a preprocessing pipeline focusing on automated artifact suppression and signal retention. This included zero-phase high-pass filtering [29], Zapline-plus line-noise removal [30], and Artifact Subspace Reconstruction (ASR) [31]. Neural activity was isolated via Independent Component Analysis (ICA) with ICLabel classification [32]. We provide both pre-ICA and IC-removed versions (extensions -preica and -icrm) to support diverse research requirements.

Training Data. This subset establishes benchmark tasks' training and testing data for ML/DL following established preprocessing protocols [33, 34]. Processing involved a 0.5 Hz FIR high-pass filter to attenuate low-frequency drifts and automated bad-channel identification via the eeglab algorithm ('clean_raw'). Spherical spline interpolation was applied to ensure uniform spatial dimensions across all recordings.

4 Benchmark

We define three core benchmark tasks (Task 1, Task 2, and Task 3) to evaluate model capacity and generalization on the L-FAME dataset. Continuous EEG recordings are segmented into four-second samples using sliding windows. We use a 50% overlap for cross-subject evaluations and 75% for intra-subject evaluations to augment data samples and ensure model convergence.Three progressive evaluation strategies measure generalization. Intra-subject evaluation measures individual-level stability via a block-wise interleaved split to mitigate temporal concept drift. Inter-subject evaluation employs a 5-fold cross-validation strictly partitioned at the subject level. In each fold, models are trained on data from 80% of the subjects and evaluated on the remaining 20% of completely unseen subjects, preventing data leakage and rigorously assessing population-level generalization. Leave-one-subject-out (LOSO) cross-validation provides a strict test for subject-independent generalization.This framework isolates performance aspects: intra-subject evaluation establishes a theoretical upper bound, inter-subject evaluation offers a baseline for model iteration, and LOSO simulates clinical scenarios with unseen subjects to expose inter-individual variability. A dynamic weighted random sampler ensures class balance; detailed technical specifications for the block-wise interleaved split, sampling, and benchmark tasks are provided in Appendices C.2 and C.3.

Table 4: **Data Statistics and Evaluation Protocols per Benchmark Task.** We detail the number of participants (N), total samples, and the specific validation strategies used for each task. Note that N varies due to longitudinal attrition ($N = 74$ at pre-intervention vs. $N = 44$ at post-intervention).

Task & Description	N	Total Time	Intra	Inter	LOSO
Task 1: Cognitive State Decoding (Rest vs. Med.)	74	14h 48mins†	✓	✓	✓
Task 2 (Pre): Technique Classification	74	9h 52mins†	–	✓	–
Task 2 (Post): Technique Classification	44	5h 52mins†	–	✓	–
Task 3: Cross-Session Adaptation	44	8h 48mins†	✓	–	–

† Approximation total duration

4.1 Task 1: Cognitive State Decoding (Rest vs. Focused Attention)

This task evaluates if EEG features can robustly distinguish between closed-eye resting (restCE01) and focused attention meditation (slMedita). Closed-eye resting serves as a proxy for mind wandering, as participants were explicitly instructed to allow their thoughts to wander freely during this phase (Appendix B.3). This task assesses model capacity to differentiate between undirected mind wandering and directed cognitive focus. We utilize data exclusively from the initial pre-intervention session to isolate acute physiological state differences and prevent confounding from long-term neuroplastic changes. This selection maximizes inter-subject diversity by including the full cohort before post-intervention attrition. Furthermore, the volume of continuous recordings provides sufficient data density for model convergence. Classification is evaluated using intra-subject, inter-subject, and leave-one-subject-out frameworks. We compare a global model against technique-specific models for HK, SA, and BF subgroups to determine if isolating specific meditation techniques effectively reduces cross-subject variance.

4.2 Task 2: Fine-Grained Technique Classification and Longitudinal Tracking

This task classifies specific meditation techniques: HK, SA, or BF, based on EEG recordings captured during practice. Intra-subject and leave-one-subject-out (LOSO) evaluations are logically excluded because each participant practices a single technique. Since individual data lack multi-class labels,

and LOSO test sets would render standard performance metrics undefined, we utilize an inter-subject framework via 5-fold cross-validation. To investigate longitudinal effects, we perform this classification using pre-intervention recordings ($N = 74$) and post-intervention follow-ups ($N = 44$). Comparing performance across these temporal milestones quantifies whether long-term training induces a divergence or convergence in the neural signatures associated with different techniques.

4.3 Task 3: Cross-Session Generalization and Domain Adaptation

This task evaluates model robustness across a six-week longitudinal gap, addressing signal non-stationarity and temporal concept drift. We utilize Task 1 intra-subject models, trained on pre-intervention data, within a transfer learning framework. To evaluate temporal drift independently of inter-subject variability, the evaluation is restricted to an intra-subject paradigm across two tiers

Zero-Shot Generalization Pre-trained models from Task 1 are applied directly to unseen post-intervention data to establish a baseline for six-week temporal shift degradation.

Few-Shot Calibration Pre-trained models are fine-tuned using a minimal, class-balanced calibration subset of post-intervention data. Subsequent evaluation on remaining data quantifies performance recovery.

This benchmark identifies the nature of longitudinal drift. Rapid recovery after calibration suggests fundamental stability of neural representations, attributing degradation to superficial domain shifts like electrode placement or impedance. Conversely, failure to recover indicates that the six-week intervention induced neuroplasticity, fundamentally altering functional neural representations and rendering historical decision boundaries obsolete.

5 Baseline Models

To comprehensively benchmark the L-FAME dataset and provide a standard reference for the performance of the proposed tasks, we evaluate a diverse set of baseline models. These models are categorized into traditional machine learning approaches and end-to-end deep learning architectures.

5.1 Classification Methods

Traditional Machine Learning For the classification of mental states and meditation techniques, frequency-domain features typically serve as the most robust biomarkers. Therefore, our traditional baseline employs Power Spectral Density (PSD) features extracted across standard EEG frequency bands, such as theta, alpha, beta, and gamma. These handcrafted features are subsequently flattened and fed into standard classifiers, specifically Support Vector Machines (SVM). While we also evaluated the Filter Bank Common Spatial Pattern (FBCSP) coupled with an SVM as a spatial-domain reference, we use the features extracted from this method to train the SVMs following the intra-subject and inter-subject evaluations.

Deep Learning To evaluate automatic feature extraction without prior manual feature engineering, we benchmark four representative deep learning architectures, ranging from standard convolutional networks to advanced spatiotemporal models. First, we utilize Shallow ConvNet and Deep ConvNet, two classical architectures widely adopted in BCI research [35]. Shallow ConvNet extracts features analogous to band power, whereas Deep ConvNet utilizes a deeper hierarchy of generic convolutional layers to capture complex representations. Second, we deploy EEGNet, a compact and specialized convolutional neural network tailored for EEG signals [36]. It utilizes depthwise and separable convolutions to significantly reduce trainable parameters while maintaining robust cross-subject generalization. Third, we implement EEG-Conformer, a state-of-the-art architecture integrating the local feature extraction of convolutions with the global, long-range dependency modeling of Transformer self-attention mechanisms [37]. This provides an advanced baseline for complex spatiotemporal EEG decoding.

6 Results

Task 1: Cognitive State Decoding. Table 5 presents the fundamental capacity of various models to distinguish between the resting state and the focused attention meditation state. Deep learning architectures demonstrate exceptional proficiency in the intra-subject evaluation paradigm, with EEGNet achieving a peak AUC of $99.2\% \pm 1.2\%$, closely followed by EEG-Conformer at $97.2\% \pm 1.4\%$. However, a substantial generalization gap emerges when evaluating models across subjects. Inter-subject and LOSO evaluations yield significantly lower performance, hovering around 66% to 70%. This pronounced variance underscores the highly individualized nature of meditation-induced neural oscillations. Furthermore, traditional machine learning baselines relying on handcrafted features, such as PSD and FBCSP coupled with SVM or RF, exhibit optimal performance compared to deep models in the intra-subject evaluation case (yielding an estimated AUC of $\approx 99.9\%$). However, they underperform in inter-subject and LOSO evaluations, rendering them inferior to convolutional networks in generalizing to unseen subjects. As detailed in Appendix D.1, group-specific decoding reveals that the SA TA NA MA (SA) technique provides the most generalized neural signatures, yielding the highest LOSO accuracy ($69.1\% \pm 23.6\%$ via EEGNet) among the three practices.

Table 5: **Task 1: State Decoding Performance.** Mean AUC (%) and standard deviation for Intra-subject, Inter-subject, and LOSO protocols ($N = 74$).

Model	Intra	Inter	LOSO
PSD + SVM	97.1±4.0	59.4±3.8	61.3±24.4
FBCSP + SVM	**99.9±0.2**	55.8±6.2	58.2±21.4
ShallowConvNet	97.2±4.4	66.5±4.3	**70.4±20.9**
DeepConvNet	97.0±4.9	66.2±5.9	68.4±27.0
EEGNet	99.2±2.1	66.5±4.0	66.6±27.4
EEG-Conformer	97.0±4.9	**66.9±5.2**	67.5±24.5

Task 2: Fine-Grained Technique Classification Classifying the specific meditation technique directly from the EEG signals represents a significantly more complex challenge. As shown in Table 6, all models yield Precision-Recall AUC (PR-AUC) scores that, while exceeding the theoretical chance level of 33.3%, remain generally low. In the pre-intervention phase, ShallowConvNet achieves the highest baseline performance at $38.0\% \pm 6.2\%$. Interestingly, the post-intervention data ($N = 44$) exhibits an observable shift in classification dynamics, with EEGNet improving to $48.8\% \pm 8.6\%$. This subtle enhancement suggests that the six-week longitudinal training intervention may gradually consolidate technique-specific neural signatures, albeit the representations remain inherently difficult to isolate across different individuals. We also notice that when fitting an SVM to the features extracted using FBCSP we also see an improvement in the PR-AUC when testing the post meditation data. This suggests that the meditations reduced the noise of the signals, and made it more structured such that it might follow some patterns that are detectable by an appropriate model, to test this we must pursue the meditation for a longer period and test if this further reduces the noise and gives higher accuracies. To validate this, further ablation studies are presented in Appendix D.2.

Table 6: **Task 2: Technique Classification Performance.** PR-AUC (%) for Inter-subject evaluation (Pre: $N = 74$, Post: $N = 44$).

Model	Pre	Post
FBCSP + SVM	34.3±4.1	47.4±7.2
ShallowConvNet	**38.0±6.2**	45.7±9.7
DeepConvNet	34.5±3.4	45.8±10.7
EEGNet	35.0±2.9	**48.8±8.6**
EEG-Conformer	37.6±4.6	43.9±11.4

Task 3: Cross-Session Generalization And Longitudinal Adaptation Table 7 summarizes robustness against six-week temporal concept drift. Under zero-shot evaluation, state-decoding models degrade significantly, yielding AUCs between 63.1% and 67.2%. This indicates domain shifts, like impedance variations and neuroplasticity, disrupt intra-session

Table 7: **Task 3: Cross-Session Generalization and Longitudinal Adaptation.** Note that the k-shot calibration performance reported in this table corresponds to the full fine-tuning strategy (updating all layers).

	Task 3: Intra-subject Adaptation (AUC %)			
Model	Zero-shot	10-shot	30-shot	Upper Bound†
ShallowConvNet	63.9±16.8	75.2±17.5	78.3±16.6	95.4±7.2
DeepConvNet	63.1±22.7	72.8±20.2	77.9±18.4	97.0±5.8
EEGNet	63.1±17.8	**76.5±21.2**	**79.3±21.6**	95.4±8.3
EEG-Conformer	**67.2±16.5**	74.5±18.1	77.2±18.0	93.3±7.9

† representing the performance ceiling achieved through full intra-blockwise training.

decision boundaries. To mitigate this, we compared full fine-tuning (updating all weights) and linear fine-tuning (updating only Batch Normalization layers). With just 30 samples, full fine-tuning recovers EEGNet AUC to 79.3%. This rapid recovery implies stable underlying representations, attributing initial losses to superficial signal shifts. Nevertheless, a 16% to 18% gap from the intra-session upper bound persists. This deficit likely reflects deeper longitudinal neuroplasticity, highlighting opportunities for enhanced fine-tuning to fully capture these adaptations.

7 Discussion and Future Work

The L-FAME benchmark provides a standardized framework to evaluate EEG-based models across three progressively challenging tasks: the decoding of cognitive states (Task 1), fine-grained technique classification (Task 2), and cross-session generalization (Task 3). Across all tasks, a consistent pattern emerges: current models achieve high intra-subject performance but degrade significantly when generalizing across subjects or time. Although Task 1 intra-subject AUC exceeds 97%, cross-subject performance plateaus between 66% and 70%. This gap of approximately 30% reveals a fundamental discrepancy between standard laboratory evaluations and real-world deployment requirements. Furthermore, Task 2 remains an unresolved cross-subject challenge even after the intervention. Task 3 indicates that longitudinal drift is substantial but remains largely correctable through minimal calibration. Appendices D.1 to D.3 provide detailed subgroup analyses, representational geometry experiments, and adaptation ablations for each task.

These results align with the established perspective that meditation-induced EEG states exhibit strong within-individual signatures [19, 7]. They also confirm that inter-individual neural variability [24, 38] and temporal non-stationarity [25] remain primary bottlenecks. The difficulty of cross-subject technique classification appears to conflict with prior studies reporting technique-specific EEG signatures [9, 39, 40]. However, those studies primarily established group-level effects by aggregating within-subject relative changes, rather than evaluating generalized decision boundaries via end-to-end cross-subject classification. This distinction highlights a critical gap between statistical neurophysiological characterization and deploying robust decoding models. Notably, post-intervention representational geometry analyses (UMAP, RSA) demonstrate that six weeks of training initiates the differentiation of technique-specific neural representations. Specifically, the similarity margin between within-class and between-class representations expands from 0.30 to 1.10 (Appendix D.2). This expansion suggests that while these signatures are emerging, they remain unstable across subjects. An informative asymmetry also emerges: Breath Focus, which relies on somatic attention, generates more separable cross-subject representations than the two mantra-based techniques. In contrast, HK and SA representations remain mutually confusable after the intervention. This observation implies that the underlying cognitive substrate (somatic versus phonological) is potentially more determinative of cross-subject decoding performance than the specific content of the mantra [3, 20].

