A MULTI-INSTITUTIONAL MULTIMODAL EEG BENCHMARK FOR FOUNDATION MODEL GENERALIZATION AND EARLY NEUROLOGICAL DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in deep learning have accelerated the development of foundation models (FMs) for electroencephalography (EEG), with significant efforts devoted to assembling EEG datasets and training large-scale models. However, existing EEG datasets remain highly fragmented and non-standardized, with limited regional diversity since most originate from the United States. Similarly, current EEG foundation models are trained on different datasets without consistent protocols, making it difficult to compare architectures fairly. Moreover, all existing models are trained exclusively on unimodal EEG signals, limiting their clinical utility, as many downstream diagnostic tasks, such as detecting neurodegenerative diseases, require integration of additional modalities beyond EEG. To address these limitations, we introduce, for the first time M-EEG, a multimodal EEG dataset comprising over 6000 patients collected from two major hospitals outside the US. In parallel, we unify all existing public EEG datasets into a single standardized corpus, enabling the first rigorous benchmarking of state-of-the-art EEG foundation model architectures under consistent pretraining and fine-tuning pipelines. Finally, using our multimodal EEG dataset, we design and evaluate a multimodal diagnostic model, demonstrating that integrating auxiliary modalities (e.g., blood biomarkers and clinical notes) with EEG substantially improves downstream prediction accuracy, for instance, achieving a 27.64% gain in Alzheimer's disease risk prediction.

1 INTRODUCTION

Background. Recent breakthroughs in deep learning have catalyzed the development of foundation models (FMs) for electroencephalography (EEG) Wang et al. (2025; 2024a;b); Yang et al. (2023); Kostas et al. (2021), with the goal of learning transferable neural representations across diverse clinical and cognitive tasks. In parallel, efforts have been made to assemble large-scale clinical EEG corpora from multiple hospitals (Khan et al., 2022; Zhang et al., 2018; Sun et al., 2025), aiming to broaden regional and clinical diversity and to better capture the inherently non-stationary, low signal-to-noise characteristics of EEG. Despite these encouraging developments, existing EEG datasets and foundation models continue to face significant limitations.

Limitations of existing EEG datasets and foundation models. On the dataset side, available corpora remain fragmented: most are heavily US-centric (Obeid & Picone, 2016; Sun et al., 2025), task-specific (Zhang et al., 2018), or involve relatively few subjects (Khan et al., 2022). Such constraints exacerbate overfitting when applying self-supervised pretraining methods, such as mask prediction (Wang et al., 2024b;a; 2025; Yang et al., 2023) or contrastive learning (Yang et al., 2023; Kostas et al., 2021), which depend critically on a wide subject pool to generate reliable positive and negative pairs. Moreover, most datasets lack integration with minimally invasive modalities such as blood-based biomarkers, which could be combined with EEG to strengthen diagnostic accuracy. The recently introduced Harvard Electroencephalography Database (Sun et al., 2025) partially addresses these limitations by releasing nearly three million hours of data from four hospitals, yet it remains entirely US-based and thus insufficient for studying regional diversity at scale.

Concerning the EEG foundation models, current models (e.g., EEGPT(Wang et al., 2024a), BIOT(Yang et al., 2023), CBraMOD(Wang et al., 2025)) exhibit two fundamental limitations: limited regional diversity and restricted clinical relevance. First, most models are pretrained on only a handful of public datasets—largely from single regions, resulting in poor generalization across demographic, clinical, and recording variations. Performance drops sharply when evaluated on unseen regional datasets, underscoring their vulnerability to distribution shifts (See Fig. 2). Dataset heterogeneity in sampling rates, channel layouts, and annotation protocols further complicates the establishment of standardized pretraining pipelines, reinforcing the need for a harmonized and globally representative corpus. Second, existing foundation models are trained exclusively on unimodal EEG signals, whereas real-world diagnosis of complex brain disorders, such as Alzheimer's disease, often requires multimodal integration, including minimally invasive biomarkers like blood-based tests. As illustrated in Table 5, incorporating auxiliary signals substantially improves disease risk prediction performance over EEG alone, reinforcing the need for multimodal foundation modeling. Yet, there remains a scarcity of public EEG datasets that are both regionally diverse and enriched with complementary clinical modalities.

Our approach. To address these gaps, we present VEEG, a large-scale, clinically annotated EEG dataset collected from two major hospitals outside of US, comprising $1,170$ hours of EEG recordings from $6,081$ patients. To our knowledge, this is the largest non-US clinical EEG corpus to date, offering significant improvements in geographic coverage, subject diversity, and clinical complexity. In addition, a unique subset of VEEG includes paired EEG, blood biomarkers, and clinical notes, enabling the first non-US multimodal benchmark for EEG–lab fusion.

Building on VEEG, we conduct a standardized benchmarking study of state-of-the-art EEG foundation models under identical pretraining and fine-tuning protocols across diverse clinical tasks drawn from both US-based and non-US datasets. Our findings demonstrate that pretraining on VEEG yields stronger generalization across regions and diseases, with clear gains on challenging diagnostic tasks such as early Alzheimer's risk prediction.

Contributions. Our contributions are summarized as follows:

- We release M-EEG, a large-scale clinical EEG corpus with 1,170 hours from 6,081 patients at two major hospitals, marking the largest non-US EEG dataset by subject count and improving the diversity of EEG pretraining resources. Furthermore, we curate a subset of M-EEG that integrates EEG signals with blood-based biomarkers and clinical notes, establishing the first non-US multimodal EEG benchmark and opening new avenues for research in EEG-laboratory data fusion.
- We standardize all existing EEG datasets to construct a unified large-scale corpus and establish a benchmark to compare state-of-the-art EEG foundation model architectures on this dataset. To the best of our knowledge, this is the first standardized large-scale EEG corpus, and our work represents the first systematic benchmarking of EEG foundation models on a common dataset using consistent pretraining and fine-tuning pipelines, thereby enabling rigorous and dataset-independent comparison.
- We introduce a multimodal EEG model for early disease diagnosis. Experiments conducted on our proposed multimodal EEG dataset, validated through Alzheimer's risk prediction, demonstrate that incorporating additional modalities substantially enhances prediction accuracy.

2 Existing datasets and EEG foundation models

2.1 Current Pretraining Corpora

Table 1 provides an overview of major EEG datasets used for representation learning, emphasizing their scale, geographic coverage, and any multimodal extensions. The field currently relies on a patchwork of hospital-based clinical EEG corpora as the backbone for foundation model pretraining.

Foremost among these is the Temple University Hospital (TUH) corpus (Obeid & Picone, 2016), which at roughly 24,000 hours of recordings from a single US hospital has underpinned much of the recent progress in self-supervised EEG representation learning (Wang et al., 2025; Han et al., 2025).

More recently, the Harvard Electroencephalography Database (HEEDB) (Sun et al., 2025) introduced an unprecedentedly large corpus on the order of millions of EEG hours, drawn from multiple US hospitals and enriched with patient metadata and auxiliary modalities, integrating demographics, medication records, lab values, and free-text clinical notes (including blood-based biomarkers). This rich multimodal resource significantly expanded data scale and scope; however, it remains entirely US-based, exacerbating a persistent regional diversity gap in EEG data. Beyond the United States, only a few smaller clinical corpora have been released. For example, the NMT-Scalp dataset from Pakistan (Khan et al., 2022) provides valuable clinical EEG data but remains limited in scale, with relatively few hours and subjects compared to TUH or HEEDB.

In addition to clinical datasets, a variety of laboratory or task-specific EEG datasets have been used for representation learning. Notable examples include SEED (Zheng & Lu, 2015) for emotion recognition, PhysioNet MI (Goldberger et al., 2000) for motor imagery, M3CV (Huang et al., 2022) for cognitive workload, HGD (Schirrmeister et al., 2017) for brain–computer interface trials, and SHHS (Zhang et al., 2018) for sleep monitoring. While each contributes valuable data for its specific domain, these datasets are relatively small in scale (often involving only tens of subjects or a few dozen hours) and narrow in clinical scope. Moreover, they are typically single-modality (EEG only) and collected under disparate protocols.

2.2 Existing EEG Foundation Models

2.2.1 Unimodal EEG-Based Foundation Models

EEG foundation models aim to learn general-purpose neural representations from large corpora without relying on task-specific labels. Table 6 summarizes representative architectures and their original pretraining data.

