Standing on the Shoulders of Giants: Rethinking EEG Foundation Model Pretraining via Multi-Teacher Distillation

Chenqi Li, Yu Liu, Shuo Zhang, Timothy Denison, Tingting Zhu
Department of Engineering Science
University of Oxford
{chenqi.li,yu.liu,shuo.zhang,timothy.denison,tingting.zhu}@eng.ox.ac.uk

Abstract

Pretraining for electroencephalogram (EEG) foundation models has predominantly relied on self-supervised masked reconstruction, a paradigm largely adapted from and inspired by the success of vision and language foundation models. However, unlike images and text, EEG datasets are notoriously expensive to collect and characterized by low signal-to-noise ratio. These challenges introduce difficulties in scaling the EEG foundation models and capturing the underlying neural semantics through reconstruction. In this work, we ask the question: can we stand on the shoulders of well-established foundation models from well-represented modalities to bootstrap the pretraining of EEG foundation models? We first demonstrate that mainstream foundation models, such as those from vision and time series, transfer surprisingly well to the EEG domain. Motivated by this observation, we propose the multi-teacher distillation pretraining (MTDP) framework for pretraining EEG foundation models via two-stage multi-teacher distillation. In the first stage, we introduce a learnable gating network to fuse representations from diverse teachers (e.g., DINOv3 and Chronos) via a masked latent denoising objective. In the second stage, we distill the fused representation into an EEG foundation model. Extensive evaluations across 2 backbone architectures, 9 downstream tasks, and 12 datasets demonstrate that multi-teacher distillation pretraining outperforms self-supervised pretraining on the majority of datasets, while remaining competitive even when the pretraining data is reduced to 25%.

1 Introduction

Electroencephalogram (EEG) records the spontaneous electrical activity of the brain through electrodes placed on the scalp [34]. It has become a fundamental tool for assisting clinical diagnosis, advancing neuroscience and cognitive science research, developing brain-computer interface (BCI), and is considered the gold standard for sleep staging [7] and epilepsy diagnosis [13]. Although EEG recordings have traditionally been analyzed by highly trained specialists, deep learning has not only succeeded in automating such analysis, but also in unlocking new applications for EEG where data were not traditionally read by humans. Deep learning models, such as convolutional neural networks and recurrent neural networks, have demonstrated exceptional performance across a variety of tasks: seizure detection [3], sleep staging [1], emotion recognition [23], and motor imagery classification [4]. However, early EEG-based deep learning models are often tailored to specific tasks and datasets, with limited ability to generalize beyond strictly defined experimental setups.

Following the remarkable success of foundation models (FMs) in computer vision and natural language processing, there is growing interest in developing similar FMs for EEG. This presents a paradigm shift away from training task- and dataset- specific models, to learning universal neural

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

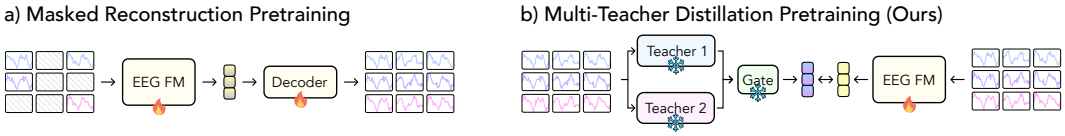

Figure 1: Comparison of EEG foundation model pretraining. a) The conventional self-supervised pretraining where EEG foundation model reconstructs missing patches in the temporal, frequency, or latent-domain. b) The proposed framework pretrains EEG foundation models by distilling cross-domain priors from well-established foundation models across diverse, well-represented modalities.

representations for models that can be adapted to a wide range of downstream tasks [43, 47]. In particular, the pretraining of existing EEG FMs predominantly relies on masked reconstruction, where the model learns to predict missing patches of the input signal. Subsequently, these pretrained models are fine-tuned with a classification head, demonstrating state-of-the-art performance across diverse benchmarks [18, 39, 30].

Unlike natural language or images, EEG data are significantly more challenging to collect and share due to strict privacy constraints. Consequently, existing EEG datasets remain orders of magnitude smaller than the internet-scale corpora used to train vision [33] and language [14] FMs, thereby challenging the scaling of EEG FMs. Furthermore, EEG signals are characterized by an exceptionally low signal-to-noise ratio, by which reconstruction-based objectives inherently prioritize the modelling of artifacts and noise, effectively directing the model away from meaningful neural dynamics. Therefore, the scarcity of data and the sensitivity to noise call for more data-efficient representation learning strategies that leverage external priors, rather than relying solely on masked reconstruction. This naturally raises the question: can we stand on the shoulders of well-established FMs from well-represented modalities to bootstrap the pretraining of an EEG FM?