The L-FAME benchmark highlights three concrete open problems for the research community. First, the persistent performance gap between intra-subject and cross-subject evaluations motivates the development of domain generalization and transfer learning methods. These methods must be tailored to spontaneous, internally driven cognitive states, which are poorly supported by existing benchmarks based on motor imagery or stimulus-evoked paradigms. Second, the near-chance performance in the cross-subject classification of techniques indicates that general-purpose architectures lack the inductive biases required to capture technique signatures that remain invariant across individuals. The integration of meditation/cognition-specific priors may be necessary. Third, Task 3 demonstrates that models can transfer over time despite the expected longitudinal EEG drift. This drift can be corrected using as few as 30 calibration samples through full fine-tuning (Appendix D.3), which presents a practical pathway toward the efficient deployment of longitudinal brain-computer interfaces without complete model retraining. More broadly, the L-FAME benchmark serves as a standardized resource for the investigation of training-induced neuroplasticity and the personalized monitoring of meditation. Both directions hold increasing relevance for the field of computational mental health [13, 12].

8 Conclusion and Limitations

We introduced L-FAME, a novel 6-week longitudinal EEG dataset and benchmark designed to evaluate neural decoding across three focused attention meditation practices. Our extensive baselines

reveal a critical insight: while deep learning models achieve exceptional intra-subject accuracy, they struggle significantly with cross-subject generalization due to inherent inter-individual variability. Furthermore, we demonstrated that longitudinal temporal drift is substantial but can be effectively mitigated via minimal few-shot calibration.

Limitations & Future Work: Although attrition reduced the longitudinal cohort to 44 participants, statistical evaluations confirm that dropouts occurred completely at random (MCAR), thereby preserving the overall integrity of the sample (see Appendix A.2). Nonetheless, reliance on self-reported practice adherence, sparse temporal sampling (two discrete time points), the lack of a control group, and a restricted college-aged demographic limit the broader clinical generalizability of the findings. We plan to continue data collection in future work to supplement the dataset with appropriate control cohorts. We anticipate that this open-access resource will facilitate the development of meditation-specific architectures and robust domain-adaptation methods for continuous neurophysiological monitoring.

References

[1] Ahmad Chaddad, Yihang Wu, Reem Kateb, and Ahmed Bouridane. Electroencephalography signal processing: A comprehensive review and analysis of methods and techniques. *Sensors*, 23(14):6434, 2023.

[2] Tarik S Bel-Bahar, Anam A Khan, Riaz B Shaik, and Muhammad A Parvaz. A scoping review of electroencephalographic (eeg) markers for tracking neurophysiological changes and predicting outcomes in substance use disorder treatment. *Frontiers in human neuroscience*, 16:995534, 2022.

[3] Robert H Logie, Annalena Venneri, Sergio Della Sala, Thomas W Redpath, and Ian Marshall. Brain activation and the phonological loop: The impact of rehearsal. *Brain and Cognition*, 53(2):293–296, 2003.

[4] Jiaxing Fan, Lin Dong, Gang Sun, and Zhize Zhou. A deep learning approach for mental fatigue state assessment. *Sensors*, 25(2):555, 2025.

[5] Xi Liu, Pang-Ning Tan, Lei Liu, and Steven J Simske. Automated classification of eeg signals for predicting students' cognitive state during learning. In *Proceedings of the international conference on web intelligence*, pages 442–450, 2017.

[6] Shaohua Tang, Yutong Liang, and Zheng Li. Mind wandering state detection during video-based learning via eeg. *Frontiers in human neuroscience*, 17:1182319, 2023.

[7] Julio Rodriguez-Larios, Kian Foong Wong, and Julian Lim. Assessing the effects of an 8-week mindfulness training program on neural oscillations and self-reports during meditation practice. *Plos one*, 19(6):e0299275, 2024.

[8] Zongpai Zhang, Wen-Ming Luh, Wenna Duan, Grace D Zhou, George Weinschenk, Adam K Anderson, and Weiying Dai. Longitudinal effects of meditation on brain resting-state functional connectivity. *Scientific reports*, 11(1):11361, 2021.

[9] Angqi Li, Julio Rodriguez-Larios, Mengsen Zhang, Taosheng Liu, Barry H Cohen, and Saiprasad Ravishankar. Not all mantra meditations are equal: Emergence of divergent alpha oscillatory dynamics across mantras. *bioRxiv*, pages 2026–02, 2026.

[10] Emilee E Burgess, Steven Selchen, Benjamin D Diplock, and Neil A Rector. A brief mindfulness-based cognitive therapy (mbct) intervention as a population-level strategy for anxiety and depression. *International Journal of Cognitive Therapy*, 14(2):380–398, 2021.

[11] Bassam Khoury, Manoj Sharma, Sarah E Rush, and Claude Fournier. Mindfulness-based stress reduction for healthy individuals: A meta-analysis. *Journal of psychosomatic research*, 78(6):519–528, 2015.

[12] RuiFang Lyu. Deep learning approaches for eeg-based healthcare applications: a comprehensive review. *Frontiers in Human Neuroscience*, 19:1689073, 2026.

[13] Zixiang Liu and Juan Zhao. Leveraging deep learning for robust eeg analysis in mental health monitoring. *Frontiers in neuroinformatics*, 18:1494970, 2025.

[14] Tracy Brandmeyer, Arnaud Delorme, and Helané Wahbeh. The neuroscience of meditation: classification, phenomenology, correlates, and mechanisms. *Progress in Brain Research*, 244:1–29, 2019.

[15] Leonardo Novelli, Devon Stoliker, Tamrin Barta, Matthew D Greaves, Sidhant Chopra, James Jackson, Jessica Kwee, Martin L Williams, and Adeel Razi. Psiconnect: A multimodal neuroimaging study of psilocybin-induced changes in brain and behaviour. *bioRxiv*, pages 2025–04, 2025.

[16] Yi-Yuan Tang, Britta K Hölzel, and Michael I Posner. The neuroscience of mindfulness meditation. *Nature Reviews Neuroscience*, 16(4):213–225, 2015.

[17] Vinay Jayaram and Alexandre Barachant. Moabb: trustworthy algorithm benchmarking for bcis. *Journal of neural engineering*, 15(6):066011, 2018.

[18] Fabien Lotte, Marco Congedo, Anatole Lécuyer, Fabrice Lamarche, and Bruno Arnaldi. A review of classification algorithms for eeg-based brain–computer interfaces. *Journal of neural engineering*, 4(2):R1–R13, 2007.

[19] Baoxiang Shang, Feiyan Duan, Ruiqi Fu, Junling Gao, Hinhung Sik, Xianghong Meng, and Chunqi Chang. Eeg-based investigation of effects of mindfulness meditation training on state and trait by deep learning and traditional machine learning. *Frontiers in human neuroscience*, 17:1033420, 2023.

[20] Antoine Lutz, Heleen A Slagter, John D Dunne, and Richard J Davidson. Attention regulation and monitoring in meditation. *Trends in cognitive sciences*, 12(4):163–169, 2008.

[21] Luis Fernando Nicolas-Alonso and Jaime Gomez-Gil. Brain computer interfaces, a review. *sensors*, 12(2):1211–1279, 2012.

[22] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

[23] Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on autonomous mental development*, 7(3):162–175, 2015.

[24] Gan Huang, Zhiheng Zhao, Shaorong Zhang, Zhenxing Hu, Jiaming Fan, Meisong Fu, Jiale Chen, Yaqiong Xiao, Jun Wang, and Guo Dan. Discrepancy between inter-and intra-subject variability in eeg-based motor imagery brain-computer interface: Evidence from multiple perspectives. *Frontiers in neuroscience*, 17:1122661, 2023.

[25] Dean J Krusienski, Moritz Grosse-Wentrup, Ferran Galán, Damien Coyle, Kai J Miller, Elliott Forney, and Charles W Anderson. Critical issues in state-of-the-art brain–computer interface signal processing. *Journal of neural engineering*, 8(2):025002, 2011.

[26] Tracy Brandmeyer and Arnaud Delorme. Reduced mind wandering in experienced meditators and associated eeg correlates. *Experimental brain research*, 236:2519–2528, 2018.

[27] Peera Wongupparaj. Eeg absolute and relative powers during mindfulness meditation: Data from thai buddhist monks. *Mendeley Data V1*, 2024.

[28] Krzysztof J Gorgolewski, Tibor Auer, Vince D Calhoun, R Cameron Craddock, Samir Das, Eugene P Duff, Guillaume Flandin, Satrajit S Ghosh, Tristan Glatard, Yaroslav O Halchenko, et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific data*, 3(1):1–9, 2016.

[29] Andreas Widmann, Erich Schröger, and Burkhard Maess. Digital filter design for electrophysiological data – a practical approach. *Journal of Neuroscience Methods*, 250:34–46, 2015.

[30] Alain de Cheveigné. Zapline-plus: a flexible and easy-to-use tool for automatic and robust removal of power line artifacts. *NeuroImage*, 216:116561, 2020.

[31] Tim Mullen. Cleanrawdata: Artifact subspace reconstruction on matlab. EEGLAB plugin, 2012.

[32] Laura Pion-Tonachini, Ken Kreutz-Delgado, and Scott Makeig. Iclabel: an automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, 198:181–197, 2019.

[33] Arnaud Delorme. Eeg is better left alone. *Scientific reports*, 13(1):2372, 2023.

[34] Roman Kessler, Alexander Enge, and Michael A Skeide. How eeg preprocessing shapes decoding performance. *Communications Biology*, 8(1):1039, 2025.

[35] 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. *Human brain mapping*, 38(11):5391–5420, 2017.

[36] 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.

[37] Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022.

[38] Saskia Haegens, Helena Cousijn, George Wallis, Paul J Harrison, and Anna C Nobre. Inter-and intra-individual variability in alpha peak frequency. *Neuroimage*, 92:46–55, 2014.

[39] Bianca Ventura, Yasir Çatal, Angelika Wolman, Andrea Buccellato, Austin Clinton Cooper, and Georg Northoff. Intrinsic neural timescales exhibit different lengths in distinct meditation techniques. *Neuroimage*, 297:120745, 2024.

[40] Julio Rodriguez-Larios, Eduardo A Bracho Montes de Oca, and Kaat Alaerts. The eeg spectral properties of meditation and mind wandering differ between experienced meditators and novices. *NeuroImage*, 245:118669, 2021.

[41] James F Cavanagh and Michael J Frank. Frontal theta as a mechanism for cognitive control. *Trends in cognitive sciences*, 18(8):414–421, 2014.

[42] Tim Lomas, Itai Ivtzan, and Cynthia HY Fu. A systematic review of the neurophysiology of mindfulness on eeg oscillations. *Neuroscience & Biobehavioral Reviews*, 57:401–410, 2015.

[43] Jean-Philippe Lachaux, Eugenio Rodriguez, Jacques Martinerie, and Francisco J Varela. Measuring phase synchrony in brain signals. *Human brain mapping*, 8(4):194–208, 1999.

[44] Sheldon Cohen. Perceived stress in a probability sample of the united states. 1988.

[45] Ernst Bohlmeijer, Peter M Ten Klooster, Martine Fledderus, Martine Veehof, and Ruth Baer. Psychometric properties of the five facet mindfulness questionnaire in depressed adults and development of a short form. *Assessment*, 18(3):308–320, 2011.

[46] Wolf E Mehling, Michael Acree, Anita Stewart, Jonathan Silas, and Alexander Jones. The multidimensional assessment of interoceptive awareness, version 2 (maia-2). *PloS one*, 13(12):e0208034, 2018.

Appendix

This appendix provides supplementary material organized into four sections. The first section presents comprehensive dataset documentation adhering to the Datasheets for Datasets framework, which encompasses detailed participant demographics (Section A.1), attrition analysis with MCAR validation (Section A.2), and the complete BIDS-compliant dataset structure alongside its processing derivatives (Section A.3). The second section details the extended experimental paradigms, including criteria for participant inclusion and exclusion (Section B.1), three meditation techniques and the corresponding training protocols (Section B.2), procedures for EEG sessions (Section B.3), and outcomes of the psychometric questionnaires (Section B.4). The third section outlines the methodology and the data processing pipeline in detail, encompassing parameters for EEG preprocessing (Section C.1), configurations for model training (Section C.2), protocols for benchmark tasks (Section C.3), and details regarding hardware and software implementations. The fourth section reports additional experimental results for all three benchmark tasks, featuring ablations of the evaluation protocols (Section D.1), longitudinal representational analyses (Section D.2), and learning curves for few-shot adaptation (Section D.3).

A Dataset Documentation (Datasheets for Datasets)

A.1 Detailed Participant Demographics

Baseline Demographic Homogeneity A total of 74 participants meeting the inclusion criteria (no history of neurological or psychiatric disorders, and no or minimal prior meditation experience) were initially recruited and randomly assigned to either the Breath Focus meditation (BF), Hare Krishna mantra meditation (HK) or SA-TA-NA-MA mantra meditation (SA) group (see in Table 8). Statistical analyses via one-way ANOVA and Pearson χ^2 tests confirm that the three groups were completely balanced at the pre-test phase. Specifically, no significant differences were observed across groups regarding age ($F(2, 71) = 0.15$, $p = 0.864$), sex distribution ($\chi^2(2) = 0.52$, $p = 0.770$), or handedness ($\chi^2(2) = 0.02$, $p = 0.988$, with one HK participant excluded from this specific metric due to missing data) in Figure 2 first column. The overall cohort primarily consisted of right-handed young adults (mean age roughly 22 years) with a slightly higher proportion of females (62.2%).