Two open-source efforts, **BENDR** (Kostas et al., 2021) and **CBraMOD** (Wang et al., 2025), were trained exclusively on the TUH clinical corpus, leveraging the breadth of U.S. hospital EEG recordings to drive self-supervised learning objectives. These works established TUH as the standard backbone for EEG foundation modeling. By contrast, **EEGPT** (Wang et al., 2024a) expanded beyond a single corpus by pretraining on a composite of multiple laboratory datasets, including PhysioNet MI, SEED, M3CV, HGD, and TSU to capture a wider spectrum of motor imagery and cognitive tasks. Similarly, **LaBraM** (Jiang et al., 2024) aggregated a heterogeneous collection of public corpora (e.g., TUEG subsets, BCIC IV-1, EmoBrain, Inria BCIC, SPIS Resting) together with private data, aiming to maximize training diversity through scale and variety. Another line of work has drawn on large-scale clinical cohorts beyond TUH. **BIOT** (Yang et al., 2023), for instance, leverages both SHHS, a population-level sleep study, and a small subset of HEEDB collected at Massachusetts General Hospital to pretrain a transformer architecture designed for cross-dataset generalization. Unlike models tied to narrowly defined tasks, BIOT emphasizes scalability across heterogeneous clinical EEG corpora, though its training sources remain limited to US-based datasets (with only a small subset of HEEDB included).

Despite their architectural differences and varying objectives, a common limitation is that each foundation model was developed using a distinct, and often narrow, pool of pretraining data. This inconsistency makes reported improvements difficult to attribute: performance gains may arise as much from the scale, scope, or bias of the underlying corpus as from innovations in model design. Consequently, direct comparison across models remains problematic without a unified and standardized pretraining benchmark.

2.2.2 Toward Multimodal EEG Foundation Models

In clinical practice, EEG is rarely interpreted in isolation. Neurologists routinely contextualize EEG findings with additional information such as blood biomarkers (indicating infection, inflammation, or metabolic abnormalities), routine laboratory test results, and clinical notes that capture patient history and diagnostic impressions. In many neurological disorders, further confirmation may require complex and costly procedures such as MRI, which highlights the value of minimally invasive signals that can complement EEG in a more accessible way. These auxiliary data sources provide critical context that can help disambiguate EEG abnormalities and improve diagnostic accuracy.

Table 1: **Existing EEG pretraining corpora.** BBB denotes blood-based biomarkers. Dataset names are color-coded as follows: blue for general clinical EEG corpora, brown for task-specific corpora, and **bold** for our contribution (**M-EEG**).

Dataset name	Region	# Hours	# Subjects	# Sites	# Channels	Sampling (Hz)	Modalities	
							BBB	Clinical notes
HEEDB (Sun et al., 2025)	US	3 000 000	109 178	4	22–57	200–512	✓	✓
TUEG (Obeid & Picone, 2016)	US	24 000	10 874	1	31	250–256	✗	✗
NMT Scalp (Khan et al., 2022)	Pakistan	625	60	1	19	200	✗	✗
M3CV (Huang et al., 2022)	China	90	106	1	64	250	✗	✗
SEED series (Zheng & Lu, 2015)	China	200 (total)	8–20	1	62	1000	✗	✗
PhysioNet MI (Goldberger et al., 2000)	US	47	109	1	64	160	✗	✗
Inria BCIC (Margaux et al., 2012)	France	30	26	1	56	200	✗	✗
BCIC IV-1 (Blankertz et al., 2007)	Europe	8	7	1	59	1000	✗	✗
HGD (Schirrmeister et al., 2017)	China	15	154	1	128	500	✗	✗
Raw EEG Data (Trujillo, 2020)	US	34	48	1	64	256	✗	✗
Grasp and Lift (Luciw et al., 2014)	UK	12	12	1	32	500	✗	✗
EmoBrain (Savran[1] et al., 2006)	Germany	5	16	1	64	1024	✗	✗
Resting State (Trujillo et al., 2017)	US	3	22	1	72	256	✗	✗
SPIS Resting (Torkamani-Azar et al., 2020)	China	1	10	1	64	2048	✗	✗
Target vs Non-Target (Korczowski et al., 2019)	France	16	43	1	32	512	✗	✗
TSU (Wang et al., 2016)	China	14	35	1	64	250	✗	✗
SHHS (Zhang et al., 2018)	US	43 446	5 804	–	2	125	✗	✗
Siena Scalp (Detti, 2020)	Italy	30	14	1	29	512	✗	✗
M-EEG	Outside of US	1 170	6 081	2	22–44	200, 500	✓	✓

Despite this reality, existing EEG foundation models remain strictly unimodal, trained only on raw EEG signals without auxiliary modalities. This limitation reduces their clinical utility: a model that sees only EEG may miss critical disease indicators that would be apparent if combined with supporting evidence such as blood tests or clinical reports.

Extending pretraining corpora beyond EEG is therefore essential for developing foundation models that generalize across diverse clinical scenarios. Incorporating modalities such as blood-based biomarkers and textual clinical records into EEG representation learning can capture patterns more consistent with real-world diagnostic reasoning (Moretti, 2015; Chetty et al., 2024), potentially improving performance on tasks like early detection of neurodegenerative diseases or prognostication after brain injury.

These considerations motivate the collection of multimodal EEG datasets that combine electrophysiological signals with complementary clinical information. In the next section, we present **M-EEG**, a multi-institutional dataset that pairs EEG recordings with blood biomarkers and clinical notes, and introduce a unified benchmarking framework for evaluation. Together, these contributions expand regional coverage, integrate multimodal context, and enable fair, standardized assessment of EEG foundation models.

3 Multi-institutional multimodal EEG Dataset

In the following, we introduce a multi-institutional EEG dataset that has been systematically compiled and meticulously curated to support advanced research in computational neuroscience. The dataset comprises three main components.

The first is **M-EEG** (Section 3.1), our in-house multimodal dataset collected outside the United States, which includes synchronized EEG recordings alongside corresponding blood test results. This multimodal dataset not only enhances the diversity of existing EEG data populations, thereby improving the generalizability of EEG foundation models (as demonstrated in Section 4.3), but also leverages its multimodal nature to boost performance on downstream tasks, as will be further discussed in Section 4.4.

The second component, **P-EEG** (Section 3.2), is a unified public dataset constructed through the aggregation and harmonization of multiple publicly available EEG datasets. It is designed specifically for the pretraining of EEG foundation models. By standardizing data formats and preprocessing

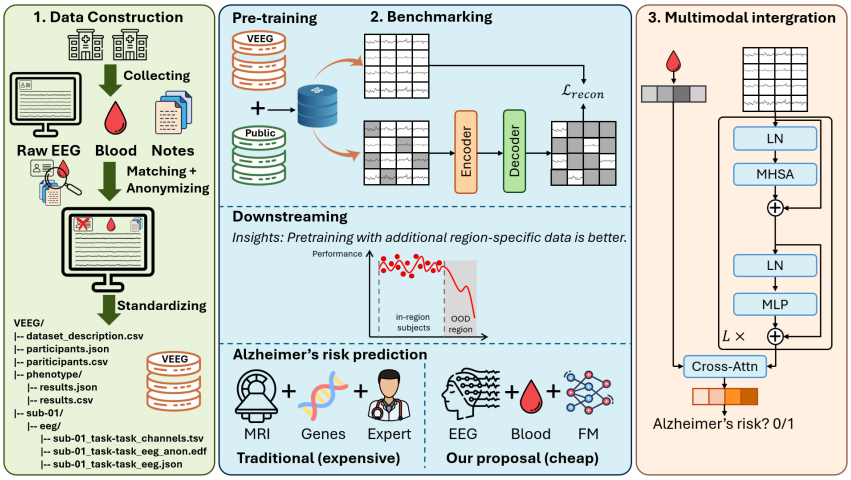

Figure 1: Overview of M-EEG. (1) **Data construction:** raw EEG, blood biomarkers, and clinical notes collected from two hospitals are anonymized and standardized into BIDS format. (2) **Benchmarking:** M-EEG enables large-scale pretraining and standardized evaluation of EEG foundation models, with downstream results showing that region-specific data improves *regional robustness*. (3) **Multimodal integration:** M-EEG includes paired EEG–blood data, allowing exploration of multimodal foundation models for clinical tasks such as early disease risk prediction.

pipelines, this unified corpus offers a robust, scalable, and reproducible benchmark for training, evaluating, and comparing foundation models in EEG-based machine learning research.

Finally, the **T-EEG** component is derived from publicly available task-oriented datasets. It is specifically curated to evaluate the performance of foundation models on a range of targeted downstream tasks.