In this work, we start by showing that mainstream FMs are surprisingly good at extracting EEG representations through a series of linear probing experiments. We then propose the multi-teacher distillation pretraining (MTDP) framework to pretrain EEG FM via two-stage multi-teacher distillation. In the first stage, we introduce a learnable gating network to weigh the importance of each teacher model and fuse teacher representations in an unsupervised manner. The second stage pretrains EEG FM by learning from the fused teacher representations through distillation. Figure 1 provides a comparison of masked reconstruction pretraining used in existing EEG FMs and the proposed MTDP framework. Our contributions can be summarized as follows:

- We show that DINOv3, a vision FM, trained on a different modality, transfers surprisingly well to EEG tasks, outperforming specialized EEG FMs under the linear probing setting.
- We explore multi-teacher knowledge distillation from vision, time series, and audio FMs as a data-efficient alternative to large-scale self-supervised pretraining for EEG.
- To leverage the complementary representations from teacher FMs of different modalities, we present a learnable gating network to weigh each teacher via a masked latent denoising objective and fuse teacher representations in an unsupervised manner.
- Extensive evaluations across 2 backbone architectures, 9 downstream tasks, and 12 datasets demonstrate that multi-teacher distillation pretraining outperforms self-supervised pretraining on the majority of datasets, while maintaining competitive performance with only 25% of the pretraining data.

2 Related Works

2.1 EEG Foundation Model

Deep learning models for automated EEG analysis have undergone a paradigm shift from task-specific supervised learning to FMs that are capable of achieving state-of-the-art results on a variety of downstream tasks. Early efforts demonstrated the viability of learning universal biosignal representations by aggregating diverse datasets for pretraining [43], followed by subsequent work that scaled the pretraining to larger datasets [47, 45, 18]. The pretraining strategy for EEG FMs focuses on self-supervised learning, specifically masked reconstruction, aiming to reconstruct masked segments

in the temporal domain [43, 8, 39], frequency domain [18], or latent representation spaces [9, 18]. Concurrently, research has focused on architectural designs to handle the heterogeneous nature of EEG electrode configurations by introducing topology-agnostic mechanisms to generalize across arbitrary electrode montages [30, 12], to improve scaling by exploring efficient state-models [10, 36], or to incorporate joint learning with additional modalities such as iEEG [46], MEG [41], and other biosignals [28]. However, despite these advances in scaling and architecture, alternative pretraining objectives to masked reconstruction remain underexplored [26].

2.2 Multi-Teacher Distillation

Although single-teacher distillation [16] is effective, multi-teacher distillation seeks to synthesize diverse perspectives from different teacher models, through collaborative frameworks [31] or dynamic weighing of teacher contributions [27]. Recent literature has increasingly explored distillation from Mixture-of-Experts (MoE) architectures [42, 2, 20, 22], which can be viewed as an extension of multi-teacher frameworks, where a gating mechanism dynamically selects the most relevant teachers. A notable application is AM-RADIO [32], which employs multi-teacher distillation to amalgamate the unique strengths of each teacher into a unified model. However, existing multi-teacher distillation methods combine teacher output using one of two methods: (1) learned weighted average (2) uniform average. Methods that try to learn to weigh each teacher require labels, which are unsuitable for FM pretraining due to the lack of labels. Methods that average the teacher uniformly assume all teachers are equally valuable, which is rarely the case. To address this, we propose a learnable gate to weigh teachers in an unsupervised manner, outperforming uniform averaging.

3 Preliminary Experiment

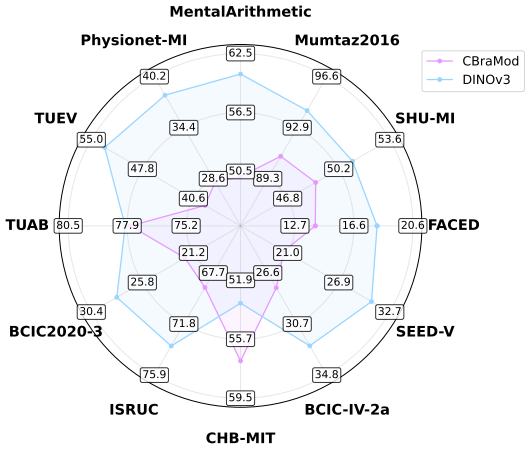

Figure 2: Linear probing performance of CBraMod and DINOv3 on EEG downstream tasks. Balanced Accuracy (%).

To evaluate the quality of representation from pretrained EEG FMs versus mainstream FMs, we perform a linear probing analysis comparing the state-of-the-art EEG FM CBraMod [39] against the vision FM DINOv3 [33]. We extract representations from EEG signals using both models as frozen feature extractors. The cross-modal adaptation protocol used to project EEG signals into a format compatible with DINOv3 is detailed in Section 5.1.2. Finally, we evaluate the linear separability of these representations by training a logistic regression classifier as a proxy for the quality of the captured neural dynamics. By restricting the classifier to a linear architecture, we ensure that the performance reflects the intrinsic quality of the representations rather than the capacity of the downstream head to compensate for poor representations.

Figure 2 compares the performance of the linear probing in a variety of downstream EEG tasks. DINOv3 consistently outperforms CBraMod across all datasets except CHB-MIT, suggesting that, compared to EEG FMs, large-scale vision FMs generalize well to EEG tasks. These findings provide strong motivation to reconsider the traditional masked reconstruction pretraining for EEG FMs, and instead explore leveraging the knowledge from mainstream FMs.