Consistent Attrition and Preserved Post-Test Balance The study experienced a notable overall attrition rate of 40.5% between the pre-test and post-test phases, reducing the final sample size to 44 participants. However, the dropout rates were distributed evenly across the intervention groups, ranging from 38.7% in the HK group to 43.8% in the BF group. Importantly, homogeneity tests on the post-test sample indicate that this attrition was non-selective and did not introduce demographic bias. The remaining participants across the three groups continued to show no significant differences in age ($F(2, 41) = 0.01$, $p = 0.991$), sex ($\chi^2(2) = 0.02$, $p = 0.992$), and handedness ($\chi^2(2) = 0.53$, $p = 0.768$) in Figure 2 second column. This preserved demographic parity across time points ensures that downstream physiological analyses (such as EEG signal evaluation) remain robust against confounding demographic variables, thereby supporting the interpretability and fairness of subsequent evaluations.

A.2 Attrition Cohorts Analysis and MCAR Validation

Baseline Demographics and Group Assignments Exhibit No Significant Correlation with Attrition To determine whether participant attrition was biased by baseline characteristics, we compared the demographic profiles of completers ($n = 44$) and dropouts ($n = 30$) across four key dimensions (Fig. 3). Pearson's χ^2 tests of independence were conducted for categorical variables, revealing no significant associations between dropout status and sex ($p = 0.366$), intervention group ($p = 0.946$), or handedness ($p = 0.601$). For the continuous variable of age, a Mann-Whitney U test indicated no significant difference in the age distribution between the two cohorts ($p = 0.424$). These results demonstrate robust baseline parity, confirming that neither biological factors nor the specific nature of the assigned meditation protocol (HK, SA, or BF) served as a primary catalyst for dropout. The absence of demographic or group-level divergence supports the assumption that the observed attrition is likely missing completely at random (MCAR) relative to these factors, validating the subsequent focus on intrinsic neurophysiological markers for predicting participant adherence.

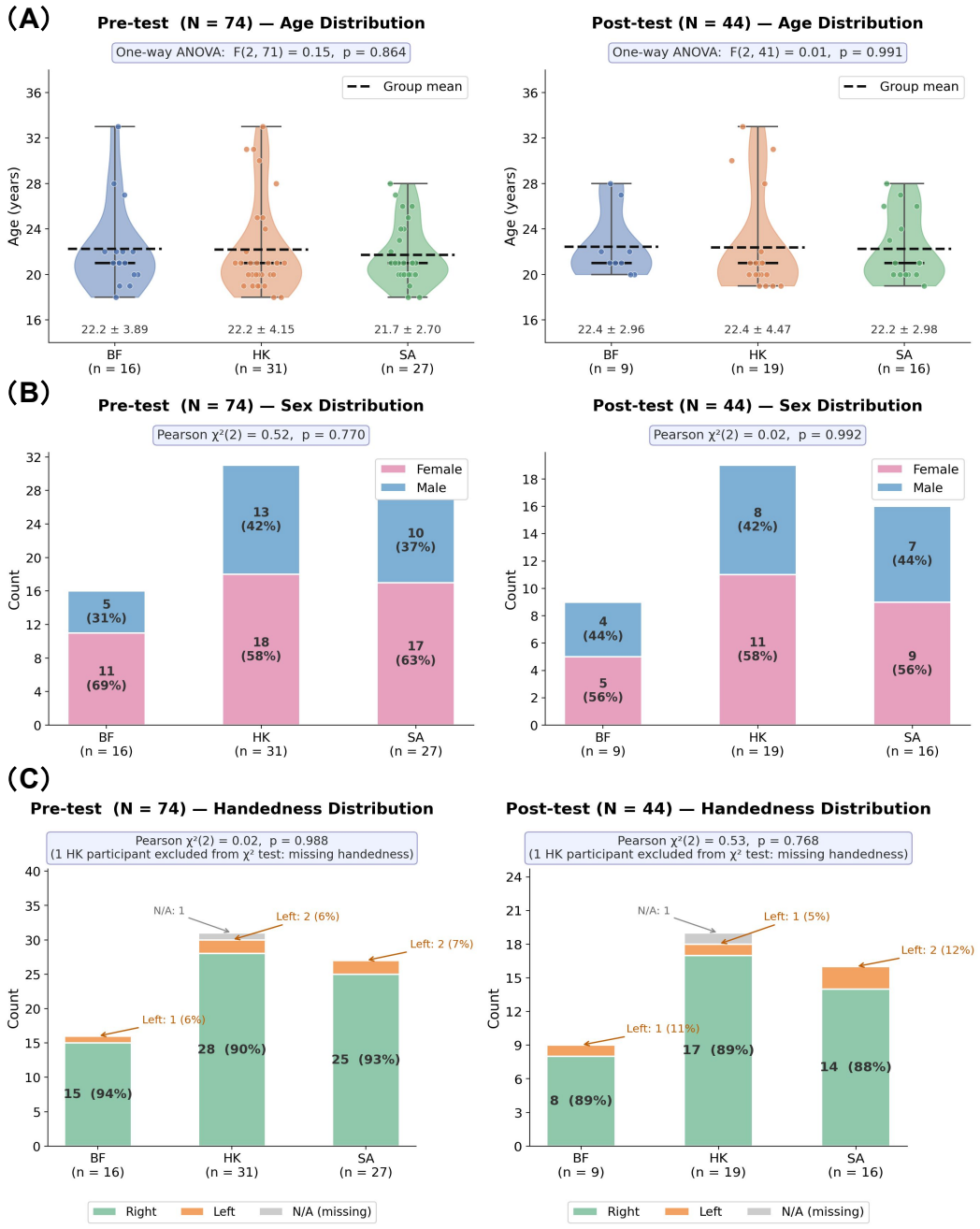

Figure 2: Distributions of (A) age, (B) sex, and (C) handedness demonstrate consistent demographic homogeneity across intervention groups at both pre-test and post-test phases.

Table 8: Participant Demographics and Homogeneity Test Results by Group and Assessment Phase

Phase	Variable	Group			Total
		BF	HK	SA	
Pre-test $N = 74$	N	16	31	27	74
	Age, M±SD	22.25 ± 3.89	22.19 ± 4.15	21.74 ± 2.70	22.04 ± 3.58
	Sex				
	Female, n (%)	11 (68.8)	18 (58.1)	17 (63.0)	46 (62.2)
	Male, n (%)	5 (31.2)	13 (41.9)	10 (37.0)	28 (37.8)
	Handedness				
	Right, n (%)	15 (93.8)	28 (90.3)	25 (92.6)	68 (91.9)
	Left, n (%)	1 (6.2)	2 (6.5)	2 (7.4)	5 (6.8)
	Attrition, n (%)	7 (43.8)	12 (38.7)	11 (40.7)	30 (40.5)
Post-test $N = 44$	N	9	19	16	44
	Age, M±SD	22.44 ± 2.96	22.37 ± 4.47	22.25 ± 2.98	22.34 ± 3.62
	Sex				
	Female, n (%)	5 (55.6)	11 (57.9)	9 (56.2)	25 (56.8)
	Male, n (%)	4 (44.4)	8 (42.1)	7 (43.8)	19 (43.2)
	Handedness				
	Right, n (%)	8 (88.9)	17 (89.5)	14 (87.5)	39 (88.6)
	Left, n (%)	1 (11.1)	1 (5.3)	2 (12.5)	4 (9.1)

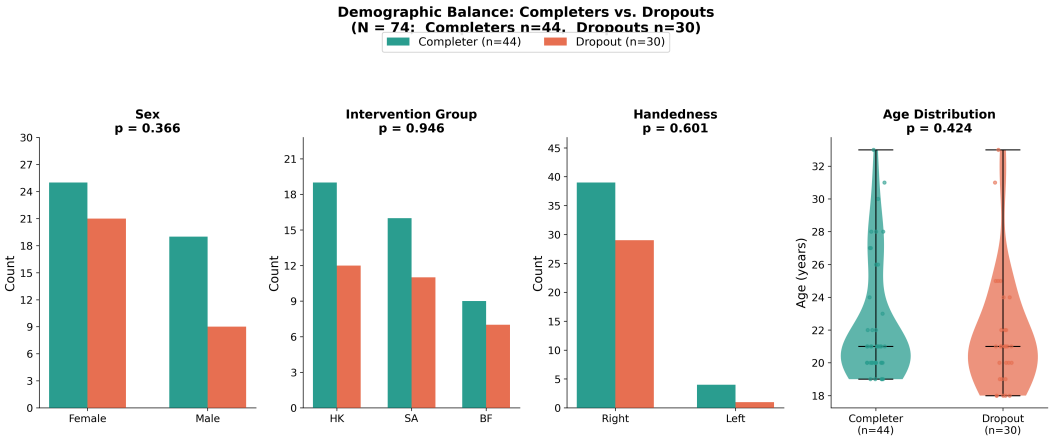

Figure 3: Analysis of Group Assignment and Demographics as Potential Causes of Dropout and Attrition, Detailing the Distributions of Sex, Group, Handedness, and Age With Corresponding p-Values for Each Analysis

Pre-Intervention EEG Oscillatory Profiles and Connectivity Fail to Predict Participant Attrition To rigorously evaluate whether baseline neurophysiological states bias experimental retention, we analyzed 30 pre-intervention EEG features across both univariate and multivariate dimensions (Fig. 5, Fig. 4). Univariate comparisons of absolute power spectral density (PSD) across five canonical frequency bands ($\delta, \theta, \alpha, \beta, \gamma$) under resting-state and active meditation conditions revealed no significant differences between completers ($n = 44$) and dropouts ($n = 30$) after FDR correction (all $q > 0.05$, Fig. 3). Similarly, frontal θ power and Phase Locking Value (PLV) for key electrode pairs (F3–F4, Fz–Pz, mean frontal) exhibited functional homogeneity between cohorts [41, 42, 43]. To assess the collective discriminative power of these features, a logistic regression classifier with ℓ_2 regularization was evaluated via 10-fold stratified cross-validation. The resulting mean AUC was 0.538 ± 0.157, which departs from the 0.50 chance baseline by only 0.038 units (Fig. 4).

The combination of non-significant univariate band-level differences and near-chance multivariate classification performance provides direct evidence against a Missing Not At Random (MNAR) mechanism. These results indicate that baseline neural representations carry effectively no predictive information regarding subsequent dropout, thereby validating the Missing Completely At Random (MCAR) assumption for the L-FAME dataset.

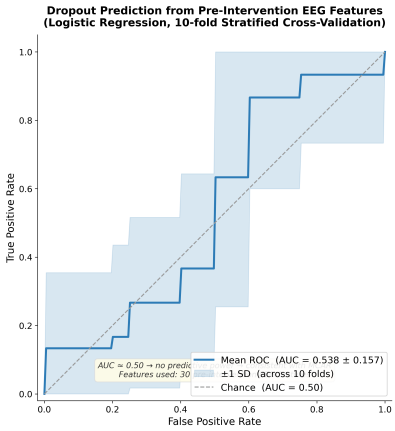

Figure 4: Receiver Operating Characteristic (ROC) curve for dropout prediction using pre-intervention EEG features.

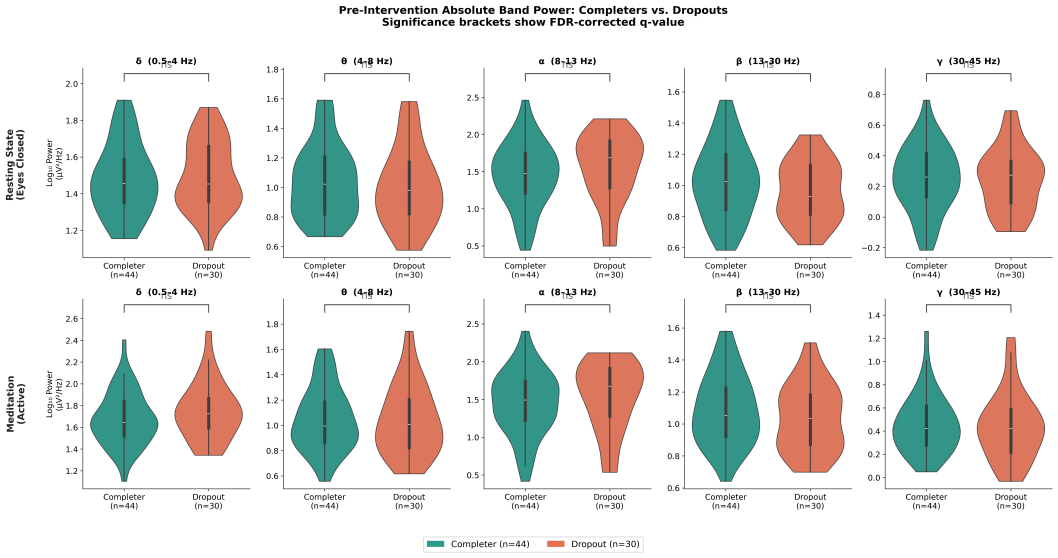

Figure 5: **Pre-intervention absolute EEG band power distributions for completers vs. dropouts.** Comparison of $\log_{10}$-transformed absolute power spectral density (PSD) between participants who completed the study ($n = 44$, teal) and those who dropped out ($n = 30$, orange). Results are shown for five canonical frequency bands under both Resting State (top row) and Active Meditation (bottom row) conditions. Significance brackets denote FDR-corrected q-values; "ns" indicates no statistically significant differences ($q > 0.05$)

A.3 Dataset Structure and Processing Derivatives

The Longitudinal Meditation Benchmark adheres to the Brain Imaging Data Structure (BIDS) v1.9.0 specification. It comprises recordings from 74 participants (`sub-01` through `sub-74`).