3.1 M-EEG: AN IN-HOUSE MULTI-INSTITUTIONAL, MULTIMODAL EEG DATASET

We construct M-EEG, a multi-institutional, multimodal EEG dataset, collected from two major hospitals located outside the United States. The primary objective of this dataset is to enhance the diversity of existing EEG datasets, both in terms of geographical representation (regional diversity) and data modality. Using this dataset, we demonstrate that regional diversity plays a critical role in improving EEG representation learning for foundation models, while incorporating additional modalities beyond EEG, such as blood biomarkers, significantly boosts the accuracy of brain-related disease prediction.

The construction of M-EEG involved several key steps: (1) raw data acquisition, (2) cross-modality synchronization, and (3) standardized data preprocessing. We present more details about each step in Appendix A.7 and Figure 1.

M-EEG advances beyond prior corpora by providing the largest non-US clinical EEG cohort to date, comprising 6,081 patients and **1,170** hours of recordings collected over a multi-year period across two hospitals. Each record follows a standardized 10–20 montage with 22–44 channels and sampling rates of 200 or 500 Hz. In addition to raw EEG, the dataset includes paired blood-based biomarkers (BBB) and clinical notes, enabling multimodal representation learning. The cohort covers a wide spectrum of neurological conditions such as epilepsy, encephalopathy, sleep disorders, and neurodegenerative diseases, reflecting real-world clinical diversity. All data are harmonized into a BIDS-compliant release to ensure accessibility and reproducibility.

3.2 P-EEG: A UNIFIED EEG CORPUS FOR FOUNDATION MODEL PRETRAINING

To establish a fair and comprehensive benchmark for foundation model pretraining, we aggregate multiple publicly available EEG datasets and integrate them with our proprietary VEEG dataset to construct a unified corpus, referred to as **P-EEG**, specifically tailored for the training and evaluation of EEG foundation models.

Although a wide range of public EEG datasets exist, each is originally created for distinct research purposes. Therefore, we carefully select only those datasets that align with the objectives and requirements of foundation model training. In the following sections, we detail the criteria used for dataset selection and describe the preprocessing pipeline employed to harmonize and standardize the selected datasets into a coherent and consistent format.

3.2.1 DATASET SELECTION

We selected datasets from Table 1 based on two main criteria: (i) a focus on patient-based clinical recordings rather than task-specific paradigms, and (ii) the ability to ensure both biological and regional diversity while maintaining sufficient EEG channel coverage.

Specifically, we excluded task-oriented datasets, highlighted in brown in Table 1, as they are tailored to narrow cognitive or motor tasks, which can bias representation learning toward predefined downstream objectives. Although the SHHS dataset (Zhang et al., 2018) offers a large sample size, it records only two EEG channels in a sleep-specific context, limiting its applicability for general-purpose pretraining. We also deferred the inclusion of the HEEDB dataset (Sun et al., 2025) due to its massive scale and the ongoing integration process, reserving it for future work.

As a result, the unified dataset, P-EEG, comprises three complementary corpora: the Temple University EEG (TUEG) dataset (Obeid & Picone, 2016), the NMT Scalp EEG dataset from Pakistan (Khan et al., 2022), and our newly introduced dataset, M-EEG. Together, these datasets span multiple hospitals, geographic regions, and acquisition protocols, forming a diverse yet clinically grounded corpus for the training and evaluation of EEG foundation models.

3.2.2 DATA PREPROCESSING AND HARMONIZATION

Our preprocessing largely follows CBraMOD (Wang et al., 2025) to reduce variability and remove noise. We discard the first and last minute of TUEG recordings, retain 19 common 10–20 channels, and apply a 0.3–75 Hz band-pass filter plus a 60 Hz notch filter. Signals are resampled at 200 Hz, segmented into 30 s windows, and normalized to $[-1, 1]$ after excluding samples with amplitudes above $100, \mu V$ (Yin et al., 2025). For NMT-Scalp (Khan et al., 2022) and M-EEG, we apply the same pipeline but use a 50 Hz notch filter and Independent Component Analysis (ICA) (Makeig et al., 1995) to further suppress artifacts.

3.3 T-EEG: A TASK-ORIENTED EEG BENCHMARK FOR DOWNSTREAM EVALUATION

Downstream BCI Tasks and Datasets. T-EEG serves as a task-oriented benchmark designed to systematically evaluate the generalization of EEG foundation models across diverse downstream applications. We include six representative tasks spanning seven EEG datasets, as summarized in Table 7. The benchmark covers well-established challenges in brain–computer interface and clinical EEG analysis: motor imagery (BCIC-2a (Blankertz et al., 2007)), sleep staging (SleepEDF (Kemp et al., 2000)), seizure detection (TUEV (Obeid & Picone, 2016)), and abnormal EEG classification (TUAB (Obeid & Picone, 2016)). To evaluate robustness under regional shifts, we further incorporate A&MISP (Ma Thi et al., 2025), ALS (Ngo et al., 2024), and N-FM (Neurought, 2023), which introduce distinct recording conditions and subject populations. Finally, to assess multimodal integration, we include the external PEARL dataset (Dzianok & Kublik, 2024) for Alzheimer's risk prediction, where paired EEG and blood biomarkers enable evaluation of multimodal representation learning.

Preprocessing pipeline. Given the heterogeneity of real-world EEG collections, the datasets in T-EEG vary substantially in sampling frequency, number of channels, and segment duration. To ensure fair comparison, we establish a standardized preprocessing pipeline: linear channel mappings are applied when necessary to align with the pretrained 19-channel montage, and signals are adaptively truncated or segmented around task-specific annotations to extract meaningful samples. Table 7 details the preprocessing setup for each dataset, with further descriptions provided in Appendix A.

Table 2: Performance of EEG foundation models pretrained on the unified corpus P-EEG and finetuned on task-oriented dataset T-EEG. Results are reported on representative downstream benchmarks.

Task	Architecture	Balanced Accuracy ↑	Cohen's Kappa / AUPR ↑	Weighted F1 / AUROC ↑
BCIC-2a	**CBraMOD**	0.4978	0.3304	0.4856
	EEGPT	**0.5374**	**0.3823**	**0.5138**
TUEV	**CBraMOD**	0.4449	0.5114	0.7394
	EEGPT	**0.5217**	**0.5581**	**0.7680**
TUAB	**CBraMOD**	0.6175	0.4384	0.6897
	EEGPT	**0.8018**	**0.8800**	**0.8826**
Sleep-EDFx	**CBraMOD**	**0.7512**	**0.7258**	**0.7978**
	EEGPT	0.6585	0.5963	0.6976

4 EEG Foundation Model Benchmarking

In this section, using the UEEG dataset, we conduct a series of experiments to address three key research questions: (1) How do state-of-the-art EEG foundation models compare in performance? (Section 4.2); (2) How effective is the VEEG dataset for pretraining EEG foundation models? (Section 4.3); (3) To what extent does incorporating multimodality improve performance on EEG-related downstream tasks? (Section 4.4).

4.1 Experiment Settings

Baselines. We include two state-of-the-art EEG foundation models as baselines. (1) **CBraMOD** (Wang et al., 2025), a reconstruction-based model was originally pretrained on TUH (TUEG). (2) **EEGPT** (Wang et al., 2024a), a multi-corpus model was originally pretrained on laboratory datasets including PhysioNet MI (Goldberger et al., 2000), SEED (Zheng & Lu, 2015), M3CV (Huang et al., 2022), HGD (Schirrmeister et al., 2017), and TSU (Wang et al., 2016).

Tasks. We evaluate foundation models on the downstream tasks defined in T-EEG (section 3.3), spanning both multiclass and binary classification settings. More details for each task are described in Appendix A.

Metrics. To ensure consistent and interpretable evaluation across tasks, we report performance using metrics tailored to the nature of each dataset. For **multiclass classification** tasks (BCIC-2a, SleepEDF, TUEV, A&MISP, ALS, N-FM), we compute Balanced Accuracy, Cohen's Kappa, and Weighted F1, which account for class imbalance and provide a comprehensive view of classification quality. For **binary classification** tasks (TUAB and PEARL), we report Balanced Accuracy together with AUROC and AUPR, as these metrics are more informative under skewed class distributions.

4.2 Model Comparison

We begin by comparing representative EEG foundation model architectures under a unified pretraining setup. Specifically, all models are pretrained on the P-EEG dataset and then finetuned on the T-EEG dataset.

We report results on four widely recognized tasks, BCIC-2a, TUEV, TUAB, and SleepEDF, spanning distinct BCI tasks, including motor imagery, seizure detection, abnormal EEG classification, and sleep staging. Together, these benchmarks cover both cognitive and clinical applications and provide complementary perspectives on model generalization. Results are summarized in Table 2.