4 Methodology

4.1 Problem Definition

Let $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{N}$ denote an unlabeled EEG dataset containing N samples. Each sample $\mathbf{x} \in \mathbb{R}^{C \times T}$ represents a multi-channel recording with C channels across T time steps. We define an EEG FM $f_\theta : \mathbb{R}^{C \times T} \to \mathbb{R}^d$, parameterized by θ, which serves as an encoder that maps the EEG sample $\mathbf{x}$ into a d-dimensional latent representation $\mathbf{h} \in \mathbb{R}^d$.

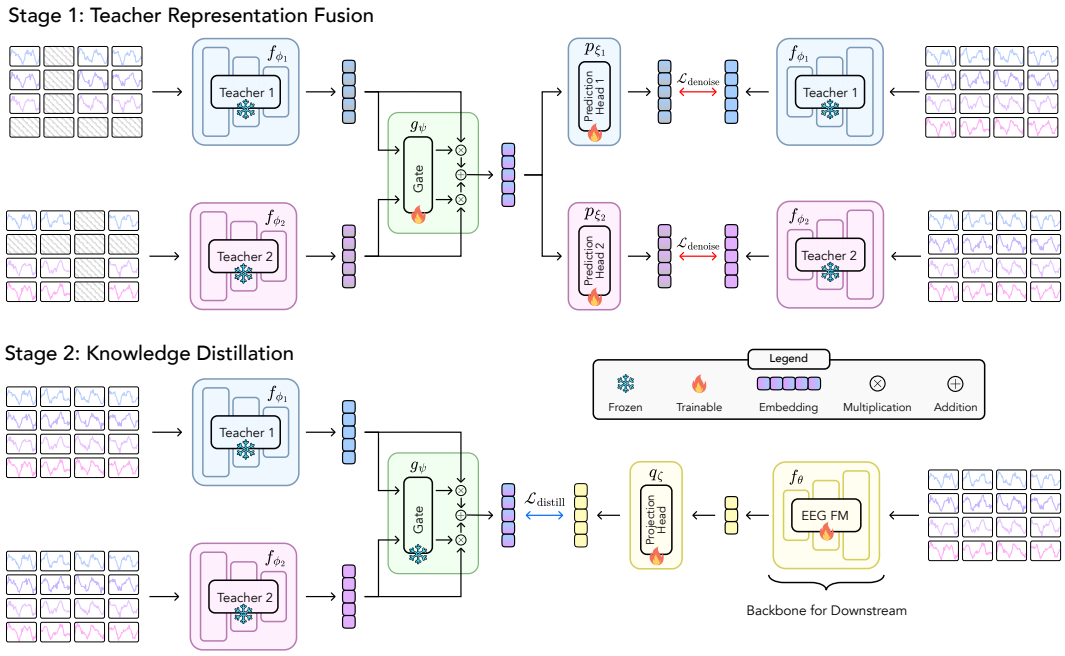

Figure 3: Overview of two-stage multi-teacher distillation pretraining (MTDP). Stage 1: teacher representation fusion. A learnable gating network is introduced to weigh and fuse representations from frozen teacher models. The gate is trained via a masked latent denoising objective. Stage 2: knowledge distillation. The fused teacher representation acts as the target to pretrain the student EEG foundation model. The distillation loss is minimized to align the student representations with the fused representations.

The objective of EEG FM pretraining is to optimize the parameters θ to maximize generalization performance across a set of N_{task} downstream tasks $\mathcal{T} = \{\mathcal{T}_t\}_{t=1}^{N_{\text{task}}}$. Each task $\mathcal{T}_t$ is associated with a dataset $\mathcal{D}_t$ partitioned into training $\mathcal{D}_t^{\text{train}}$, validation $\mathcal{D}_t^{\text{val}}$, and test $\mathcal{D}_t^{\text{test}}$ subsets. The training set consists of N_{train} EEG-label pairs, $\mathcal{D}_t^{\text{train}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_{\text{train}}}$, with the validation and test sets defined analogously. The pretraining optimization objective is formulated as follows:

$$\theta^* = \arg\max_{\theta} \mathbb{E}_{\mathcal{T}_t \in \mathcal{T}} \left[\mathcal{P}_t(h_{\eta_t} \circ f_{\theta_t}, \mathcal{D}_t^{\text{test}}) \right] \tag{1}$$

where $\mathcal{P}_t$ denotes the performance metric (e.g., accuracy or F1-score) for the task $\mathcal{T}_t$, and h_{η_t} is a task-specific head parameterized by η_t. The operator $\circ$ denotes function composition. The adapted parameters θ_t and η_t are derived from the pretrained initialization θ and random initialization η via an adaptation algorithm $\mathbb{A}$:

$$(\theta_t, \eta_t) = \mathbb{A}(\theta, \eta, \mathcal{D}_t^{\text{train}}, \mathcal{D}_t^{\text{val}}) \tag{2}$$

The algorithm $\mathbb{A}$ typically corresponds to linear probing, where $\theta_t = \theta$, or fine-tuning. Crucially, the pretraining phase lacks prior knowledge of $\mathcal{T}$, requiring the encoder f_θ to learn universal representations of neural dynamics agnostic to tasks.

4.2 Two-Stage Multi-Teacher Distillation Pretraining

The core idea of the proposed MTDP framework is to pretrain EEG FMs by leveraging the knowledge from mainstream FMs. We achieve this by injecting the knowledge of well-established FMs from well-represented modalities (teachers) into an EEG FM (student) via a specialized distillation process. The choice of EEG FM architecture is flexible, which can be an existing EEG FM, like CBraMod.