Root-level metadata. The root directory encapsulates standard BIDS metadata files, including `dataset_description.json` for license and software provenance, `participants.tsv` and `participants.json` for per-subject demographics (such as group, age, sex, and handedness), `croissant_metadata.json` serving as the ML dataset card, and a comprehensive `README` file.

Raw EEG recordings. The directory for each subject contains two longitudinal sessions: `ses-premedita` (pre-training) and `ses-posmedita` (post-training). As depicted in Figure 6, of the

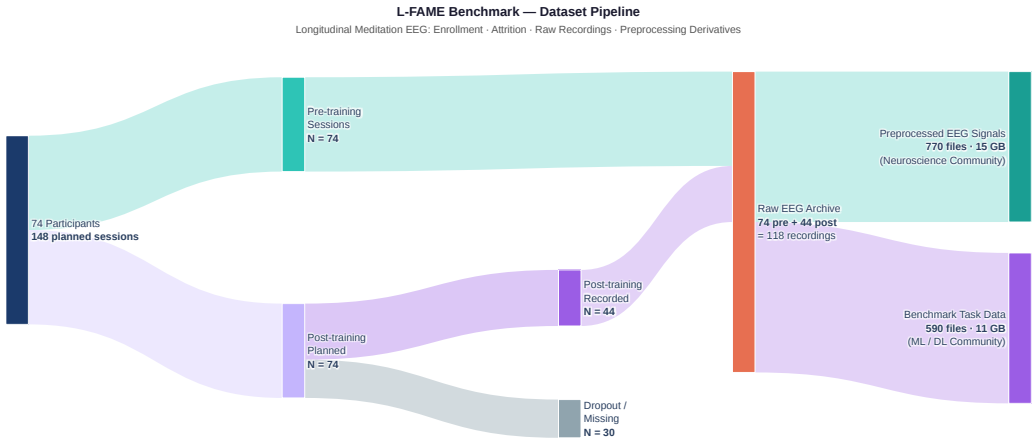

Figure 6: Diagram of the Longitudinal Meditation Benchmark pipeline. Link widths are proportional to the recording counts at each stage. The diagram illustrates the progression from raw data acquisition, accounting for participant attrition, to the generation of the two preprocessed derivative tiers.

148 planned sessions, all 74 pre-training sessions and 44 post-training sessions were successfully recorded. This attrition yields a total of 118 raw BrainVision `.eeg` files. Within each session, the EEG data are maintained in the BrainVision format, which consists of a binary data file (`.eeg`, approximately 114 to 125 MB per session), a text header (`.vhdr`), and a marker file (`.vmrk`). These are accompanied by BIDS sidecar files detailing channel information, event metadata, electrode positions based on the CapTrak coordinate system, and acquisition parameters. Furthermore, four experimental conditions are recorded during each session: a sustained meditation task (`task-Medita`, `task-slMedita`) alongside resting-state blocks with eyes open (`task-restOE`) and eyes closed (`task-restCE01`, `task-restCE02`).

Preprocessed derivatives. The `derivatives/` directory houses two distinct tiers of processed data. Both processing pipelines operate on all 118 sessions across the five task segments, initially producing 590 base files per tier (Figure 6). The first tier, `eeglab_preproc/` (15 GB), contains data preprocessed via EEGLAB in `.set` and `.fdt` file pairs. The file count in this tier is elevated because it retains an intermediate post-ICA artifact-removal stage (`_preproc_icrm`) for the three tasks that require ICA cleaning (`task-slMedita`, `task-restCE01`, `task-restCE02`). The second tier, `ml_preproc_data/` (11 GB, 590 files), discards these intermediate pairs, storing only the final cleaned epoch tensor as a single NumPy (`.npy`) array per session per task.

B Extended Experimental Paradigms and Questionnaires

B.1 Participant Inclusion and Exclusion Criteria

To ensure sample homogeneity and data integrity, participants were screened based on several physiological, neurological, and logistical criteria. Inclusion required an intermediate level of English proficiency or higher and a current affiliation with Michigan State University (e.g., as a student, faculty, or staff member). Candidates were excluded if they reported significant vision, speech, or hearing impairments, or a history of significant head injuries or neurological disorders, including epilepsy, seizures, or stroke. Furthermore, participants were required to be free of any medications known to alter brain function and were mandated to abstain from alcohol and THC-containing substances for 24 hours preceding the experimental session. Additional screening factors included hand dominance, summer availability, and documented meditation experience. Specifically, screening ensured that participants had little to no prior meditation experience (defined as having practiced only once, or for a few days at least five years prior). Final eligibility was also contingent upon the ability to have reliable transportation to the testing facility.

B.2 Meditation Techniques and Training & Practice Protocol

The present study employed three distinct meditative practices to investigate the neurophysiological correlates across the participants: breath focus meditation, Hare Krishna mantra meditation, and SA-TA-NA-MA mantra meditation. The practice of breath focus meditation requires the sustained attention of the practitioner on the somatic sensations of respiration, functioning primarily as an exercise in cognitive control and the continuous monitoring of sensory input. By repeatedly redirecting the focus of the mind back to the breath upon the detection of mind-wandering episodes, this technique actively recruits the executive control networks of the brain. Both the Hare Krishna and SA-TA-NA-MA mantra meditations involve the continuous mental or vocal repetition of a specific sequence of syllables, which engages the articulatory rehearsal component of working memory and modulates the default mode network through constant auditory and cognitive engagement.

Prior to the formal recording of the EEG data, all participants underwent a standardized training protocol to ensure the accurate execution of each technique. The duration of the training session was approximately twenty-five minutes, utilizing a standardized audio script rather than live expert guidance to maintain absolute consistency across all subjects. During this phase, the participants were seated comfortably in a sound-attenuated room to minimize environmental distractions and physiological artifacts. The instructions detailed the required posture, the precise mechanics of the chants, and the standardized method for redirecting attention when distracted, thereby establishing a uniform experiential baseline for the participants before the commencement of the experimental trials. Following the initial baseline assessments, participants were instructed to practice daily following a gradually increasing schedule designed to enhance achievability: 5 minutes daily for the first week, 10 minutes daily for the second week, and 15 minutes daily from the third week onwards.

Furthermore, to monitor adherence and ensure practice quality throughout the intervention, participants maintained a daily online journal. This log systematically recorded the practice time of day, total duration, post-meditation thoughts or feelings, and a 1 to 5 self-rated assessment of overall focus quality.

B.3 EEG Session and Task Procedures

During both the baseline and post-intervention EEG recording sessions, participants completed a standardized sequence of five tasks while seated comfortably. The total duration of the session was approximately one hour, encompassing the preparation and the recording. Prior to the commencement of each block, the research team provided oral instructions and manually inserted event markers into the continuous EEG recording to precisely denote the onset and offset of the tasks. To minimize oculomotor artifacts, participants were instructed to maintain eye closure throughout the sequence, with the exception of the initial baseline task.

1. **Eyes-Open Resting State (restOE) - 2 mins:** For this initial baseline, participants were instructed to keep their eyes open and maintain a relaxed state.

2. **Eyes-Closed Resting State 1 (restCE01) - 4 mins:** For this pre-task baseline, participants were instructed to *"close your eyes and let your mind wander."*

3. **Active Meditation (Medita) - 8 mins:** The instructions for this block were dependent on the assigned group. Participants in the mantra groups (SA and HK) were instructed to *"close your eyes throughout the task, chant the assigned mantra out loud, and focus on the mantra."* Participants in the BF group were instructed to *"close their eyes and perform alternate nostril breathing and focus on their breath"*.

4. **Eyes-Closed Resting State 2 (restCE02) - 4 mins:** This post-meditation resting interval utilized instructions identical to those of restCE01.

5. **Silent Meditation (slMedita) - 8 mins:** Participants in the SA and HK groups were instructed to *"repeat the mantra in your mind, like inner speech, and focus on that with eyes closed throughout the task time."* Participants in the BF group were instructed to *"close your eyes and focus on your breathing"*.

B.4 Psychometric Questionnaires

To establish a comprehensive psychological profile and provide a standardized context for the interpreted neural activity, a battery of psychometric assessments was administered following the experimental trials. The first instrument administered was the Perceived Stress Scale (PSS) [44]. The primary function of this scale is to evaluate the degree to which situations in the daily life of the individual are appraised as stressful. By quantifying the unpredictable and uncontrollable aspects of the life of the respondent, the PSS provides a critical measure of the current psychological load, which serves as a potential covariate in the analysis of the efficacy of the meditation intervention.

Following the assessment of stress, the participants completed the Short Form of the Five Facet Mindfulness Questionnaire (FFMQ-SF) [45]. The purpose of this instrument is to characterize the inherent capacity of the individual for trait mindfulness. The study evaluates this disposition through five distinct subcategories: observing, describing, acting with awareness, non-judging of inner experience, and non-reactivity to inner experience. We compare the pre- and post-intervention scores to assess changes in trait mindfulness.

Finally, the participants completed the second version of the Multidimensional Assessment of Interoceptive Awareness (MAIA-2) [46]. This assessment is utilized to evaluate the subjective perception of the individual regarding internal bodily sensations. It captures the multidimensional nature of interoception through domains such as the noticing of somatic sensations, the regulation of psychological distress through somatic attention, and the emotional response to bodily states. The administration of MAIA-2 facilitates a deeper understanding of the interaction between the somatic awareness of the individual and the specific cognitive demands of the different meditative paradigms, thereby providing a robust psychological framework for the observed EEG data.

B.4.1 Psychological Assessment Outcomes

This section presents the quantitative outcomes of the psychological assessments administered during the pre-intervention and post-intervention sessions. Specifically, we evaluate the longitudinal shifts in participant responses across three validated instruments: the Perceived Stress Scale (PSS) to measure stress reduction, the Multidimensional Assessment of Interoceptive Awareness version 2 (MAIA-2) to assess interoceptive body awareness, and the Five Facet Mindfulness Questionnaire: Short-Form (FFMQ-SF) to quantify trait mindfulness. The subsequent paragraphs detail the statistical variations and psychometric profile shifts for the SA, HK, and BF groups to establish the psychological efficacy of the respective interventions.

Table 9: Questionnaire descriptive statistics (mean $\pm$ standard deviation) by arm for the strict cross-instrument paired cohort. The scores for the MAIA-2 and FFMQ-sf represent the average of the total summed scores for each group. Note: one subject all questionnaires are missing for both pre- and post-intervention resulted in $n = 73$.

		Unpaired (All)		**Paired Cohort** ($n = 43$)		
Measure	**Group**	n	**Pre**	n	**Pre**	**Post**
PSS $^{\downarrow}$	SA	26	16.2 ± 4.6	16	17.7 ± 5.6	13.8 ± 6.0
	HK	31	18.4 ± 5.8	19	19.0 ± 6.0	15.6 ± 5.0
	BF	16	19.2 ± 6.3	8	20.5 ± 7.4	15.1 ± 7.1
MAIA-2 $^{\uparrow}$	SA	26	2.84 ± 0.66	16	2.88 ± 0.65	3.44 ± 0.25
	HK	31	2.58 ± 0.56	19	2.65 ± 0.44	3.10 ± 0.57
	BF	16	2.75 ± 0.66	8	2.86 ± 0.79	3.65 ± 0.53
FFMQ-SF $^{\uparrow}$	SA	26	3.22 ± 20.46	16	3.24 ± 0.53	3.36 ± 0.47
	HK	31	3.05 ± 0.45	19	3.20 ± 0.39	3.38 ± 0.33
	BF	16	3.27 ± 0.51	8	3.22 ± 0.52	3.55 ± 0.49

Quantitative Changes in the Perceived Stress Scale Figure 7 illustrates the total scores of the Perceived Stress Scale (PSS) for the SA, HK, and BF groups during both the pre-intervention and post-intervention sessions. Within the plots, each data point represents the raw PSS total score of an

individual participant, accompanied by a probability density curve and a box plot to indicate the data distribution.

Table 9 details the differences in scores between the two sessions. The comparative analysis between the pre-intervention and post-intervention stages relies on a paired cohort of 43 participants who completed the questionnaire assessments at both time points. Although the table includes an unpaired column detailing the overall pre-intervention baseline results for the entire initial participant pool, all longitudinal comparisons and subsequent quantitative findings discussed throughout this section, including the analyses of the MAIA-2 and FFMQ-SF questionnaires, derive exclusively from this paired cohort. Within this cohort, the mean PSS total scores demonstrate a consistent reduction from the pre-intervention session to the post-intervention session across all groups. Specifically, the mean total score for the SA group significantly decreased ($p < 0.05$) from 17.7 at pre-intervention to 13.8 at post-intervention. Furthermore, the mean total score for the HK group significantly decreased from 19.0 to 15.6, and the corresponding score for the BF group non-significant decreased from 20.5 to 15.1.