Overall, EEGPT tends to outperform CBraMOD across diverse tasks, likely because its auxiliary alignment loss mitigates mode collapse and yields more discriminative representations, whereas CBraMOD relies solely on masked prediction

4.3 Impacts of regional data

As illustrated in Fig. 2, on BCIC-2a, which shares characteristics with the pretraining data described in Table 6, both CBraMOD and EEGPT achieve justifiable performance (balanced accuracy: 0.49 vs. 0.51, Kohen's kappa: 0.32 vs. 0.34, weighted F1: 0.47 vs. 0.49). In contrast, on A&MISP,

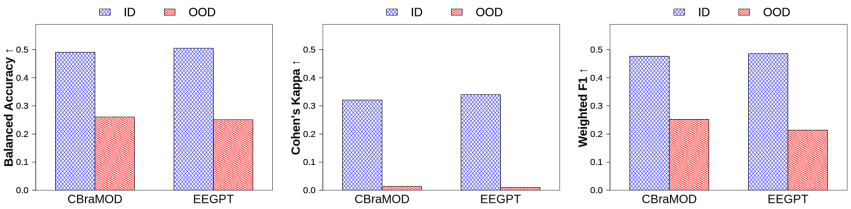

Figure 2: Performance comparison on 4-class motor imagery tasks under in-region (ID) and out-of-region (OOD) settings. BCIC-2a serves as the ID dataset, whereas A&MISP represents the OOD dataset from the region represented in M-EEG.

collected under different regional conditions, performance collapses, with balanced accuracy and F1 reduced by nearly 50% and kappa by more than 95%. To examine regional robustness, we split P-EEG into two subsets: an out-of-region set collected from the same geographic area as M-EEG, and an in-region set collected elsewhere. We then design two experiments: (1) adding M-EEG should not downgrade the performance of models trained on the in-region subset (Table 3), and (2) adding M-EEG should improve the performance of models trained on the out-of-region subset (Table 4).

Table 3 shows that incorporating M-EEG does not degrade performance on the *in-region* subset. Across BCIC-2a, TUAB, and TUEV, most metrics either improve or remain stable. For instance, CBraMOD gains +17.20% balanced accuracy on TUEV and +4.41% on TUAB, while EEGPT improves by +6.39% on BCIC-2a. The few decreases (e.g., EEGPT on Sleep-EDFx, below 3% on secondary metrics) are marginal and do not alter the overall trend. These results confirm that adding M-EEG preserves accuracy on benchmarks that have traditionally anchored EEG foundation model comparisons, ensuring continuity with prior work and demonstrating that regional diversity does not harm in-region tasks.

Table 4 highlights the *out-of-region* subset, where the benefits of M-EEG pretraining are pronounced. Both CBraMOD and EEGPT consistently improve, with substantial relative gains on A&MISP (+8.37% balanced accuracy and +190% Cohen's κ for EEGPT) and ALS (+3.74% BA and +19.43% κ for EEGPT). Even on the high-performing N-FM dataset, where baselines approach ceiling, CBraMOD achieves a +3.92% improvement in balanced accuracy. These findings show that regional coverage not only maintains comparability on in-region tasks but also directly enhances robustness when models are transferred to populations and recording conditions absent from US-centric corpora.

4.4 Impacts of multimodality data

Multimodal fusion. We integrate blood test results with EEG via a simple cross-attention module: blood biomarkers are projected into the EEG embedding space and used as queries to attend over EEG tokens. More details are presented in Appendix A.3.

Experiments results. Table 5 reports Alzheimer's risk prediction on the PEARL dataset across three tasks: MSIT, SMT, and RST. Incorporating blood-based biomarkers alongside EEG consistently improves performance for both CBraMOD and EEGPT. On MSIT, adding BBB yields relative gains of +27.6% balanced accuracy and +37.4% AUPR for CBraMOD, and comparable improvements for EEGPT (+25.1% and +37.6%). Importantly, this +27.6% gain is observed in a setting where the unimodal EEG baseline already achieved balanced accuracy above 0.5, i.e., better than random guessing, underscoring the substantial added value of multimodal integration.

Our preliminary findings demonstrate clear improvements in risk prediction, motivating future work on developing foundation models that seamlessly integrate EEG with other minimally invasive modalities.

5 Conclusion

In this study, we present M-EEG, a novel multimodal EEG dataset collected from two hospitals outside the United States. To support large-scale modeling, we further curated and standardized

Table 3: Comparison of EEG foundation models pretrained on the original datasets versus those trained on **P-EEG**, considering datasets from the different regions with M-EEG.

Tasks	Architectures		Balanced Accuracy ↑		Cohen's Kappa / AUPR ↑		Weighted F1 / AUROC ↑	
			Performance	Gain	Performance	Gain	Performance	Gain
BCIC-2a	**CBraMOD**	Base	0.4907		0.3210		0.4766	
		P-EEG	0.4978	+1.45%	0.3304	+2.93%	0.4856	+1.89%
	EEGPT	Base	0.5051		0.3402		0.4860	
		P-EEG	0.5374	+6.39%	0.3823	+12.38%	0.5138	+5.10%
TUEV	**CBraMOD**	Base	0.3796		0.4734		0.7162	
		P-EEG	0.4449	+17.20%	0.5114	+8.03%	0.7394	+3.24%
	EEGPT	Base	0.5431		0.5361		0.7481	
		P-EEG	0.5217	-3.93%	0.5581	+4.10%	0.7680	+ 2.66%
TUAB	**CBraMOD**	Base	0.5914		0.5685		0.6230	
		P-EEG	0.6175	+4.41%	0.6167	+8.48%	0.6527	+4.77%
	EEGPT	Base	0.7891		0.8749		0.8708	
		P-EEG	0.8018	+1.61%	0.8800	+0.58%	0.8826	+1.36%
Sleep-EDFx	**CBraMOD**	Base	0.7390		0.7316		0.8000	
		P-EEG	0.7512	+1.65%	0.7258	-0.79%	0.7978	-0.28%
	EEGPT	Base	0.6356		0.6117		0.7062	
		P-EEG	0.6585	+3.60%	0.5963	-2.52%	0.6976	-1.22%

Table 4: Comparison of EEG foundation models pretrained on the original datasets versus those trained on **P-EEG**, considering datasets from the same region as M-EEG.

Tasks	Architectures		Balanced Accuracy		Cohen's Kappa		Weighted F1	
			Performance	Gain	Performance	Gain	Performance	Gain
A&MISP	**CBraMOD**	Base	0.2604		0.0136		0.2523	
		P-EEG	0.2715	+4.26%	0.0286	+110.29%	0.2494	-1.14%
	EEGPT	Base	0.2507		0.0100		0.2138	
		P-EEG	0.2716	+8.37%	0.0290	+190.00%	0.2234	+4.49%
ALS	**CBraMOD**	Base	0.3706		0.1930		0.4047	
		P-EEG	0.3715	+0.24%	0.2018	+4.56%	0.4019	-0.69%
	EEGPT	Base	0.3448		0.1549		0.3733	
		P-EEG	0.3577	+3.74%	0.1850	+19.43%	0.3843	+2.95%
N-FM	**CBraMOD**	Base	0.9192		0.9183		0.9187	
		P-EEG	0.9553	+3.92%	0.9548	+3.97%	0.9551	+3.96%
	EEGPT	Base	0.9979		0.9979		0.9978	
		P-EEG	0.9989	+0.10%	0.9990	+0.11%	0.9989	+0.11%

Table 5: Alzheimer's risk prediction on the PEARL dataset. We compare unimodal EEG (w/o BBB) with multimodal EEG plus blood-based biomarkers (w/ BBB) with teal denotes the relative improvements over the EEG-only baseline.

Tasks	Architectures		Balanced Accuracy		AUPR		AUROC	
			Performance	Gain	Performance	Gain	Performance	Gain
PEARL–MSIT	**CBraMOD**	w/o BBB	0.5283		0.5523		0.5877	
		w/ BBB	0.6743	+27.64%	0.7588	+37.39%	0.7779	+32.36%
	EEGPT	w/o BBB	0.4615		0.4285		0.4063	
		w/ BBB	0.5774	+25.11%	0.5895	+37.57%	0.5976	+47.08%
PEARL–SMT	**CBraMOD**	w/o BBB	0.5296		0.4692		0.5040	
		w/ BBB	0.6288	+18.73%	0.6774	+44.37%	0.7156	+41.98%
	EEGPT	w/o BBB	0.4746		0.4132		0.4222	
		w/ BBB	0.5627	+18.56%	0.6109	+47.85%	0.5651	+33.85%
PEARL–RST	**CBraMOD**	w/o BBB	0.4504		0.4445		0.4580	
		w/ BBB	0.6960	+54.52%	0.7772	+74.84%	0.7783	+69.93%
	EEGPT	w/o BBB	0.4366		0.3925		0.3949	
		w/ BBB	0.5753	+31.77%	0.5985	+52.48%	0.5483	+38.85%

existing public EEG datasets into two complementary resources: P-EEG, designed for pretraining EEG foundation models, and T-EEG, a suite of task-oriented datasets tailored for finetuning models on specific applications. Leveraging these datasets, we conducted a comprehensive evaluation of the two most advanced EEG foundation models to date. Beyond benchmarking, we also investigated the benefits of pretraining on M-EEG, and demonstrate that incorporating multimodal EEG substantially boosts downstream predictive performance, most notably in Alzheimer's disease detection. In the future, we plan to further enrich **M-EEG** through larger-scale, longitudinal data collection and to explore foundation models that integrate EEG with multiple minimally invasive modalities, aiming toward clinically reliable multimodal foundation models.