EEG signals are high-dimensional and non-stationary, characterized by a complex interplay of spatial, temporal, and spectral components that reflect the underlying neurophysiological activities. Although Section 3 shows that vision FMs have demonstrated promising potential to extract useful EEG representations, a single teacher often lacks the breadth to capture all of the aforementioned

dimensions simultaneously. To address this, our MTDP framework seeks to synthesize complementary insights from FMs of different modalities that serve to guide the EEG model toward a holistic understanding of the multifaceted nature of EEG more effectively than any single teacher could alone.

To leverage the complementary representations from different modalities, our MTDP framework has two stages:

- **Stage 1: Teacher Representation Fusion.** We introduce a learnable gating network to weigh the importance of each teacher model using a masked latent denoising objective. The goal is to synthesize a unified teacher representation in an unsupervised manner.
- **Stage 2: Knowledge Distillation.** The unified teacher representation serves as the distillation target for a randomly initialized EEG FM student. The goal is to transfer the synthesized knowledge from different teachers into a compact EEG FM.

An overview of the method is illustrated in Figure 3.

4.2.1 Stage 1: Teacher Representation Fusion

The main challenge in the fusion of teacher representation is to identify the relative importance of each teacher in an unsupervised manner. Although teachers can be weighted using task-specific objective functions, such functions are unavailable during unsupervised pretraining when downstream tasks are unknown. To address this issue, we design a masked latent denoising objective to automatically weigh each teacher. Intuitively, the objective encourages the fusion of masked teacher representations such that it can best predict each of the unmasked teacher representations.

Given an EEG sample $\mathbf{x}$ from the pretraining dataset $\mathcal{D}$, we construct a masked version $\tilde{\mathbf{x}}$ by applying a binary mask $\mathbf{m} \in \{0,1\}^{C\times T}$:

$$\tilde{\mathbf{x}} = \mathbf{x} \odot \mathbf{m}, \tag{3}$$

where $\odot$ denotes element-wise product. We apply two types of masking strategies:

- **Segment Masking**: A contiguous window between one and two seconds is masked across all channels to encourage the model to learn temporal dynamics.
- **Channel Dropout**: A random channel is dropped for all timesteps to capture spatial correlation and redundancy.

Let $\{f_{\phi_k}\}_{k=1}^{K}$ be a set of K pretrained teacher models, where each model is parametrized by ϕ_k. For each teacher k, we extract the representation from the unmasked and masked EEG samples:

$$\mathbf{h}_k = f_{\phi_k}(\mathbf{x}) \quad \text{and} \quad \tilde{\mathbf{h}}_k = f_{\phi_k}(\tilde{\mathbf{x}}) \tag{4}$$

Let g_ψ be the gating network parametrized by ψ. The gate takes as input the concatenated masked teacher representations and outputs a scalar weight for each teacher:

$$\mathbf{w} = [w_1, w_2, \ldots, w_K] = g_\psi([\tilde{\mathbf{h}}_1; \tilde{\mathbf{h}}_2; \ldots; \tilde{\mathbf{h}}_K]) \tag{5}$$

The fused teacher representation is computed as the weighted sum of the teacher representations:

$$\tilde{\mathbf{h}}_{\text{fused}} = \sum_{k=1}^{K} w_k \cdot \tilde{\mathbf{h}}_k \tag{6}$$

The fused representation $\tilde{\mathbf{h}}_{\text{fused}}$ amalgamates the representations from different teachers and will serve as the distillation target for stage 2, facilitating the transfer of knowledge from multiple teachers into an EEG FM student. The gating network is trained using the masked latent denoising objective:

$$\mathcal{L}_{\text{denoise}} = \sum_{k=1}^{K} \|p_{\xi_k}(\tilde{\mathbf{h}}_{\text{fused}}) - \mathbf{h}_k\|_2^2, \tag{7}$$

where $\{p_{\xi_k}\}_{k=1}^{K}$ is a set of prediction heads parametrized by ξ_k for each teacher.

4.2.2 Stage 2: Knowledge Distillation

With the pretrained gating network, the objective of stage 2 is to distill the fused teacher representation into the student model. For an EEG sample $\mathbf{x}$ from the pretraining dataset, we

compute the representation from each teacher $\mathbf{h}_k = f_{\phi_k}(\mathbf{x})$ and the weight for each teacher $\mathbf{w} = [w_1, \ldots, w_K] = g_\psi([\mathbf{h}_1; \mathbf{h}_2; \ldots; \mathbf{h}_K])$. The fused representation is computed as:

$$\mathbf{h}_{\text{fused}} = \sum_{k=1}^{K} w_k \cdot \mathbf{h}_k \tag{8}$$

To train the student model, we minimize the distillation loss as the cosine similarity between the student representation and the target fused teacher representation:

$$\mathcal{L}_{\text{distill}} = 1 - \frac{\langle q_\zeta \circ f_\theta(x), \mathbf{h}_{\text{fused}} \rangle}{||q_\zeta \circ f_\theta(x)||_2 \cdot ||\mathbf{h}_{\text{fused}}||_2}, \tag{9}$$

where q_ζ is a learnable projection head parameterized by ζ to match the dimensionality of the student representation with that of the fused teacher representation. $\langle \cdot, \cdot \rangle$ denotes the inner product operator. For downstream fine-tuning, teacher models $\{f_{\phi_k}\}_{k=1}^K$, gating network g_ψ, prediction heads $\{p_{\xi_k}\}_{k=1}^K$, and projection head q_ζ are discarded. A detailed description of the MTDP framework is provided in Algorithm 1.