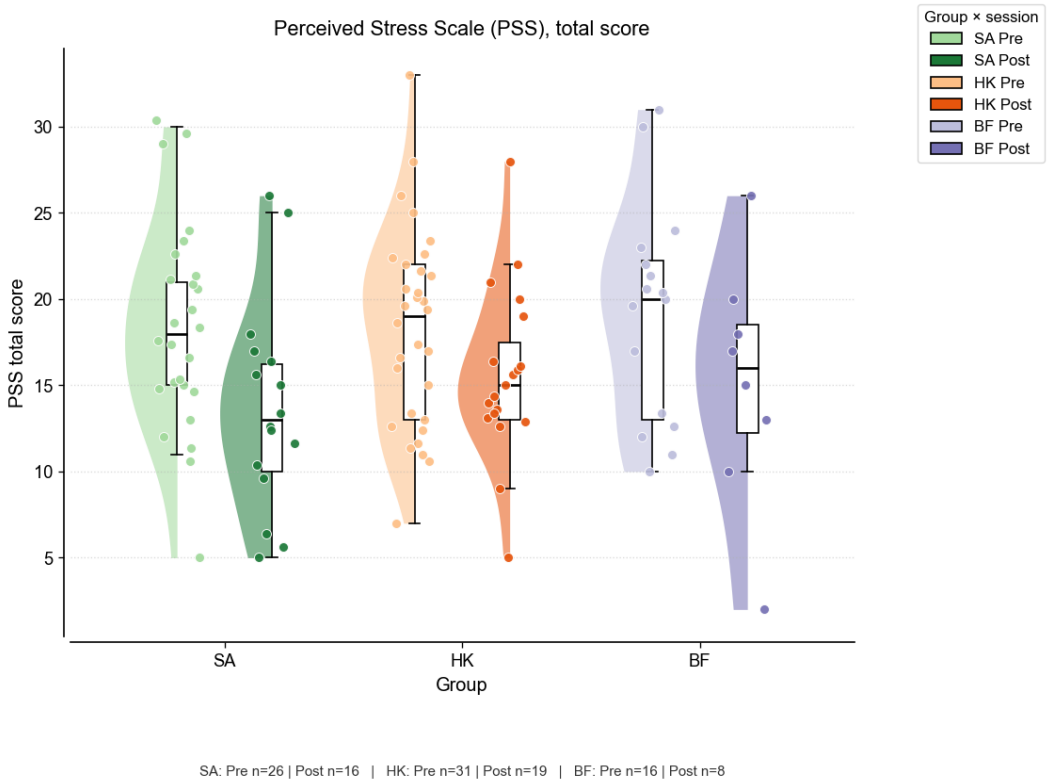

Figure 7: Total PSS scores at baseline and follow-up for SA, HK, and BF.

Quantitative Changes in the Multidimensional Assessment of Interoceptive Awareness The Multidimensional Assessment of Interoceptive Awareness version 2 (MAIA-2) evaluates bodily awareness across eight subscales. These include Noticing (the awareness of comfortable, uncomfortable, and neutral sensations), Not-Distracting (the tendency not to ignore pain or discomfort), Not-Worrying (the absence of emotional distress when experiencing discomfort), Attention Regulation (the ability to sustain and control attention on bodily sensations), Emotional Awareness (the recognition of connections between bodily sensations and emotional states), Self-Regulation (the ability to manage psychological distress by attending to bodily sensations), Body Listening (the active receptivity to physiological signals for insight), and Trusting (the perception of the body as safe and trustworthy). Figure 8 presents normalized radar plots for the SA, HK, and BF groups, illustrating shifts in these subscale profiles from pre- to post-intervention.

Subscale profiles varied across the study groups after the intervention. The SA group improved across most domains, particularly in Not-Distracting, Trusting, and Attention Regulation. The HK group

showed similar overall gains, primarily in Attention Regulation, Self-Regulation, and Emotional Awareness. The BF group improved in Self-Regulation, Emotional Awareness, and Trusting, but showed negligible change in Noticing. Table 9 details these quantitative changes. Overall MAIA-2 item means increased for all groups: from 2.88 to 3.44 for the SA group, from 2.65 to 3.10 for the HK group, and from 2.86 to 3.65 for the BF group.

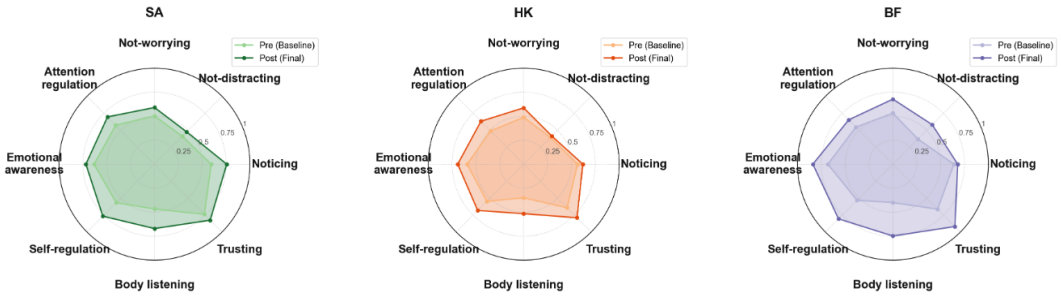

Figure 8: Mean MAIA-2 scores at baseline and follow-up for SA, HK, and BF on eight domains.

Quantitative Changes of the Five Facet Mindfulness Questionnaire: Short-Form The Five Facet Mindfulness Questionnaire: Short-Form (FFMQ-SF) assesses trait mindfulness across five distinct subscales using a five-point Likert scale, ranging from 1 (never or very rarely true) to 5 (very often or always true). These subscales encompass Observing (the conscious perception of internal and external sensory experiences), Describing (the ability to clearly label internal experiences with words), Acting with Awareness (the maintenance of focused attention on current activities without operating on autopilot), Non-judging of Inner Experience (the adoption of a non-evaluative attitude toward personal thoughts and emotions), and Non-reactivity to Inner Experience (the capacity to allow thoughts and emotions to pass without getting entangled or reacting impulsively). Similar to the visualization approach employed for the MAIA-2, Figure 9 presents normalized radar plots for the SA, HK, and BF groups, depicting relative shifts in these facet profiles from pre-intervention to post-intervention.

Distinct variations in facet profiles are observable across the study groups following the intervention. In the SA group, post-intervention scores demonstrate minor improvements across all facets expect slightly decrease in Describe facet. The HK group exhibits a general expansion of the mindfulness profile, with prominent enhancements in all five facets. Conversely, the BF group displays a distinct pattern characterized by notable increases in Describing, Observing, Acting with Awareness, and Non-reactivity, coupled with a discernible decrease in the Non-judging facet. These visual profile shifts are quantitatively supported by the data in Table 9, which presents the average of the total summed scores for each group. Specifically, the mean score for the SA group increases from 3.24 at pre-intervention to 3.36 at post-intervention. Furthermore, the mean score for the HK group increases from 3.20 to 3.38, and the mean score for the BF group rises from 3.22 to 3.55.

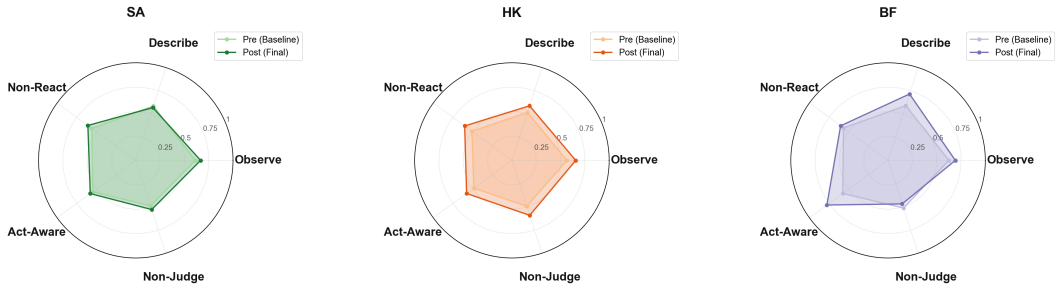

Figure 9: Mean FFMQ-SF scores at baseline and follow-up for SA, HK, and BF on five facets.

C Detailed Methodology and Data Processing

C.1 Preprocessing Details

For data in cleaned EEG in Section 3.3, the High-pass conditioning utilized a 1 Hz zero-phase Butterworth filter. Zapline-plus removed 60 Hz interference via spectral integration without distorting neural oscillations. ASR was configured with a 25-standard-deviation burst detection threshold and 0.2 maximum bad-channel tolerance. Channels flagged by ASR were interpolated using spherical splines to maintain the 64-channel manifold. A zero-valued reference was appended before common average re-referencing across all electrodes plus the FCz channel, totaling 65 channels. ICLabel components were rejected if the artifact probability reached ≥ 0.9. For the training data used for Benchmark tasks, the algorithm simply preprocessed the data with a spatial correlation threshold of 0.9 and a line-noise criterion of 4 standard deviations.

C.2 Model Training Details

Temporal Partitioning And Class Balancing At 250 Hz, the recordings yield 64×1000 feature matrices across 64 channels. For intra-subject evaluations, the data are partitioned into 20-second blocks, alternating between training and testing sets in an $80\%/20\%$ ratio. Unlike standard chronological splitting, this block-wise strategy is utilized exclusively in the intra-subject setting to account for non-stationary temporal dynamics throughout the session. To prevent data leakage between adjacent blocks, 4-second safety gaps are inserted (Figure 10). Additionally, to address duration discrepancies between resting and meditation states, a dynamic weighted random sampler is implemented to ensure a 1:1 class exposure per mini-batch during training.

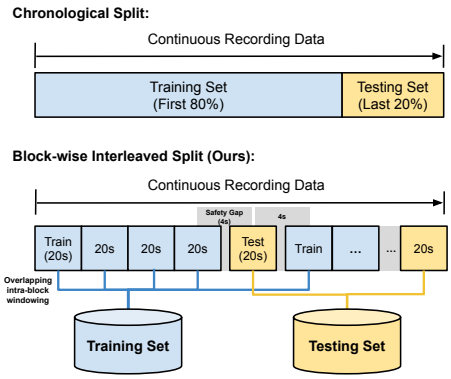

Figure 10: Chronological split and block-wise interleaved split(Ours) for intra-subject data splitting.

Training traditional SVMs on FBCSP and PSD extracted features. For FBCSP, we use a sampling frequency of 250 Hz, number of components $k = 4$ which is the number of features extracted from every frequency band, we filter the EEG reading for 9 bands, namely, (4,8), (8,12), (12,16), (16,20) (20,24), (24,28), (28,32), (32,36), and (36,40) Hz each. The input dimension of the EEG reading to algorithm is $N * C * T$, and the output dimension is $N * Bk$, where N is the number of trials which varies based on the evaluation, $C = 64$ is the number of channels, and $T = 1000$ is the time steps (number of snapshots taken during EEG reading), $B = 9$ represents the 9 bands extracted from each EEG readin g, and $k = 4$ is the number of features extracted from each band using the band's specific filter w_b. For the PSD feature extraction, we use the same sampling frequency, but different frequency bands of delta (1, 4), theta (4, 8), alpha (8, 13), beta (13, 30), and gamma (30, 45) and the output dimension of the features is $N * BC$ because PSD measures frequency power for each channel and uses those as features. We train an SVM on top of features extracted from both of these methods, the SVM has a linear kernel and parameter $C = 1.0$.

C.3 Benchmark Task Detailed Methodology

Data and Cohort Specifications for Task 1. This paragraph supplements the methodological discussion in Subsection Task 1: Cognitive State Decoding. The initial session includes the complete cohort ($N = 74$), avoiding the attrition observed in the post-intervention phase ($N = 44$). Continuous recordings comprise a four-minute resting baseline and an eight-minute meditation task. The global model is trained on all 74 subjects to establish a comparison baseline for the subgroup-specific models.

Evaluation Logic And Attrition For Task 2 This paragraph supplements the methodology for fine-grained technique classification and longitudinal tracking. For this specific task, a stratified inter-subject cross-validation approach is the only viable evaluation strategy, whereas intra-subject and

Leave-One-Subject-Out (LOSO) paradigms are fundamentally inapplicable. Because each participant was assigned to practice exclusively one specific meditation technique, an individual subject's dataset inherently lacks multi-class variance. Consequently, both an intra-subject split and a LOSO approach would inevitably yield a test set containing only a single class label.

Calibration Protocols for Task 3. This paragraph supplements the methodology described in Subsection 4.3. We evaluate the model performance under zero-shot, 10-shot (consisting of 8.5 seconds of data for both resting and meditation states), and 30-shot (consisting of 18.5 seconds for each state) configurations. The full post-session data from the intra-blockwise training of Task 1 is utilized as the performance ceiling. To maintain sufficient signal density, the data is processed using 4-second windows with an 87.5% overlap. We implement two distinct strategies for the post-training adaptation process. The first strategy involves freezing the convolutional layers and updating only the normalization layers. This approach is adopted because normalization layers perform linear transformations, which facilitate efficient domain alignment with minimal risk of overfitting on small datasets. The second strategy permits full fine-tuning, allowing the update of all model parameters. A comprehensive comparison of these two strategies is presented in the subsequent results section.

C.4 Implementation Details

Hardware, Software, and Parallelization All experiments were conducted on a server equipped with eight NVIDIA RTX A5000 GPUs (24 GB VRAM each). Models were implemented in PyTorch 2.6.0 with CUDA 11.8. Each training run was launched as an independent single-GPU process and assigned to a dedicated device via a custom scheduling script, enabling up to eight concurrent runs without inter-GPU communication overhead. All computations were performed in standard float32 precision.

Optimization and Hyperparameter Search All models were optimized using Adam ($\beta_1 = 0.9, \beta_2 = 0.999$) with cross-entropy loss and a constant learning rate throughout training. Hyperparameters were selected via Optuna with 50 trials per model–task combination, maximizing validation balanced accuracy; unpromising trials were pruned using the Median Pruner (warm-up steps = 10). The search covered learning rate $\in [10^{-5}, 10^{-3}]$ (log-uniform; restricted to $[10^{-5}, 5 \times 10^{-4}]$ for Conformer), weight decay $\in [10^{-5}, 10^{-2}]$ (log-uniform), dropout rate $\in [0.05, 0.7]$, temporal kernel size, and model-specific architectural parameters (e.g., embedding size and attention depth for Conformer). For Tasks 1 and 2, final models were trained for up to 250 epochs with a batch size of 32 and early stopping (patience = 50 epochs) based on validation balanced accuracy. For Task 3, only the classification head was fine-tuned for 30 epochs with a batch size of 4 to mitigate overfitting in the few-shot regime.