REFERENCES

Benjamin Blankertz, Guido Dornhege, Matthias Krauledat, Klaus-Robert Müller, and Gabriel Curio. The non-invasive berlin brain–computer interface: fast acquisition of effective performance in untrained subjects. *NeuroImage*, 37(2):539–550, 2007.

Chowtapalle Anuraag Chetty, Harsha Bhardwaj, G. Pradeep Kumar, T. Devanand, C. S. Aswin Sekhar, Tuba Aktürk, Ilayda Kiyi, Görsev Yener, Bahar Güntekin, Justin Joseph, and Chinnakkaruppan Adaikkan. Eeg biomarkers in alzheimer's and prodromal alzheimer's: a comprehensive analysis of spectral and connectivity features. *Alzheimer's Research & Therapy*, 16(1): 236, 2024. doi: 10.1186/s13195-024-01582-w. URL https://alzres.biomedcentral.com/articles/10.1186/s13195-024-01582-w. Open access.

Paolo Detti. Siena scalp eeg database. *physionet*, 10:493, 2020.

Piotr Dzianok and Ewa Kublik. Pearl-neuro database: Eeg, fmri, health and lifestyle data of middle-aged people at risk of dementia. *Scientific Data*, 11(1):276, 2024. doi: 10.1038/s41597-024-03106-5.

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

Danny Dongyeop Han, Ahhyun Lucy Lee, Taeyang Lee, Yonghyeon Gwon, Sebin Lee, Seongjin Lee, David Keetae Park, Shinjae Yoo, Jiook Cha, and Chun Kee Chung. Diver-0 : A fully channel equivariant eeg foundation model, 2025. URL https://arxiv.org/abs/2507.14141.

Gan Huang, Zhenxing Hu, Weize Chen, Shaorong Zhang, Zhen Liang, Linling Li, Li Zhang, and Zhiguo Zhang. M3cv: A multi-subject, multi-session, and multi-task database for eeg-based biometrics challenge. *NeuroImage*, 264:119666, 2022.

Weibang Jiang, Liming Zhao, and Bao liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QzTpTRVtrP.

B. Kemp, A.H. Zwinderman, B. Tuk, H.A.C. Kamphuisen, and J.J.L. Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000. doi: 10.1109/10.867928.

Hassan Aqeel Khan, Rahat Ul Ain, Awais Mehmood Kamboh, Hammad Tanveer Butt, Saima Shafait, Wasim Alamgir, Didier Stricker, and Faisal Shafait. The nmt scalp eeg dataset: an open-source annotated dataset of healthy and pathological eeg recordings for predictive modeling. *Frontiers in neuroscience*, 15:755817, 2022.

Louis Korczowski, Martine Cederhout, Anton Andreev, Grégoire Cattan, Pedro Luiz Coelho Rodrigues, Violette Gautheret, and Marco Congedo. *Brain Invaders calibration-less P300-based BCI with modulation of flash duration Dataset (bi2015a)*. PhD thesis, GIPSA-lab, 2019.

Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. Bendr: using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

Matthew D Luciw, Ewa Jarocka, and Benoni B Edin. Multi-channel EEG recordings during 3,936 grasp and lift trials with varying weight and friction. *Scientific Data*, 1(1):1–11, 2014.

Chau Ma Thi, Hoang-Anh Nguyen The, Kien Nguyen Minh, Long Vu Thanh, Hieu Nguyen Dinh, Nhu-Y Huynh Thi, Thanh-Huong Ha Thi, Trong-Nghia Hoang Tien, Doan-Truc Au Dao, Kim-Long Nguyen Hoang, Vy Huynh Kha, and Tuyet-Linh Le Hoang. Uet175: Eeg dataset of motor imagery tasks in vietnamese stroke patients. *Frontiers in Neuroscience*, Volume 19 - 2025, 2025. ISSN 1662-453X. doi: 10.3389/fnins.2025.1580931. URL https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2025.1580931.

Scott Makeig, Anthony Bell, Tzyy-Ping Jung, and Terrence J Sejnowski. Independent component analysis of electroencephalographic data. In D. Touretzky, M.C. Mozer, and M. Hasselmo (eds.), *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1995. URL https://proceedings.neurips.cc/paper_files/paper/1995/file/754dda4b1ba34c6fa89716b85d68532b-Paper.pdf.

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

Davide Vito Moretti. Association of eeg, mri, and regional blood flow biomarkers is predictive of prodromal alzheimer's disease. *Neuropsychiatric Disease and Treatment*, 11:2779–2791, 2015. doi: 10.2147/NDT.S93253. URL https://doi.org/10.2147/NDT.S93253.

Neurought. 94 vietnamese characters eeg dataset (female). https://www.kaggle.com/datasets/neurought/94-vietnamese-characters-eeg-dataset-female, 2023. Accessed: 2025-09-25.

Thi Duyen Ngo, Hai Dang Kieu, Minh Hoa Nguyen, The Hoang-Anh Nguyen, Van Mao Can, Ba Hung Nguyen, and Thanh Ha Le. An eeg & eye-tracking dataset of als patients & healthy people during eye-tracking-based spelling system usage. *Scientific Data*, 11(1):664, 2024. ISSN 2052-4463. doi: 10.1038/s41597-024-03501-y. URL https://doi.org/10.1038/s41597-024-03501-y.

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

Arman Savran[1], Koray Ciftci[1], Guillame Chanel, Javier Cruz Mota, Luong Hong Viet, Bülent Sankur[1], Lale Akarun[1], Alice Caplier, and Michele Rombaut. Emotiondetection in the loop from brain signals and facial images. *Proceedings of the eNTERFACE 2006 Workshop*, 2006.

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*, aug 2017. ISSN 1097-0193. doi: 10.1002/hbm.23730. URL http://dx.doi.org/10.1002/hbm.23730.

Khalid F Shad, Yadollah Aghazadeh, Sajjad Ahmad, and Bernard Kress. Peripheral markers of alzheimer's disease: Surveillance of white blood cells. *Synapse*, 67(9):541–543, 2013. doi: 10.1002/syn.21667.

Chenxi Sun, Jin Jing, Niels Turley, Callison Alcott, Wan-Yee Kang, Andrew J. Cole, Daniel M. Goldenholz, Alice Lam, Edilberto Amorim, Catherine Chu, Sydney Cash, Valdery Moura Junior, Aditya Gupta, Manohar Ghanta, Bruce Nearing, Fábio A. Nascimento, Aaron Struck, Jennifer Kim, Shadi Sartipi, Alexandra-Maria Tauton, Marta Fernandes, Haoqi Sun, Grace Bayas, Kaileigh Gallagher, Joost B. Wagenaar, Nishant Sinha, Christopher Lee-Messer, Christine Tsien Silvers, Bharath Gunapati, Jonathan Rosand, Jurriaan Peters, Tobias Loddenkemper, Jong Woo Lee, Sahar Zafar, and M. Brandon Westover. Harvard electroencephalography database: A comprehensive clinical electroencephalographic resource from four boston hospitals. *Epilepsia*, 2025.

Mastaneh Torkamani-Azar, Sumeyra Demir Kanik, Serap Aydin, and Mujdat Cetin. Prediction of reaction time and vigilance variability from spatio-spectral features of resting-state eeg in a long sustained attention task. *IEEE journal of biomedical and health informatics*, 24(9):2550–2558, 2020.

Logan Trujillo. Raw EEG Data, 2020. URL https://doi.org/10.18738/T8/SS2NHB. Dataset.

Logan T Trujillo, Candice T Stanfield, and Ruben D Vela. The effect of electroencephalogram (eeg) reference choice on information-theoretic measures of the complexity and integration of eeg signals. *Frontiers in neuroscience*, 11:425, 2017.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=lvS2b8CjG5.