5 Experiments

5.1 Experimental Setups

Our experiments are designed to understand the gain of the proposed multi-teacher distillation pretraining framework over the self-supervised masked reconstruction pretraining framework adopted in existing EEG FMs. Therefore, we reuse the setup from existing EEG FMs and replace the self-supervised pretraining with multi-teacher distillation pretraining.

5.1.1 Teacher and Student Models

Student Models. To demonstrate the generalizability of the proposed framework, we conduct experiments using student backbones based on Transformer [37] and Mamba [15] architectures, which represent the vast majority of existing EEG FMs. We selected CBraMod [39] and EEGMamba [40], as they represent the current state-of-the-art performance for their respective architectures according to independent third-party benchmarks [25]. The architecture and corresponding hyperparameters are detailed in Table A1-A2.

Teacher Models. We adopt DINOv3-ViT-B/16 [33] and Chronos-Bolt-Base [5] as teachers. DINOv3-ViT-B/16 is distilled from DINOv3-7B, which is pretrained on 1.7 billion images. It is based on the ViT-B/16 [11] architecture with 87M parameters, consisting of a 12-layer transformer encoder. Chronos-Bolt-Base is a zero-shot time series forecasting model pretrained on nearly 100 billion time series observations. Chronos-Bolt-Base is based on the T5-Efficient-Base [35] architecture with 205M parameters, consisting of a 12-layer transformer encoder and 12-layer transformer decoder. We extract representations from the encoder only, as the decoder is designed for generative forecasting.

5.1.2 Preprocessing EEG for Teachers

While numerous strategies are available for preprocessing EEG signals for DINOv3 and Chronos, we adopt the Occam Razor's principle, prioritizing minimal computational overhead. DINOv3 expects images, which are represented as tensors of shape $\mathbb{R}^{3\times H\times W}$, where H and W are the height and width. EEG signals are tensors of shape $\mathbb{R}^{C\times T}$, which can be considered as single-channel images with height C and width T. We preprocesss EEG via min-max normalization to 0-1. We apply min-max normalization to scale the EEG signal to the $[0, 1]$ range, followed by a replication of the data over three channels. We also experimented with the spectrogram and continuous wavelet transform to provide additional phase-frequency information, but they introduced significant computational overhead and did not result in significant differences in performance. Chronos expects univariate time series, which are tensors of shape $\mathbb{R}^L$, where L is the sequence length. The EEG channels are processed independently, and the mean representation across all channels is used.

5.1.3 Pretraining and Downstream Datasets

Pretraining Dataset and Preprocessing. Following CBraMod and EEGMamba, we adopt the Temple University Hospital EEG Corpus (TUEG) for pretraining [29], in order to ensure fair comparison

between the self-supervised masked reconstruction pretraining and the multi-teacher distillation pretraining. Section B.1 describes the preprocessing pipeline.

Downstream Datasets. To assess the performance and generalizability of the proposed method, we conduct evaluations on 12 publicly available downstream datasets. We follow the preprocessing pipeline and subject-independent dataset splits adopted in CBraMod [39] and EEGMamba [40]. Table 1 summarizes all the downstream datasets.

Table 1: Overview of downstream datasets

Task	Dataset	Rate	# Channels	Duration	# Samples	Label
Motor Imagery Classification	PhysioNet-MI	160 Hz	64	4s	9,837	4-class
	SHU-MI	250 Hz	32	4s	11,988	2-class
	BCIC-IV-2a	250 Hz	22	4s	5088	4-class
Emotion Recognition	FACED	250 Hz	32	10s	10,332	9-class
	SEED-V	1000 Hz	62	1s	115,001	5-class
Sleep Staging	ISRUC	200 Hz	6	30s	90,187	5-class
Seizure Detection	CHB-MIT	256 Hz	16	10s	326,993	2-class
Imagined Speech Classification	BCIC2020-3	256 Hz	64	3s	6,000	5-class
Mental Disorder Diagnosis	Mumtaz2016	256 Hz	19	5s	7,143	2-class
Mental Stress Detection	MentalArithmetic	500 Hz	20	5s	1,707	2-class
Event Type Classification	TUEV	250 Hz	16	5s	113,353	6-class
Abnormal Detection	TUAB	250 Hz	16	10s	409,455	2-class

5.1.4 Training and Evaluation Details

Pretraining Stage 1 Teacher Representation Fusion. The gating network consists of a 2-layer MLP with ReLU activation between the layers and softmax activation to normalize the weights. The prediction heads consist of a single fully connected layer. Mean Squared Error (MSE) loss is adopted as the masked latent denoising objective.