To ensure reproducibility, all random seeds across PyTorch, NumPy, and Optuna samplers were fixed to a designated value, and deterministic algorithms were enforced where supported.

D Additional Experimental Results

D.1 Benchmark Task1 Supplementary Material

This supplementary section provides a multidimensional analysis of the decoding performance for Benchmark Task 1, extending beyond standard aggregate metrics to dissect the critical bottlenecks in EEG-based classification. Specifically, we present detailed investigations into four key aspects: (1) the profound performance discrepancy between subject-specific and cross-subject evaluation protocols, (2) cohort-specific biases and cross-group generalization dynamics across three distinct sub-populations (SA, HK, and BF), (3) the impact of temporal non-stationarity evaluated through a rigorous chronological-split protocol versus standard block-randomized cross-validation, and (4) subject-level heterogeneity, which isolates specific algorithmic vulnerabilities to traditionally challenging participants.

The primary objective of including these granular analyses is to expose the hidden variances, temporal drifts, and dataset biases that are frequently obscured by conventional mean-accuracy reporting. For researchers utilizing this dataset, these results serve as a diagnostic map of the data's inherent complexities. They demonstrate that idealized intra-subject evaluations represent a theoretical upper bound, whereas inter-subject and temporal-causal protocols reflect true real-world robustness.

Performance Degradation for Cross-Subject and Superior Generalization of Convolutional Architectures The experimental results from Table 10 reveal a substantial performance discrepancy between subject-specific and cross-subject evaluation strategies. In the Block-wise and Chrono-wise intra-subjects splitting settings, most models achieve high decoding accuracy, with EEGNet reaching 99.2% and 96.8% AUC, respectively. However, when transitioning to Inter-Subject and LOSO protocols, all models experience a drastic decline in performance. This significant drop underscores the severe challenge posed by inter-subject variability and distribution shifts in EEG signals, indicating that models heavily rely on subject-specific temporal features when trained and evaluated on individualized data splits.

Comparing traditional machine learning approaches with deep neural networks reveals a stark contrast in cross-domain generalization capabilities. The FBCSP + SVM pipeline achieves near-perfect metrics in the Intra-Block setting, recording an AUC of 99.9% and an accuracy of 99.4%. Nevertheless, its performance plummets to near-chance levels, specifically 55.8% AUC and 53.8% accuracy, under the Inter-Subject protocol. Same thing happens when the FBCSP and PSD feature extraction methods are combined with an SVM in the LOSO setting. The AUC for FBCSP + SVM drops to 51.8% while the PSD + SVM drop to 61.3%. This shows that these methods struggle when presented with data that wasn't included when fitting the SVM, and hence they have low ability to generalize, making them best suited to use for Intra-Subject evaluations.

In contrast to traditional methods, deep learning models such as ShallowConvNet and EEG-Conformer demonstrate superior generalizability and maintain a much stronger baseline in cross-subject scenarios (inter-subjects and LOSO). It is worth noting that EEG-Conformer explicitly incorporates a shallow convolutional module as its initial feature extraction head prior to applying its global self-attention mechanism. The strong performance of both architectures indicates that CNN learned spatial-temporal representations offer greater robustness to subject-specific noise than traditional hand-crafted spatial filters. Furthermore, the observation that a relatively simple architecture like ShallowConvNet achieves highly competitive results suggests that deeper convolutional models may be highly susceptible to overfitting when deployed across distinct individuals. Consequently, utilizing fewer convolutional layers appears to act as an implicit regularization mechanism, successfully preventing the network from memorizing idiosyncratic, subject-specific artifacts and thereby preserving cross-population generalizability.

Evaluating Cohort Biases And Cross-Group Generalization Dynamics To investigate how different population distributions affect model robustness, the dataset is partitioned into three distinct sub-datasets: SA with $n = 27$, HK with $n = 31$, and BF with $n = 16$. The fundamental purpose of this group-specific experiment is to isolate cohort characteristics and reveal how internal data variance and sample size fundamentally influence model generalization. For the intra-subject protocol, the displayed metrics represent the average decoding accuracy computed across all individuals strictly within their respective group. Conversely, for the inter-subject and leave-one-subject-out protocols, the dataset is completely re-partitioned, meaning models are independently trained and tested exclusively within each of the three isolated subsets. While the intra-subject results display uniformly excellent performance across all groups, with architectures like EEGNet universally exceeding 98% area under the curve, the cross-subject evaluations expose profound inter-group disparities. The SA cohort consistently demonstrates the most superior generalization capability, achieving the highest metrics across most deep learning architectures, peaking at a 69.1% area under the curve for EEGNet under the leave-one-subject-out protocol. Surprisingly, despite possessing the largest participant pool of thirty-one subjects, the HK group yields noticeably inferior cross-subject performance compared to the SA group. Meanwhile, the BF cohort, severely constrained by the smallest sample size of sixteen subjects, exhibits the most severe performance degradation, with inter-subject area under the curve dropping to as low as 48.2%. This nonlinear relationship between subject count and testing accuracy highlights a critical insight: raw sample size does not strictly dictate cross-subject robustness. Instead, the superior generalization of the SA group strongly suggests that the most likely primary factor is the more pronounced inherent distinction between the SA task itself and MW, which facilitates easier feature extraction. Secondarily, cohort-specific factors such as intrinsic data quality, demographic homogeneity, or distinct experimental conditions also play a vital role in forming transferable feature representations. These findings underscore the absolute necessity of identifying and addressing both task-inherent variances and dataset biases before deploying models across diverse populations.

Table 10: **Task 1: Decoding Performance Across Protocols.** Detailed metrics for each evaluation strategy. Results are reported as Mean ± SD across subjects ($N = 74$). **Bold** indicates the best performance within each strategy group.

Strategy	Model	AUC (%)	Acc (%)	BAcc (%)	F1 (%)
Intra-Subject (block-wise)	PSD + SVM	97.1±4.0	92.9±7.2	92.1±8.0	92.0±8.1
	FBCSP + SVM	**99.9±0.2**	**99.4±1.5**	**99.3±1.9**	**99.4±1.7**
	ShallowConvNet	97.2±4.4	92.6±6.1	90.6±7.7	91.3±7.2
	DeepConvNet	97.0±8.5	88.0±15.2	86.7±14.2	85.5±18.1
	EEGNet	99.2±2.1	96.2±4.6	95.4±6.0	95.6±5.5
	EEG-Conformer	97.0±4.9	92.3±6.9	90.3±9.1	90.8±8.6
Intra-Subject (chrono-wise)	PSD + SVM	94.3±7.0	89.0±8.6	87.0±10.6	87.1±10.4
	FBCSP + SVM	**99.5±2.1**	**98.5±5.0**	**97.9±7.6**	**97.8±8.0**
	ShallowConvNet	97.2±4.4	92.6±6.1	90.6±7.7	91.3±8.0
	DeepConvNet	97.0±8.5	88.0±15.2	86.7±14.2	85.5±18.1
	EEGNet	99.2±2.1	96.2±4.6	95.4±6.0	95.6±5.5
	EEG-Conformer	97.0±4.9	92.3±6.9	90.3±9.1	90.8±8.6
Inter-Subject	PSD + SVM	59.4±3.8	57.7±4.2	56.9±2.9	55.6±3.4
	FBCSP + SVM	55.8±6.2	53.8±4.7	53.7±3.6	52.1±3.9
	ShallowConvNet	66.5±3.9	**64.9±4.9**	**61.3±3.1**	**60.7±3.7**
	DeepConvNet	66.2±5.3	62.8±5.3	59.5±2.8	59.0±3.6
	EEGNet	66.5±3.5	60.6±5.3	60.9±3.7	59.0±4.5
	EEG-Conformer	**66.9±4.7**	64.3±5.0	61.1±3.1	60.6±3.7
LOSO	PSD + SVM	61.3±4.4	56.7±2.3	57.9±2.8	55.6±12.5
	FBCSP + SVM	58.2±21.4	61.7±11.8	52.7±10.8	47.6±12.0
	ShallowConvNet	**70.4±20.8**	**63.9±16.7**	**61.5±15.0**	**56.7±17.9**
	DeepConvNet	68.4±26.8	61.3±20.4	59.8±18.2	54.6±20.7
	EEGNet	66.6±27.2	58.4±21.2	59.8±17.9	52.5±21.9
	EEG-Conformer	67.5±24.3	62.7±17.8	60.0±15.8	55.3±18.3

Impact Of Temporal Shifts On Evaluation Realism The transition from intra-block to intra-chrono evaluation protocols exposes a systematic performance degradation across all evaluated architectures, with area under the curve metrics declining by 2.1% to 4.4% as explicitly detailed in the model-level comparison (Figure 11A). This performance gap highlights the methodological trade-offs inherent to the block-wise partitioning strategy. On the positive side, the higher accuracy achieved by intra-block splitting demonstrates that this method successfully captures the non-stationary temporal dynamics throughout the session, thereby establishing a valuable performance upper bound for within-session real-time detection applications. However, this localized randomization simultaneously introduces the risk of temporal leakage, where future signal dynamics inadvertently inform the training phase. Whether this specific leakage ultimately compromises cross-session generalization accuracy remains an open question that warrants further observation. In contrast, the intra-chrono protocol enforces strict past-to-future causality, explicitly revealing the models' vulnerability to natural signal drift over time. Furthermore, individual scatter distributions and per-subject temporal gap analyses (Figure 11B and Figure 11C) highlight that this temporal vulnerability is highly heterogeneous across subjects. While the majority of data points fall below the parity line, indicating generalized performance loss, specific individuals experience severe performance collapses whereas others maintain stable accuracy regardless of the splitting strategy. Ultimately, while the intra-block evaluation establishes a theoretical upper bound for model capacity under stationary conditions, the intra-chrono approach provides a fundamentally more rigorous and realistic assessment of cross-temporal generalization for continuous, online deployment scenarios.

Analyzing Individual Variability and Cohort Biases The primary objective of this subject-wise analysis is to evaluate intra-subject decoding performance across the dataset. While models demonstrate high state-separation accuracy for most participants, a subset of subjects remains

Table 11: **Task 1: Group-specific Performance.** Mean AUC (%) and BAcc (%) ± standard deviation per site. Intra-subject results are computed from per-subject held-out evaluations within each site.

		Intra-Subject		Inter-Subject		LOSO	
Group	**Model**	**AUC**	**BAcc**	**AUC**	**BAcc**	**AUC**	**BAcc**
SA (n=27)	ShallowConvNet	98.4±2.5	93.4±5.5	61.1±4.4	57.5±4.1	66.9±22.3	59.8±13.8
	DeepConvNet	96.2±11.2	88.1±14.2	60.6±3.9	55.9±4.1	68.9±24.8	61.1±18.8
	EEGNet	**99.7±0.5**	**96.6±4.4**	60.5±9.5	56.5±6.7	**69.1±23.1**	**61.6±13.8**
	EEG-Conformer	98.6±2.6	93.2±6.7	**61.1±4.7**	**57.8±4.1**	68.8±26.3	61.2±18.2
HK (n=31)	ShallowConvNet	96.2±5.7	89.4±8.4	54.5±7.2	**54.4±5.5**	**61.2±29.3**	**57.0±14.9**
	DeepConvNet	97.4±7.3	88.2±13.0	54.1±9.5	51.8±8.3	58.3±30.7	53.8±18.6
	EEGNet	**98.8±3.1**	**94.7±6.9**	**55.5±11.6**	52.8±9.6	57.5±31.1	54.2±23.0
	EEG-Conformer	96.2±5.8	89.3±8.7	54.0±9.8	52.2±8.4	60.1±31.7	53.6±19.6
BF (n=16)	ShallowConvNet	97.1±3.2	88.1±7.9	55.2±3.4	53.9±2.5	51.5±17.1	52.0±7.9
	DeepConvNet	97.7±3.9	81.3±14.9	50.4±8.9	52.1±4.7	52.4±24.2	53.3±11.5
	EEGNet	**99.1±1.2**	**94.8±6.0**	48.2±9.0	49.1±6.3	49.2±21.6	51.2±8.0
	EEG-Conformer	95.9±5.4	87.1±11.4	**55.6±5.9**	**54.9±3.1**	**55.9±20.7**	**53.9±8.4**

challenging (Figure 12). This difficulty can be attributed to several factors. First, for novice practitioners engaging in silent meditation for the first time, frequent mind-wandering episodes can blur the neural distinction between meditation and resting states. Second, inherently poor signal quality or atypical EEG patterns in certain subjects complicate feature extraction. As a result, deep learning classifiers such as the Deep Convolutional Network (DCN) exhibit noticeably reduced classification accuracy on this subset.

To systematically quantify this variability, subjects are categorized into quartiles based on their average decoding accuracy across all groups: the top 25% are defined as resilient subjects (easiest to decode), and the bottom 25% are defined as vulnerable subjects (hardest to decode). Stratifying these distributions reveals distinct cohort differences. The SA group ($n = 27$) exhibits the most robust intra-subject state separation, featuring the highest proportion of resilient subjects at 37.0% and the lowest proportion of vulnerable subjects at 18.5%. In contrast, the BF group ($n = 16$) is heavily skewed toward the difficult end, with 68.7% of its participants falling into the bottom half of performance (Q1 and Q2) and only 6.2% classified as resilient. Similarly, the HK group ($n = 31$) presents a higher vulnerable proportion (29.0%) than the SA cohort. These disparities align with the initial difficulty of the respective practices for novices. The SA task uses a simple, four-syllable mantra that facilitates sustained attention during early training stages, whereas the BF task demands continuous breath monitoring and the HK task involves a longer, more complex mantra. Both BF and HK tasks are therefore more susceptible to mind-wandering during early training stages.