Jiquan Wang, Sha Zhao, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Generalizable sleep staging via multi-level domain alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 265–273, 2024b.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBramod: A criss-cross brain foundation model for EEG decoding. In *The International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NPNUHgHF2w.

Yijun Wang, Xiaogang Chen, Xiaorong Gao, and Shangkai Gao. A benchmark dataset for ssvep-based brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1746–1752, 2016.

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

Jin Yin, Aiping Liu, Chang Li, Ruobing Qian, and Xun Chen. A gan guided parallel cnn and transformer network for eeg denoising. *IEEE Journal of Biomedical and Health Informatics*, 29 (6):3930–3941, 2025. doi: 10.1109/JBHI.2023.3277596.

Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The national sleep research resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, 25(10):1351–1358, 2018.

Pengfei Zhang, Qian Wang, Shu-Dong Chen, Qihao-Hua Guo, Xia-Pheng Cao, Lan Tan, and Jin-Tai Yu. Peripheral immune cells and cerebrospinal fluid biomarkers of alzheimer's disease pathology in cognitively intact older adults: The cable study. *Journal of Alzheimer's Disease*, 87(3):721–730, 2022. doi: 10.3233/JAD-220236.

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.

A APPENDIX

A.1 DOWNSTREAM DATASET DESCRIPTION

- **BCIC-2A: Motor Imagery Task** The dataset contains EEG recordings from 10 participants performing four motor-imagery tasks: left hand (Class 1), right hand (Class 2), feet (Class 3), and tongue (Class 4). Each participant completed two sessions on separate days, with 288 trials per session.
- **A&MISP: Motor imagery tasks in Vietnamese** his dataset comprises 220 EEG recording sessions (2,640 trials) from 30 Vietnamese stroke patients (aged 43–78), recorded at 128 Hz using 22 motor-cortex channels (Emotiv EPOC Flex) under clinical conditions. Trials were validated via ERD% $\geq$ 30% to ensure neural engagement, and rich clinical metadata is included (NIHSS, mRS, Oxford muscle strength, lesion location).
- **ALS: Motor imagery tasks** This dataset contains raw 32-channel EEG recordings sampled at 256 Hz from six ALS patients each contributing up to ten sessions over three to five months and 170 healthy controls (one session each)
- **SleepEDF: Sleep Stage Detection Task** This dataset better probes model generalization: unlike trial- or event-based BCI data, sleep staging is continuous and requires long-duration stage labels. SleepEDF comprises 197 overnight recordings (78 healthy participants) with EEG, EOG, chin EMG, and event annotations.
- **TUEV: Event type classification** This dataset is subset of TUEG that contains annotations of EEG segments as one of six classes: (1) spike and sharp wave (SPSW), (2) generalized periodic epileptiform discharges (GPED), (3) periodic lateralized epileptiform discharges (PLED), (4) eye movement (EYEM), (5) artifact (ARTF) and. The EEG signals contain 23 channel sat 256Hz and are segmented into 112,4915-second samples.
- **TUAB: Abnormal detection** This dataset is a corpus of EEGs which are 23-channel and sampled at 256 Hz. All data have been annotated as normal or abnormal. There are total 409,455 10-second samples that we use for binary classification to predict normal/abnormal.
- **N-FM: Characters detection** This dataset consists of EEG recordings collected from 282 volunteers, aged 15 to 70, for brainwave signal classification of 94 Vietnamese characters. The data was collected using the NeuroSky Mindwave Mobile 2 headset in a controlled, noise-minimized environment to ensure high-quality signals.
- **PEARL: Alzheimer's risk prediction**: This dataset comprises data from 192 self-reported healthy middle-aged individuals (50–63 years), with an equal gender distribution. While 79 subjects are publicly accessible, the dataset includes both EEG and fMRI modalities, along with supplementary information such as blood test results, demographic profiles, and health status.

Table 6: Summary of recent state-of-the-art architectures for EEG Foundation Models and their original corresponding pretraining data.

Architectures	**Pretraning Datasets**
CBraMOD	TUEG
EEGPT	PhysioNet MI, HGD, TSU, SEED, M3CV
LaBraM	a subset of TUEG, BCIC IV-1, EmoBrain, Grasp and Lift, Inria BCIC, Resting State, SPIS Resting, SEED, Siena Scalp, Target vs Non-Target, Raw EEG Data, Private Data
BIOT	SHHS, a tiny subset from HEEDB
BENDR	TUEG

Table 7: Summary of T-EEG and its BCI Tasks.

BCI Task	Dataset	Rate	# Channels	Duration	# Labels
Motor Imagery	BCIC-2a	250 Hz	22	4s	4
	A&MISP	128 Hz	22	4s	4
	ALS	128 Hz	19	4s	4
Sleep Staging	SleepEDF	100 Hz	2	30s	5
Seizure / Event Detection	TUEV	250 Hz	16	10s	4
Abnormal EEG Detection	TUAB	250 Hz	16	10s	2
Characters Detection	N-FM	512 Hz	1	1s	94
Alzheimer's risk prediction	PEARL	1000 Hz	19	30s	2

Table 8: Hyperparameters for **M-EEG** fine-tuning.

Hyperparameters	Settings
Epochs	50
Batch size	64
Dropout	0.1
Optimizer	AdamW
Learning rate	1e-4
Adam β	(0.9, 0.999)
Adam ϵ	1e-8
Weight decay	5e-2
Scheduler	CosineAnnealingLR
Cosine cycle epochs	50
Minimal learning rate	1e-6
Clipping gradient norm	1

A.2 Fine-tuning on downstream tasks

We load the pre-trained weights of **M-EEG** and replace the reconstruction head with a task-specific head which is composed of multi-layer perceptrons. Here the learned EEG representations are flattened and fed into the task-specific head for downstream tasks. Then we fine-tune **M-EEG** in downstream datasets. **We employ binary cross-entropy (BCE) loss for binary classification, cross-entropy loss for multi-class classification.** More hyperparameters for **M-EEG** fine-tuning on downstream datasets are shown in Table 8.

A.3 Details on Multimodal Fusion Finetuning

Motivation. We draw motivation from medical studies indicating that cognitive impairments, such as Alzheimer's disease, are often accompanied by measurable alterations in peripheral blood counts, reflecting changes in both the numbers and proportions of circulating cells (Shad et al., 2013; Zhang et al., 2022; Dzianok & Kublik, 2024). Importantly, blood-based biomarkers provide a low-cost and minimally invasive means of capturing such physiological signals. Inspired by this, we propose a multimodal pipeline that integrates blood test results with EEG data to facilitate earlier detection of cognitive decline and support timely clinical intervention.

Multimodal fusion finetuning. Formally, let $\boldsymbol{r} \in \mathbb{R}^m$ denote the normalized vector of blood-based biomarkers. We apply a lightweight projection network $\mathrm{MLP}(\cdot)$ that maps $\boldsymbol{r}$ into the EEG token embedding space:

$$\boldsymbol{q} = \mathrm{MLP}(\boldsymbol{r}) \in \mathbb{R}^d. \tag{1}$$

Given EEG embedded tokens $\boldsymbol{Z} = \mathcal{E}_\theta(\boldsymbol{X}) \in \mathbb{R}^{L\times d}$, we implement late fusion by treating $\boldsymbol{q}$ as a query attending to the EEG tokens:

$$\alpha = \mathrm{softmax}\left(\frac{(\boldsymbol{q}W_Q)(\boldsymbol{Z}W_K)^\top}{\sqrt{d_k}}\right), \qquad \boldsymbol{h} = \alpha(\boldsymbol{Z}W_V)W_O \in \mathbb{R}^d. \tag{2}$$

The resulting cross-modal representation $\boldsymbol{h}$ serves as input to a prediction head for downstream tasks. At a high level, we adopt cross-attention since it enables *adaptive alignment* between biomarker information and EEG dynamics: the biomarker query can selectively attend to the most informative EEG patterns rather than relying on a static combination. This flexibility is particularly important when the contribution of blood-based signals varies across patients or conditions.

A.4 More results on Alzheimer's risk prediction on the PEARL dataset

In this section, we report additional results on Alzheimer's risk prediction using the PEARL dataset. Specifically, we investigate the contribution of blood biomarkers when combined with EEG representations extracted from two foundation models (**CBraMod** and **EEGPT**). The goal is to assess (i) whether multimodal fusion with blood improves over EEG-only baselines, and (ii) how EEG compares to blood-only models in terms of predictive power.

In addition to evaluating the original checkpoints of **EEGPT** and **CBraMod**, we also pretrained both foundation models on our dataset and repeated the same experiments. This allows us to assess whether the observed multimodal gains are consistent across both the original and domain-adapted versions of the foundation models.