Pretraining Stage 2 Knowledge Distillation. The pretrained weights for the gating network from stage 1 are loaded and frozen during stage 2 knowledge distillation. The projection head consists of a single fully connected layer. Cosine similarity is adopted as the distillation objective to minimize the distance between the projected student representations and the fused teacher representation. The student weights are randomly initialized.

Training Setup. Across both stages, we adopted a batch size of 64 per GPU across four V100 GPUs, resulting in a global batch size of 256. To avoid redundant teacher inference during distillation, teacher representations were pre-computed and stored prior to the start of the distillation process. The gate is trained for one epoch in less than an hour. Distillation on a 25% subset of the TUEG dataset over 40 epochs is completed in 21 hours. The loss curves for both are illustrated in Figure A1. Hyperparameters for pretraining are detailed in Table A1-A2. During distillation, the gating network converges to nearly static weights of approximately 0.56 for DINOv3 and 0.43 for Chronos across inputs, effectively learning fixed teacher contributions rather than input-dependent dynamic routing.

Downstream Fine-tuning. Following CBraMod and EEGMamba, the last layer representations are flattened and passed to a classifier. The classifier consists of a 3-layer MLP with ELU activation and dropout. The model and classifier are jointly fine-tuned for each downstream dataset over 50 epochs. The best checkpoint is saved based on validation performance and evaluated on the test set.

Baselines and Metrics. We compare the multi-teacher distillation pretrained checkpoint against self-supervised checkpoints provided by CBraMod and EEGMamba, in order to understand the gains of the proposed pretraining framework. For reference, we also include traditional EEG models and EEG FMs in Table A4-A5. We report Balanced Accuracy, AUC-PR, and AUROC for binary classification and Balanced Accuracy, Cohen's Kappa, and Weighted-F1 for multi-class classification.

5.2 Experimental Results

5.2.1 Comparison of MTDP and self-supervised pretraining

Experimental results on 12 downstream datasets are presented in Table 2. "CBraMod 100%" and "EEGMamba 100%"[1] represents the self-supervised pretrained checkpoint provided by the authors.

[1]The results differ from the values reported in the EEGMamba paper, as we reran finetuning after identifying a bug in the EEGMamba codebase, where they used the test set (as opposed to the validation set) to save checkpoints.

Table 2: Comparison of EEG foundation model performance on 12 downstream tasks over three seeds. "25%" and "100%" indicate pretraining dataset size. Metrics acronyms: Balanced Accuracy (B. Acc.), Weighted F1 (F1-W), Cohen's Kappa (Kappa).

Methods	FACED			Physionet-MI			SHU-MI		
	B. Acc.	AUC-PR	AUROC	B. Acc.	AUC-PR	AUROC	B. Acc.	AUC-PR	AUROC
EEGMamba 25% [40]	47.95±0.66	41.26±0.65	47.89±0.47	58.23±0.58	44.41±0.60	58.43±0.44	60.41±0.35	65.52±0.92	64.71±0.64
EEGMamba 100% [40]	49.91±0.70	43.46±0.63	50.27±0.44	59.83±0.42	46.43±0.56	59.88±0.40	60.16±0.22	66.53±0.46	65.02±1.30
EEGMamba-MTDP 25%	48.83±0.51	42.09±0.75	48.63±0.79	60.62±0.55	47.49±0.73	60.93±0.63	**61.06±0.67**	**68.49±0.88**	**66.93±1.41**
EEGMamba-MTDP 100%	**52.30±0.47**	**45.87±0.53**	**51.93±0.61**	**62.40±0.69**	**49.86±0.92**	**62.51±0.58**	60.91±0.20	68.19±0.25	65.87±0.57
CBraMod 25% [39]	54.88±0.92	49.06±1.03	55.29±0.82	61.47±0.84	48.63±1.12	61.62±0.76	63.16±0.39	69.75±0.82	69.76±0.84
CBraMod 100% [39]	55.09±0.89	50.41±1.22	56.18±0.93	64.17±0.91	52.22±1.69	64.27±1.00	63.70±1.51	71.39±0.88	69.88±0.68
CBraMod-MTDP 25%	56.71±0.45	50.96±0.57	56.79±0.52	62.96±0.64	50.60±0.85	62.90±0.59	**64.11±0.28**	71.43±0.60	**70.52±0.24**
CBraMod-MTDP 100%	**57.33±0.66**	**51.67±0.63**	**57.45±0.48**	**64.91±0.92**	**53.21±1.22**	**64.97±0.88**	63.49±0.72	**71.99±0.76**	70.31±1.13