This subject-wise analysis provides a practical guide for future researchers using the L-FAME dataset. Rather than treating decoding failures as statistical noise obscured by aggregate metrics, this stratification highlights challenging edge cases. This benchmark offers two methodological paths for future studies. Researchers aiming to establish the upper bound of state separability can use these vulnerability metrics as an empirical exclusion criterion, selectively removing vulnerable subjects to isolate pure neural signatures of the target states. Alternatively, researchers focusing on algorithmic robustness can specifically target these vulnerable subjects. Future work should investigate the root causes of these intra-subject failures to determine whether performance degradation is driven by behavioral non-compliance (e.g., mind-wandering during silent meditation) or artifactual signal degradation. By explicitly reporting these individual differences, this analysis contextualizes the overall task difficulty and directs the community toward targeted algorithmic improvements.

D.2 Benchmark Task 2 Supplementary Material

This section investigates the neurophysiological differences among the three meditation techniques and evaluates how the six-week intervention affects their neural representations. To separate practice-induced neural adaptations from potential attrition bias, we analyze a matched paired-subject cohort.

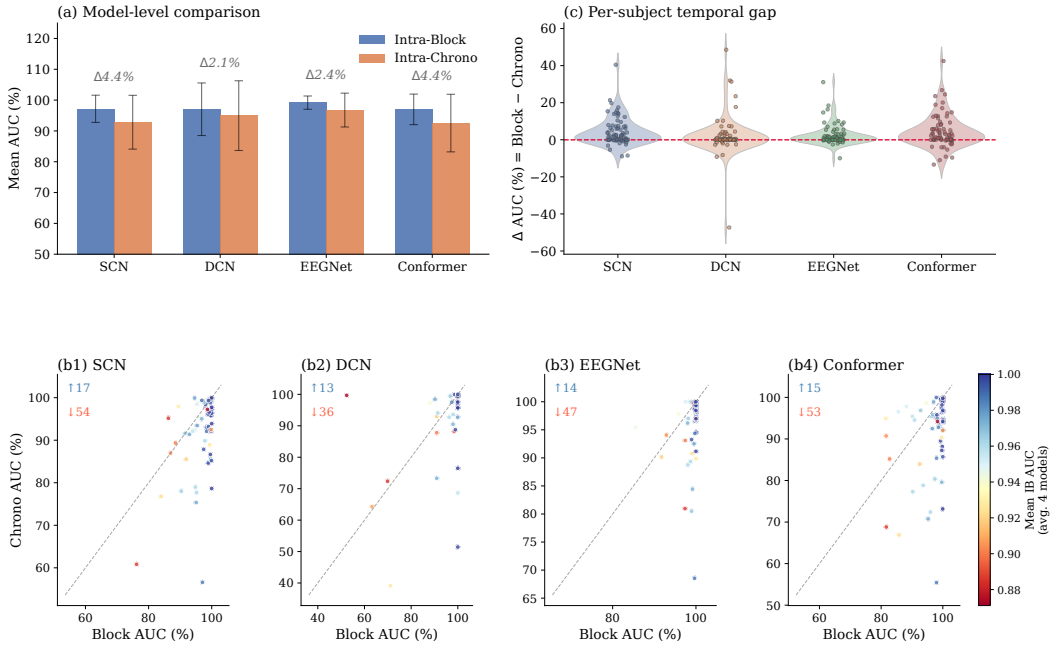

Figure 11: **Intra-Subject Evaluation Protocol Comparison: Block-Split vs. Chronological-Split** ($N = 74$) **(a)** Mean AUC (%) and standard deviation across all subjects for four models under two intra-subject protocols: Intra-Block (random fold assignment) and Intra-Chrono (temporally ordered folds). Δ denotes the performance gap between protocols. **(b1–b4)** Per-subject scatter plots for each model; each point represents one subject, with the x-axis showing Intra-Block AUC and the y-axis showing Intra-Chrono AUC. Points below the diagonal (dashed line) indicate subjects whose performance degrades under temporal splitting. Point colour encodes each subject's mean Intra-Block AUC averaged across all four models (warm: lower performers; cool: higher performers). **(c)** Distribution of ΔAUC = Block − Chrono per subject for each model (violin: kernel density; dots: individual subjects); the red dashed line marks $\Delta = 0$. SCN: ShallowConvNet; DCN: DeepConvNet.

We trace the temporal evolution of representational separability using UMAP latent space visualizations and Representational Similarity Analysis (RSA), we observe that three-class classification separability improves post-training compared to pre-training from the extracted feature space, and that the distinct meditation modalities make the BF group easier to distinguish. These observations are validated with an inter-subject One-Versus-All (OvA) classification framework. The results provide statistical evidence for a clear boundary between the somatic-directed breath-focused (BF) modality and the inner-speech practices (HK and SA).

Controlled Longitudinal Evaluation Isolates Genuine Practice Effects From Attrition Bias he primary evaluation of the second task initially compares the pre-intervention performance, trained on the full cohort of 74 subjects, against the post-intervention performance, trained on the 44 subjects who completed the six-week programme. Because these two conditions differ simultaneously in temporal session and subject composition, this unrestricted comparison is inherently confounded by potential attrition bias; highly engaged participants might artificially inflate post-intervention metrics independently of genuine neurophysiological changes. To eliminate this confound and isolate the true longitudinal impact, a strictly paired evaluation protocol was established. Specifically, the analysis was restricted exclusively to the subset of 44 subjects who possessed both pre- and post-intervention recordings. By evaluating their pre-intervention baseline directly against their post-intervention results using an identical inter-subject cross-validation protocol, this paired design ensures a rigorous comparison. Analyzing this strictly matched baseline reveals distinctly divergent behavioral patterns between the deep learning architectures. From Table 12, for the EEGNet model, all four evaluated metrics exhibit a monotonic increase, and crucially, the improvement persists even after the subject population is strictly held constant, with the area under the precision-recall curve rising from 42.9% to 48.8%. This sustained enhancement strongly indicates that the performance gain reflects the genuine

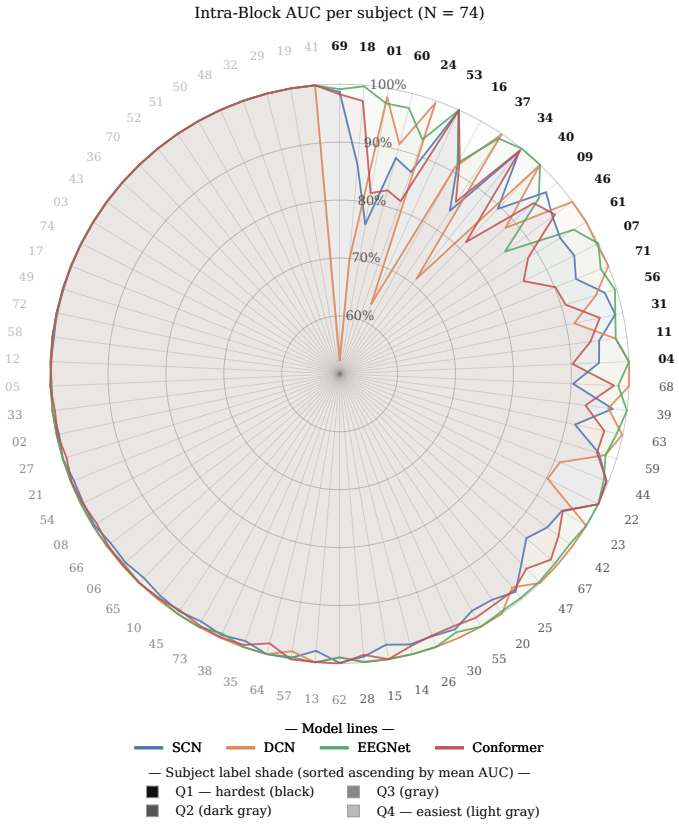

Figure 12: **Intra-Block AUC Across All Subjects and Models** ($N = 74$) Each spoke represents one subject, arranged clockwise in ascending order of mean Intra-Block AUC averaged across all four models; subjects in the lower-performing quartile (Q1) appear first in the sweep. The four coloured lines correspond to the four model architectures: SCN (ShallowConvNet), DCN (DeepConvNet), EEGNet, and Conformer (EEG-Conformer). The radial axis is linearly rescaled such that 50% AUC (chance level) maps to the origin and 100% AUC maps to the outermost ring; concentric grid circles mark 60%, 70%, 80%, 90%, and 100%. Subject labels are shaded in greyscale according to their mean-AUC performance quartile across all models: **Q1** (black, bold) denotes subjects in the lowest quartile ($\leq$25th percentile, most difficult to decode), **Q2** (dark grey), **Q3** (grey), and **Q4** (light grey, highest quartile, $\geq$75th percentile). Inward dips along a spoke indicate that at least one model achieves lower-than-average performance for that subject, highlighting subject-level heterogeneity in EEG decodability.

consolidation of technique-specific neural representations following longitudinal practice rather than mere selection bias. Conversely, the ShallowConvNet model demonstrates statistically indistinguishable performance between the controlled pre-intervention and post-intervention conditions, recording precision-recall areas of 46.9% and 45.7%, respectively. This performance plateau suggests that the apparent improvement observed in the unrestricted comparison for ShallowConvNet is predominantly driven by demographic attrition rather than algorithmic adaptation to practice-induced neural changes. Ultimately, these complementary findings validate the critical necessity of this paired-subject dataset design, as it successfully captures authentic cognitive advancements while simultaneously preventing the misinterpretation of subject selection artifacts as algorithmic improvements.

Longitudinal Intervention Enhances Representational Separability To rigorously determine whether the observed improvements in technique classification accuracy stem from sample size variations or a genuine longitudinal strengthening of technique-specific neural patterns, a paired analysis was conducted on forty-four subjects possessing both pre-intervention and post-intervention recordings. Specifically, features are extracted from the final hidden layer of an EEGNet model trained on post-intervention data and applied identically to both sessions. This strict control ensures that the resulting embedding space reflects actual changes in the neural data rather than shifting model weights. By fitting UMAP jointly on the combined pre- and post-session features ($n_{\text{neighbors}} = 30$,

Table 12: **Task 2 Performance Metrics Under Controlled Longitudinal Conditions.** The Session column denotes the temporal phase of data collection, specifically the pre-intervention (Pre) and post-intervention (Post) stages. The N value indicates the total number of subjects included in the training and evaluation cohort, where N=74 represents the full initial sample and N=44 represents the paired subset of participants who successfully completed the entire six-week meditation program.

Model	Session(N)	PR-AUC	BAcc	F1	AUC
EEGNet	Pre (74)	35.0 ± 2.9	35.0 ± 5.6	32.8 ± 5.9	52.4 ± 4.1
	Pre (44)	42.9 ± 8.7	41.4 ± 6.6	40.1 ± 6.5	56.8 ± 5.5
	Post (44)	$\mathbf{48.8 \pm 8.6}$	$\mathbf{48.0 \pm 7.4}$	$\mathbf{45.0 \pm 8.3}$	$\mathbf{64.5 \pm 7.3}$
ShallowConvNet	Pre (74)	38.0 ± 6.2	38.5 ± 4.4	33.4 ± 4.7	55.4 ± 6.4
	Pre (44)	$\mathbf{46.9 \pm 4.5}$	43.9 ± 5.1	40.5 ± 4.5	59.2 ± 6.3
	Post (44)	45.7 ± 9.7	$\mathbf{44.7 \pm 7.1}$	$\mathbf{42.9 \pm 8.0}$	$\mathbf{60.0 \pm 6.7}$

$\min\_dist = 0.15$) to establish a common coordinate system, the visualization reveals a striking representational shift. In the pre-intervention phase, the feature embeddings for the three meditation techniques remain largely entangled, indicating relatively weak class-specific neural signatures prior to structured practice. Specifically, while the HK (red cluster) and SA (blue cluster) cohorts exhibit substantial overlap, the BF (green cluster) group already begins to demonstrate an incipient separation, although with some residual intersection. In contrast, the post-intervention latent space displays a markedly more dispersed distribution with clearer categorical boundaries (Figure 13). Although a minor overlap remains visible between the HK and SA groups, the BF group becomes clearly isolated into a highly distinct cluster. This progressive separation trajectory highlights the distinct underlying cognitive mechanisms of these practices: the BF technique, predominantly involving somatic attention directed toward respiration, elicits unique neurophysiological features that become readily separable from the inner-speech-focused HK and SA techniques even after short-term training.

To formally validate the visual observations, representational similarity analysis is performed by averaging the high-dimensional feature embeddings within each category to establish class centroids, followed by computing the pairwise cosine similarities among these three centers (Figure 14). The resulting similarity matrices strictly corroborate the progressive separation trajectory. In the pre-intervention phase, the neurophysiological representations exhibit severe entanglement. The inner-speech-focused HK and SA cohorts share a remarkably high cosine similarity of 0.953 , while the BF group also maintains substantial positive correlations with HK at 0.573 and SA at 0.570. After the intervention, the representational landscape undergoes a profound transformation. The similarity between the HK and SA centroids decreases to 0.250, reflecting a measurable divergence despite their shared cognitive foundation. More importantly, the BF centroid becomes structurally independent, exhibiting near-orthogonal or inverse relationships with HK at -0.055 and SA at -0.482. This stark quantitative shift from high positive correlation to orthogonal or negative similarity rigorously confirms that the somatic-directed BF technique cultivates a profoundly distinct neural signature compared to the inner-speech modalities following structured practice.