Table 9 compares EEG-only models with multimodal EEG+Blood models (Concat and Attention fusion). The addition of blood consistently improves performance across both CBraMod and EEGPT. Among fusion strategies, attention achieves the strongest gains, suggesting that selective modality weighting is more effective than simple concatenation.

Table 10 presents blood-only baselines trained with an MLP, compared with EEG-only and EEG+Blood models. Blood-only consistently outperforms EEG-only, highlighting the stronger predictive value of blood biomarkers. However, combining EEG with blood further enhances performance, indicating that blood biomarkers, when integrated into EEG signals, provide essential complementary information for Alzheimer's risk prediction.

As shown in Tables 11 and 12, similar trends are observed when using our pretrained checkpoints. Attention-based fusion remains the best-performing strategy, and EEG-only models consistently outperform blood-only baselines. These results confirm that the benefit of incorporating blood biomarkers generalizes across both original and domain-adapted versions of EEG foundation models.

A.5 M-EEG's statistics

To further characterize our cohort, we summarize demographic distributions and recording statistics. The following subsection highlights patient demographics (age, gender), overall recording length, and site-specific acquisition configurations, providing a compact overview of the dataset composition.

- **Patient demographics**: We first summarize the demographics of patients included in M-EEG. Patient age was estimated from year of birth at the time of recording, yielding a median age of 46 years. Gender distribution comprised 61.63% female, 22.30% male, and 16.07% unspecified. In total, M-EEG includes 6081 unique patients, representing a broad and heterogeneous population across both hospitals. We included patients' demographics in Table 13 and Figure 3
- **Recording statistic**: We next characterize the EEG recordings themselves. Across both hospitals, M-EEG contains a total of 1170 hours of EEG, with a mean duration of 11.6 minutes per session. Temporal coverage spans 2019–2025, with the number of recordings generally increasing over time. The largest volumes were collected in 2024 (2258 subjects) and 2025 (2896 subjects), which contributed comparable numbers of sessions and together account for the majority of the dataset. Figure 4 summarizes these statistics, showing both the distribution of recording durations and the yearly distribution of subjects.
- **Site-specific patient demographics**: We further examined the distribution of subjects across acquisition sites. Hospital B contributed the majority of patients (5,134 subjects, 500 Hz, 44 channels), whereas Hospital A contributed 947 subjects recorded with a 200 Hz, 22-channel setup. This site-specific imbalance reflects differing patient volumes and hardware

Table 9: Alzheimer's risk prediction on the PEARL dataset. We compare unimodal EEG (baseline) with multimodal EEG plus blood-based biomarkers (Concat. and Attention). Metrics are balanced accuracy, PR-AUC and ROC-AUC. Relative improvements (%) over EEG-only are shown in the **Gain** columns, with teal denoting improvements and magenta for drops.

Task	Architecture	Metric	EEG-Only		EEG + BBB (Concat.)		EEG + BBB (Attention)	
			Perf.	Gain	Perf.	Gain	Perf.	Gain
PEARL–MSIT	CBraMOD	Acc	0.4816		0.5263	+9.28%	0.6373	+32.33%
		pr_auc	0.5597		0.6013	+7.43%	0.6863	+22.62%
		roc_auc	0.5818		0.5979	+2.77%	0.7235	+24.36%
	EEGPT	Acc	0.4550		0.4968	+9.19%	0.5560	+22.20%
		pr_auc	0.4840		0.5767	+19.15%	0.6056	+25.12%
		roc_auc	0.4035		0.4915	+21.81%	0.5023	+24.49%
PEARL–RMT	CBraMOD	Acc	0.5280		0.4982	-5.64%	0.6288	+19.09%
		pr_auc	0.4661		0.5656	+21.35%	0.6043	+29.65%
		roc_auc	0.4985		0.5946	+19.28%	0.6554	+31.47%
	EEGPT	Acc	0.4312		0.4310	-0.05%	0.5226	+21.20%
		pr_auc	0.3982		0.4462	+12.05%	0.5745	+44.27%
		roc_auc	0.3805		0.4072	+7.02%	0.5285	+38.90%
PEARL–RST	CBraMOD	Acc	0.4504		0.5606	+24.47%	0.5793	+28.62%
		pr_auc	0.3927		0.6600	+68.07%	0.6666	+69.75%
		roc_auc	0.3997		0.6098	+52.56%	0.6416	+60.52%
	EEGPT	Acc	0.3952		0.4096	+3.64%	0.5722	+44.79%
		pr_auc	0.3556		0.3910	+9.96%	0.4856	+36.56%
		roc_auc	0.3281		0.3742	+14.05%	0.4310	+31.36%

Table 10: Comparison of BBB-only, EEG-only (original baseline), and EEG plus BBB (Attention) models on the PEARL dataset. Metrics are balanced accuracy, PR-AUC and ROC-AUC. Relative improvements (%) over EEG-only are shown in the **Gain** columns, with teal denoting improvements and magenta for drops.

Task	Architecture	Metric	EEG-only		BBB-only		EEG + BBB (Attention)	
			Perf.	Gain	Perf.	Gain	Perf.	Gain
PEARL–MSIT	CBraMOD	Acc	0.4816	-11.3%	0.543		0.6373	+17.4%
		pr_auc	0.5597	+6.4%	0.526		0.6863	+30.7%
		roc_auc	0.5818	-3.5%	0.603		0.7235	+19.9%
	EEGPT	Acc	0.4550	-16.2%	0.543		0.5560	+2.4%
		pr_auc	0.4840	-7.9%	0.526		0.6056	+15.13%
		roc_auc	0.4035	-33.1%	0.603		0.5023	-16.7%
PEARL–RMT	CBraMOD	Acc	0.5280	-2.7%	0.543		0.6288	+15.8%
		pr_auc	0.4661	+11.4%	0.526		0.6043	+14.9%
		roc_auc	0.4766	-20.9%	0.603		0.5402	-10.4%
	EEGPT	Acc	0.4615	-15.0%	0.543		0.5976	+10.1%
		pr_auc	0.4203	-20.1%	0.526		0.5458	3.7%
		roc_auc	0.4050	-32.8%	0.603		0.5407	-10.33%
PEARL–RST	CBraMOD	Acc	0.5283	-2.7%	0.543		0.5877	+8.2%
		pr_auc	0.5001	-4.9%	0.526		0.5620	+6.8%
		roc_auc	0.4766	-20.9%	0.603		0.5402	-10.4%
	EEGPT	Acc	0.4615	-15%	0.543		0.5976	+10.1%
		pr_auc	0.4203	-20.1%	0.526		0.5458	+3.7%
		roc_auc	0.4050	-32.8%	0.603		0.5407	-10.3%

configurations but ensures representation from both institutions. The overall breakdown is summarized in Table 14.

Table 11: Alzheimer's risk prediction on the PEARL dataset. We compare unimodal EEG (pretrained using P-EEG) with multimodal EEG plus blood-based biomarkers (Concat. and Attention). Metrics are balanced accuracy, PR-AUC and ROC-AUC. Relative improvements (%) over EEG-only are shown in the **Gain** columns, with teal denoting improvements and magenta for drops.

Task	Architecture	Metric	EEG-Only		EEG + BBB (Concat.)		EEG + BBB (Attention)	
			Perf.	Gain	Perf.	Gain	Perf.	Gain
PEARL–MSIT	CBraMOD	Acc	0.5283		0.5515	+4.39%	0.6743	+27.64%
		pr_auc	0.5523		0.5609	+1.56%	0.7588	+37.39%
		roc_auc	0.5877		0.6148	+4.61%	0.7779	+32.36%
	EEGPT	Acc	0.4615		0.5505	+19.29%	0.5660	+22.64%
		pr_auc	0.4285		0.5319	+24.13%	0.5789	+35.10%
		roc_auc	0.4063		0.4974	+22.42%	0.5191	+27.76%
PEARL–RMT	CBraMOD	Acc	0.5296		0.5492	+3.70%	0.6213	+17.32%
		pr_auc	0.4692		0.6213	+32.42%	0.6773	+44.35%
		roc_auc	0.5040		0.6274	+24.48%	0.7156	+41.98%
	EEGPT	Acc	0.4746		0.4861	+2.42%	0.5627	+18.56%
		pr_auc	0.4132		0.5375	+30.08%	0.6109	+47.85%
		roc_auc	0.4222		0.4855	+14.99%	0.5651	+33.85%
PEARL–RST	CBraMOD	Acc	0.4375		0.6472	+47.93%	0.6960	+59.09%
		pr_auc	0.4445		0.7095	+59.62%	0.7772	+74.85%
		roc_auc	0.4580		0.6839	+49.32%	0.7783	+69.93%
	EEGPT	Acc	0.4366		0.4776	+9.39%	0.5753	+31.77%
		pr_auc	0.3925		0.4127	+5.15%	0.5985	+52.48%
		roc_auc	0.3949		0.4165	+5.47%	0.5483	+38.85%

Table 12: Comparison of BBB-only, EEG-only (pretrained using P-EEG), and EEG plus BBB (Attention) models on the PEARL dataset. Metrics are balanced accuracy, PR-AUC and ROC-AUC. Relative improvements (%) over EEG-only are shown in the **Gain** columns, with teal denoting improvements and magenta for drops.