Methods	BCIC2020-3			Mumtaz2016			MentalArithmetic		
	B. Acc.	AUC-PR	AUROC	B. Acc.	AUC-PR	AUROC	B. Acc.	AUC-PR	AUROC
EEGMamba 25% [40]	46.62±1.63	33.28±2.04	46.56±1.62	88.02±0.82	97.30±0.29	96.94±0.62	57.99±2.28	35.28±1.56	67.99±1.75
EEGMamba 100% [40]	46.27±0.81	32.83±1.01	46.26±0.83	89.17±0.18	96.41±0.11	95.39±0.19	54.98±3.23	38.80±1.86	66.51±1.59
EEGMamba-MTDP 25%	52.67±1.14	40.83±1.42	52.60±1.17	93.60±0.09	98.81±0.28	98.74±0.31	70.02±1.12	67.90±1.11	81.25±1.45
EEGMamba-MTDP 100%	**53.47±1.67**	**41.83±2.08**	**53.43±1.76**	**94.13±0.30**	**98.86±0.06**	**98.79±0.07**	**74.54±1.97**	**72.52±1.85**	**84.70±1.10**
CBraMod 25% [39]	44.31±0.91	30.39±1.13	44.27±0.88	90.98±0.78	98.26±0.35	98.09±0.46	61.81±3.35	53.52±5.94	78.59±3.17
CBraMod 100% [39]	53.73±1.08	42.16±1.63	53.83±0.96	95.60±0.56	99.23±0.32	99.21±0.25	72.56±1.32	62.67±0.99	79.05±0.73
CBraMod-MTDP 25%	60.13±0.83	50.17±1.04	60.16±0.83	93.20±0.72	98.67±0.43	98.60±0.47	71.88±2.41	67.40±3.86	82.57±0.07
CBraMod-MTDP 100%	**62.44±0.80**	**53.06±1.00**	**62.49±0.81**	**95.71±0.13**	**99.44±0.16**	**99.43±0.17**	**75.46±2.09**	**73.39±2.30**	**87.42±0.53**

Methods	ISRUC			TUEV			CHB-MIT		
	B. Acc.	AUC-PR	AUROC	B. Acc.	AUC-PR	AUROC	B. Acc.	AUC-PR	AUROC
EEGMamba 25% [40]	75.42±0.40	71.67±0.53	77.80±0.42	62.28±1.76	56.73±1.43	77.60±0.66	53.93±2.73	41.18±5.98	91.29±0.28
EEGMamba 100% [40]	75.84±0.61	72.35±0.56	78.32±0.50	60.65±0.73	58.59±2.27	78.05±1.16	56.76±2.05	35.50±0.89	89.99±1.44
EEGMamba-MTDP 25%	77.83±0.26	73.88±0.31	79.73±0.40	65.65±0.55	64.88±1.87	81.61±0.87	69.31±1.61	**47.82±2.41**	91.61±1.62
EEGMamba-MTDP 100%	**79.03±0.47**	**75.48±0.86**	**80.74±0.55**	**68.17±1.14**	**70.17±0.79**	**84.33±0.27**	**73.11±3.51**	46.65±5.09	**92.62±1.25**
CBraMod 25% [39]	77.09±0.54	73.92±0.74	79.52±0.52	63.54±0.91	64.21±0.99	81.20±0.66	72.61±0.95	44.13±5.34	90.38±1.62
CBraMod 100% [39]	78.65±1.10	74.42±1.52	80.11±0.99	66.59±1.24	67.44±1.21	83.31±0.71	**73.98±2.84**	36.89±3.82	88.92±1.54
CBraMod-MTDP 25%	79.15±0.27	75.23±0.43	80.55±0.42	66.41±0.95	69.97±0.66	84.49±0.48	71.32±8.09	60.30±2.66	**94.20±0.84**
CBraMod-MTDP 100%	**79.25±0.12**	**75.82±0.80**	**81.06±0.58**	**67.18±2.20**	**72.25±1.48**	**85.51±0.86**	71.78±3.24	**65.50±1.73**	94.03±0.29

Methods	TUAB			SEED-V			BCIC-IV-2a		
	B. Acc.	AUC-PR	AUROC	B. Acc.	AUC-PR	AUROC	B. Acc.	AUC-PR	AUROC
EEGMamba 25% [40]	78.74±0.55	86.43±1.19	86.01±0.82	40.28±0.18	25.89±0.38	41.15±0.36	41.03±1.35	21.37±1.79	34.89±3.50
EEGMamba 100% [40]	79.32±0.87	86.96±1.17	86.75±1.01	41.51±0.21	27.29±0.14	42.18±0.16	50.81±0.33	34.41±0.44	49.98±0.61
EEGMamba-MTDP 25%	79.43±0.39	87.62±0.51	86.43±0.37	41.29±0.21	26.79±0.29	41.83±0.22	53.76±0.27	38.35±0.35	52.75±0.58
EEGMamba-MTDP 100%	**80.46±0.31**	**87.78±0.42**	**86.85±0.32**	**42.06±0.21**	**27.83±0.30**	**42.81±0.23**	**55.12±0.78**	**40.16±1.04**	**54.39±0.95**
CBraMod 25% [39]	78.68±0.59	87.45±0.07	87.19±0.32	39.74±0.51	25.06±0.64	40.41±0.53	48.78±0.63	31.71±0.83	47.61±0.99
CBraMod 100% [39]	**82.49±0.22**	**92.21±0.08**	**91.56±0.11**	40.91±0.97	25.69±1.43	41.01±1.08	51.38±0.66	35.18±0.94	49.84±0.85
CBraMod-MTDP 25%	79.95±0.45	88.39±1.01	87.81±0.73	41.48±0.15	27.35±0.19	42.32±0.16	59.46±0.54	45.95±0.72	58.73±0.68
CBraMod-MTDP 100%	81.19±0.26	88.83±0.14	87.76±0.34	**41.80±0.13**	**27.70±0.20**	**42.69±0.12**	**60.24±0.61**	**46.99±0.81**	**59.35±1.21**

"MTDP 25%" and "MTDP 100%" represents the checkpoint pretrained using the proposed MTDP framework with 25% and 100% of the TUEG dataset.