Quantitative Validation Of Technique Separability Via One-Versus-All Classification To quantitatively validate the geometric separability observed in the prior representations, the three-way technique classification was decomposed into three independent binary classifiers via an One-Versus-All (OvA) approach. For each OVA model, the labeling scheme follows a strict binary dichotomy: data from the designated target group (e.g., HK, SA, or BF) are assigned to the positive class, while data from the remaining two groups are aggregated into the negative class. A rigorous 5-fold inter-subject cross-validation strategy is strictly enforced to evaluate genuine generalization capabilities and prevent data leakage. Specifically, the entire subject pool is partitioned into five mutually exclusive subsets. In each iteration of the cross-validation process, four subsets (comprising 80% of the individuals) are allocated to the training set to optimize the classifiers, while the single remaining subset (comprising 20% of the individuals) is exclusively held out for the testing phase. By iteratively ensuring that the training and testing sets contain completely non-overlapping subject pools across all five folds, this inter-subject training paradigm guarantees that the models learn universal, technique-specific cognitive traits rather than overfitting to individual-level physiological artifacts. Because of this structural skew, the binary AUC and BAcc serve as the most rigorous and sensitive metrics for evaluation. The

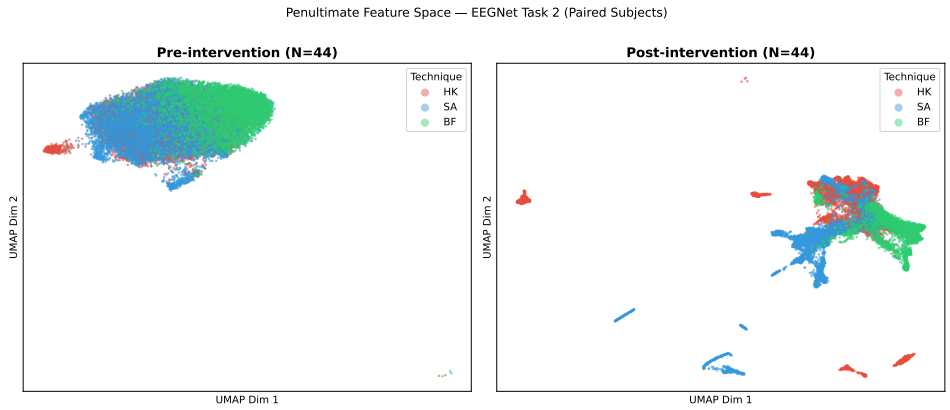

Figure 13: **UMAP visualization of EEGNet penultimate-layer representations for the 44 paired subjects before (left) and after (right) the six-week meditation intervention, with separate models trained for before and after intervention.** Each point represents one 4-second EEG window, colored by meditation technique (HK: red, SA: blue, BF: green). UMAP is fitted jointly on the combined features so that both panels share a common coordinate system. Note that some of the data points (e.g., for HK) are covered by other classes' points.

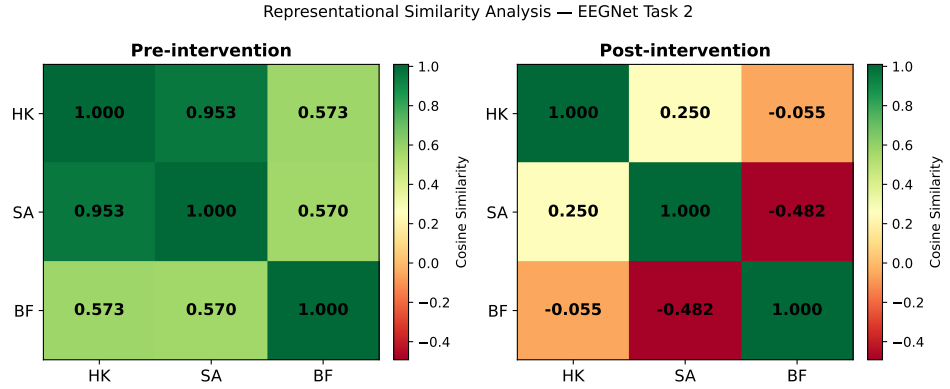

Figure 14: **Pre-intervention and post-intervention representational similarity analysis (RSA) matrices for EEGNet** illustrate the cosine similarity between the neural representation centroids of three techniques (HK, SA, and BF), where both numerical values and color gradients denote the corresponding similarity scores.

resulting performance distributions across all evaluated architectures reveal a stark asymmetry in class distinctness. When the BF cohort is isolated as the positive target class, models achieve the highest decoding metrics, distinctly surpassing theoretical chance baselines and demonstrating a statistically significant performance advantage over the alternative techniques. Conversely, when either the HK or SA cohort is designated as the singular target class, the classification performance collapses to near-chance levels. This quantitative discrepancy strictly corroborates the previously discussed representational structures. It confirms that the BF meditation technique, which relies heavily on somatic attention toward respiration, produces a highly distinct and separable neurophysiological pattern. In contrast, the HK and SA techniques, which share underlying cognitive mechanisms related to inner speech, exhibit high mutual confusability and poor independent separability. Consequently, the elevated classification efficacy when isolating the BF group demonstrates that the observed representational shifts are driven by genuine, task-specific cognitive divergences, underscoring the profound neurophysiological distinction between somatic-focused and inner-speech-focused meditation practices. As depicted in the performance boxplots (Figure 15), statistical test confirm a statistically significant advantage when the BF cohort is designated as the target class, compared to the two inner-speech-focused groups. Notably, the individual data points overlaid on these boxplots represent distinct performance observations comprehensively sampled across four model architectures, both pre- and post-intervention sessions, and five cross-validation folds. This rigorous

statistical validation further substantiates the clear neurophysiological boundary separating these distinct meditation modalities.

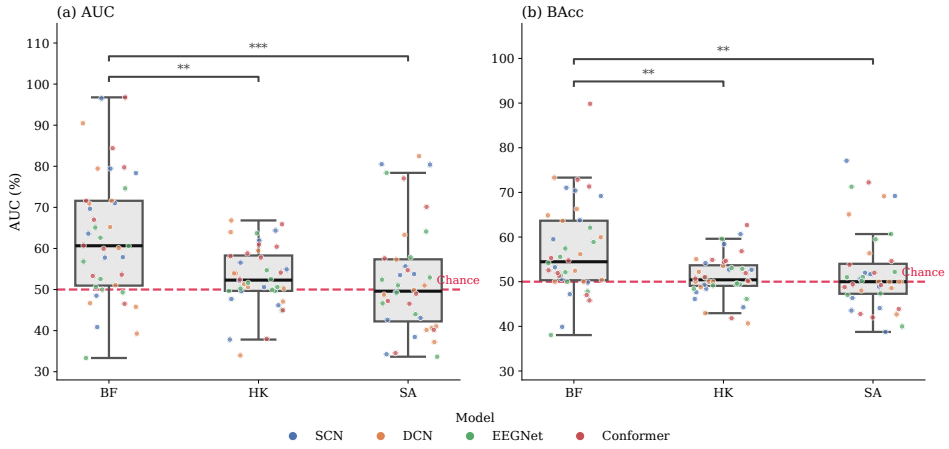

Figure 15: **Task 2 One-vs-All Classification Performance by Target Group (a)** AUC (%) and **(b)** balanced accuracy (BAcc, %) for each target group (BF, HK, SA) under the inter-subject One-Versus-All (OvA) protocol. Each box shows the median (thick black line), interquartile range (IQR), and 1.5×IQR whiskers; individual runs are overlaid as jittered dots coloured by model architecture (SCN: ShallowConvNet; DCN: DeepConvNet). Each group comprises 40 observations (4 models × 5 folds × 2 sessions). The red dashed line marks the chance level (50%) expected for a binary classifier on balanced classes. Significance brackets above the boxes indicate one-sided Mann-Whitney U tests (BF > HK and BF > SA): $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, *ns* not significant ($p \geq 0.05$).

D.3 Benchmark Task 3 Supplementary Material

Monotonic Performance Gains And The Superiority Of Full Fine-Tuning The evaluation of cross-session generalization, where models trained on pre-intervention data are tested on post-intervention sessions, reveals critical insights into few-shot adaptation strategies (Figure 16). As demonstrated by the learning trajectories across all evaluated architectures, including ShallowConvNet, DeepConvNet, EEGNet, and EEG-Conformer, the area under the curve exhibits a consistent monotonic increase as the number of available training shots grows from ten to thirty. Notably, the zero-shot reference baseline remains substantially lower than both adaptation curves, indicating that pre-intervention representations alone are insufficient for optimal cross-session transfer due to natural temporal distribution shifts. However, introducing even a minimal ten-shot calibration yields significant performance recoveries. In comparing adaptation paradigms, full fine-tuning consistently and decisively outperforms linear probing across all shot capacities and model architectures. This persistent gap highlights that updating the entire hierarchical weight structure is fundamentally necessary to capture the subtle neurophysiological changes induced over time, whereas restricting updates to the final classification layer via linear probing limits the capacity of the model to adapt to newly emerged, technique-specific data distributions.

Universal Subject-Level Efficacy And Architecture-Agnostic Adaptation The performance advantage of full fine-tuning over linear probing extends beyond aggregate metrics, demonstrating strong consistency at the individual subject level (Figure 17). An analysis of the paired scatter distribution under the 30-shot condition reveals that the vast majority of the 176 measurements (derived from 44 subjects evaluated across 4 models) lie above the parity line. The overall effectiveness of full fine-tuning suggests that updating the feature extractor is generally necessary to capture the shifts in the temporal features of EEG signals induced by short-term training interventions. However, the minority of instances where linear probing outperforms full fine-tuning warrants further explanation. Excluding extreme cases where both evaluation methods exhibit poor performance, which is likely attributable to degraded signal quality, the majority of data points favoring linear probing cluster closely around the parity line. This observation indicates that short-term meditation training did not induce substantial changes in the EEG signals for these specific subjects. Under such conditions, the initial feature representations remain adequate, allowing linear probing to achieve

competitive performance while avoiding the potential overfitting risks associated with full network updates.

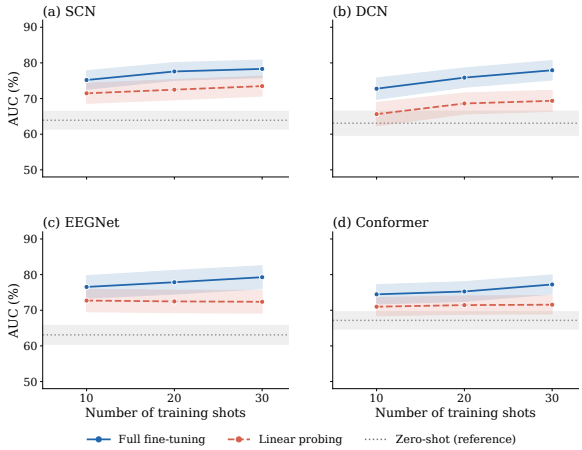

Figure 16: **Few-Shot Adaptation Learning Curves for Task 3 ($N = 44$ subjects).** Each panel shows mean AUC (%) as a function of training set size (10, 20, and 30 labeled samples) for one model architecture: **(a)** SCN (ShallowConvNet), **(b)** DCN (DeepConvNet), **(c)** EEGNet, and **(d)** EEGConformer. Solid blue lines denote full fine-tuning (all model weights updated); dashed red lines denote linear probing (backbone frozen, only the classification head retrained). Shaded bands represent ± 1 standard error of the mean across subjects. The grey dotted horizontal line and shaded region show the zero-shot baseline (mean ± 1 SE), where the pre-trained model is evaluated directly without any subject-specific adaptation. All panels share the same axes to facilitate cross-model comparison. Full fine-tuning consistently outperforms linear probing across all models and shot counts, and both strategies substantially exceed the zero-shot baseline even at 10 shots.

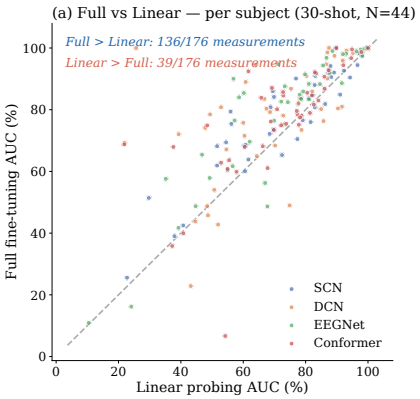

Figure 17: **Per-Subject Comparison of Full Fine-Tuning vs. Linear Probing at 30 Shots ($N = 44$ subjects).** Each point represents one subject–model pair ($44 \times 4 = 176$ observations); the x-axis shows AUC (%) under linear probing and the y-axis under full fine-tuning. Points are coloured by model architecture (SCN: ShallowConvNet; DCN: DeepConvNet). The dashed diagonal line marks equality ($y = x$): points above indicate that full fine-tuning outperforms linear probing for that subject–model pair, while points below indicate the reverse. Inset text reports the counts of pairs in each region. The systematic concentration of points above the diagonal confirms that the advantage of full fine-tuning is consistent across subjects and model architectures, rather than driven by a subset of outliers.

Acknowledgments

We would like to acknowledge Devin O'Rourke and Sidharth Chhabra from The Harmony Collective, Ypsilanti, Michigan for their expert guidance in meditation training. We also extend our gratitude to Michigan State University students Ab Basit Rafi Syed, Pratham Pradhan, Annie Wozniak, Vu Song Thuy Nguyen, Genevieve Orlewicz, and Alisia Coipel for their valuable assistance with data collection.

Ethics statement

All experimental procedures were approved by the Institutional Review Board (IRB) of Michigan State University. All participants provided written informed consent prior to the commencement of the study. The privacy rights of all human subjects have been observed throughout the research.