Task	Architecture	Metric	EEG-only		BBB-only		EEG + BBB (Attention)	
			Perf.	Gain	Perf.	Gain	Perf.	Gain
PEARL–MSIT	CBraMOD	Acc	0.5283	-2.07%	0.543		0.6743	+24.18%
		pr_auc	0.5523	+5.00%	0.526		0.7588	+44.26%
		roc_auc	0.5877	-2.54%	0.603		0.7779	+29.01%
	EEGPT	Acc	0.4615	-15.01%	0.543		0.5660	+4.24%
		pr_auc	0.4285	-18.54%	0.526		0.5789	+10.06%
		roc_auc	0.4063	-32.62%	0.603		0.5191	-13.91%
PEARL–RMT	CBraMOD	Acc	0.5296	-2.47%	0.543		0.6213	+14.42%
		pr_auc	0.4692	-10.80%	0.526		0.6773	+28.76%
		roc_auc	0.5040	-16.42%	0.603		0.7156	+18.67%
	EEGPT	Acc	0.4745	-12.62%	0.543		0.5627	+3.63%
		pr_auc	0.4132	-21.45%	0.526		0.6109	+16.14%
		roc_auc	0.4222	-29.98%	0.603		0.5651	-6.29%
PEARL–RST	CBraMOD	Acc	0.4375	-19.43%	0.543		0.6960	+28.18%
		pr_auc	0.4445	-15.49%	0.526		0.7772	+47.76%
		roc_auc	0.4580	-24.05%	0.603		0.7783	+29.07%
	EEGPT	Acc	0.4366	-19.59%	0.543		0.5753	+5.95%
		pr_auc	0.3925	-25.38%	0.526		0.5985	+13.78%
		roc_auc	0.3949	-34.51%	0.603		0.5483	-9.07%

A.6 Description of the BIDS Structure of the Database

In this study, we organized our database following the **Brain Imaging Data Structure (BIDS) specification**, version 1.8.0. BIDS is a community-driven standard that provides a uniform way to

Table 13: Summary of Patient Statistics

Category	Statistic	Value
Age	Mean ± STD	45.88 ± 18.08
	Minimum - Maximum	1 - 104
Gender	Female	61.63%
	Male	22.30%
	N/A	16.07%

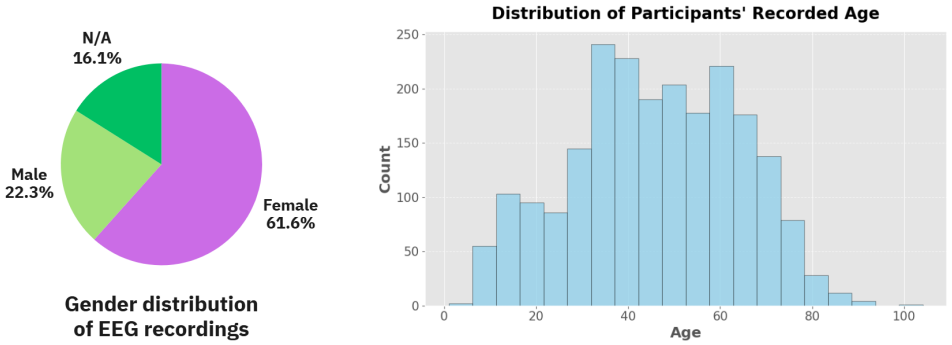

Figure 3: Patients' demographics

Table 14: Site-specific patient demographics between 2 hospitals

Site	#Subjects	Channels	Sampling Rate
Hospital A	947	22	200
Hospital B	5,134	44	500

arrange neuroimaging and physiological datasets, ensuring consistency, interoperability, and reproducibility across studies.

By adopting BIDS v1.8.0, we gain several advantages:

- **Standardization**: Data from different acquisition sites and modalities (e.g., EEG signals, clinical laboratory results) are represented in a consistent format, reducing ambiguity in interpretation.

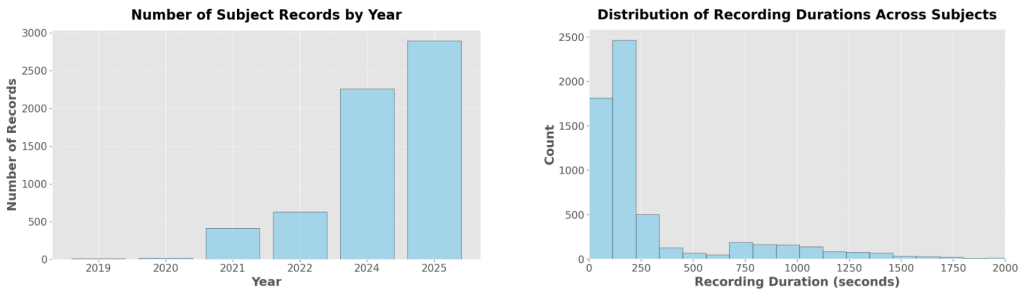

Figure 4: Recording duration and yearly distribution of subjects in M-EEG

- **Compatibility**: The dataset can be directly integrated with existing BIDS-aware software tools for preprocessing, quality control, and statistical analysis.
- **Reproducibility**: Researchers can reuse the dataset with minimal manual curation, which facilitates replication studies and meta-analyses.
- **Extensibility**: Beyond EEG recordings, our design includes phenotype-level information (e.g., laboratory test results), enabling multimodal analysis that links neurophysiological data with clinical variables.

At the top level, the dataset is structured according to the BIDS hierarchy, which includes:

- `dataset_description.json`: Contains metadata describing the dataset, its authorship, and BIDS compliance.
- `participants.tsv` and `participants.json`: Contain participant-level demographic and group information.
- `phenotype/`: Contains clinical laboratory test results in `results.tsv` and related metadata in `results.json`.
- `sub-xxxx/`: Contain subject-specific data, including an `eeg/` subfolder with EEG recordings, associated metadata, channel information, and a `sub-xxxx_scans.tsv` file documenting recording timestamps.

This organization ensures that the dataset is self-describing and can be recognized by BIDS-compatible tools without requiring additional documentation. For reproducibility and illustration purposes, a small selection of de-identified EEG recordings along with corresponding metadata will be provided as supplementary material once the manuscript is accepted.

A.7 Details on Data acquisition

Raw data acquisition. M-EEG comprises $1,170$ hours of routine clinical EEG collected from $6,081$ patients across two hospitals over multiple years. Raw signals were acquired under standard clinical protocols but with distinct hardware setups: Hospital A employed a 22-channel 10–20 montage sampled at 200 Hz, while Hospital B employed a 44-channel configuration sampled at 500 Hz. The cohort spans a broad range of neurological conditions, including epilepsy, encephalopathy, sleep disorders, and neurodegenerative diseases, reflecting the heterogeneity of real-world practice. All recordings were fully de-identified before release, with patient identifiers removed and institution-specific metadata anonymized, thereby preserving clinical fidelity while ensuring compliance with privacy and ethical standards.

Cross-modality synchronization. In addition to EEG, VEEG provides minimally invasive physiological and textual context. All routine blood-based biomarkers and de-identified clinical notes are centralized in a dedicated `phenotype/` directory. Each patient is linked to two files: `results.tsv`, containing tabular laboratory values, and `results.json`, containing free-text diagnostic notes and impressions. This design ensures consistent alignment between EEG and auxiliary modalities while remaining lightweight and machine-readable.

Standardization. EEG recordings follow BIDS conventions with sidecar JSON metadata, and auxiliary modalities are synchronized by subject identifiers in the `phenotype/` directory. This unified schema enables seamless integration of EEG, biomarkers, and clinical notes, providing a scalable foundation for multimodal representation learning.

A.8 Limitations

The robustness gains from incorporating regional data are marginal but consistent, indicating steady benefits even at limited scale. These results provide encouraging evidence that regional coverage can enhance generalization, though M-EEG remains smaller than corpora such as TUEG or HEEDB. As we expand data collection to achieve greater balance, future work will more fully explore the role of regional diversity in building robust EEG foundation models.