"CBraMod-MTDP 25%" and "EEGMamba-MTDP 25%" outperforms "CBraMod 100%" and "EEGMamba 100%" on 9/12 and 10/12 datasets respectively. When scaling the pretraining dataset to 100%, we observe improved performance across all datasets for EEGMamba, and all datasets except SHU-MI and ISRUC for CBraMod. When using CBraMod as the backbone, significant performance gains are observed for BCIC2020-3 (8.80/11.01/8.74%), MentalArithmetic (4.87/10.63/8.32%), CHB-MIT (1.44/28.95/5.29%), and BCIC-IV-2a (8.43/11.23/9.27%). When using EEGMamba as the

backbone, significant performance gains are observed for MentalArithmetic (19.45/32.14/15.96%), TUEV (9.64/13.15/7.15%), and CHB-MIT (19.59/5.17/1.35%).

These performance gains are notable, as we only replace the self-supervised pretraining algorithm with the proposed multi-teacher distillation pretraining algorithm, while maintaining a downstream pipeline that is strictly identical to the baselines. The result highlights the merit of MTDP as a promising pretraining framework for EEG FMs.

5.2.2 Ablation Studies

In Table 3, we perform ablation studies to show the effectiveness of teacher fusion from stage 1 of the pretraining presented in Section 4.2.1, as well as introducing an additional teacher, wav2vec 2.0 [6]. All experiments are conducted using CBraMod with 25% pretraining dataset over 1 seed. Set 1 distills from Chronos. Set 2 distills from DINO. Similar to AM-RADIO [32], we also distill from both teachers by summing the loss from each teacher, as denoted by Set 3. Set 4 is the full MTDP framework proposed in Section 4.2. Set 5 further extends the results from Table 2 to introduce a third teacher from the audio domain, wav2vec 2.0.

Table 3: Ablation of different distillation methods

Component	Set 1	Set 2	Set 3	Set 4	Set 5
Chronos	✓	×	✓	✓	✓
DINOv3	×	✓	✓	✓	✓
wav2vec	×	×	×	×	✓
Two-Stage	×	×	×	✓	✓
FACED	51.13	54.40	56.39	56.57	57.58
PhysioNetMI	62.57	62.62	62.56	62.74	63.19
SHU-MI	62.67	62.27	62.11	63.78	64.04
BCI2020-3	37.73	56.93	42.80	59.87	62.00

Comparing Set 1 and Set 2, we observe that distilling from Chronos is better than distilling from DINOv3 on SHU-MI, but worse on FACED, PhysioNetMI, and BCI2020-3. This suggests that the individual teachers provide representations that excel for different downstream tasks. In addition, we observe that Set 3 performs better than Set 1 and Set 2 in FACED, worse performance in SHU-MI and PhysioNetMI, and between the two in BCI2020-3. This suggests that simply summing the losses from both teachers does not always yield better performance than distilling from a single teacher. Overall, Set 4 achieves the best performance across the board compared to Sets 1-3. This highlights the effectiveness of the proposed framework in fusing teacher representations. Furthermore, Set 5 demonstrates that incorporating an audio domain teacher like wav2vec 2.0 provides complementary features that further enhance the model's performance.

5.2.3 Linear Probing Comparison

Similar to Section 3, we compare the representation quality of CBraMod when pretrained using the self-supervised masked reconstruction pretraining framework and the multi-teacher distillation pretraining framework. We extract representations from frozen CBraMod and CBraMod-MTDP 25%, then analyze the linear separability of these representations by training a logistic regression classifier. Figure 4 summarizes the linear probing performance. It is observed that CBraMod-MTDP 25% outperforms CBraMod in 10 of 12 downstream datasets. These findings suggest that the proposed framework significantly enhances the model's ability to learn generalizable representations during pretraining.

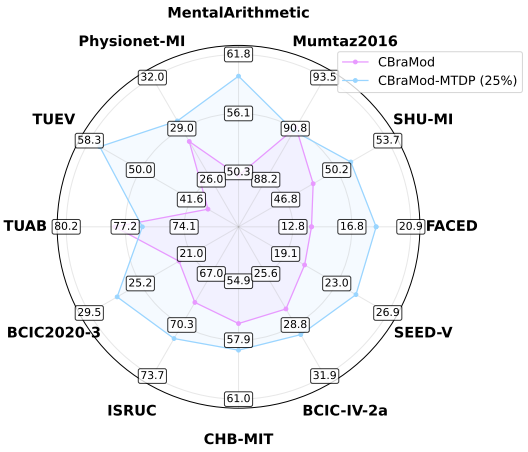

Figure 4: Linear probing performance of CBraMod and CBraMod-MTDP 25% on downstream tasks.

6 Conclusion

In this paper, we present a new two-stage paradigm for EEG FM pretraining that stands on the shoulders of well-established foundation models from other modalities. The first stage learns to fuse

out-of-domain teacher representations by weighing them using a learnable gate in an unsupervised manner. The second stage distills the fused representation from different teachers into an EEG foundation model. The proposed MTDP framework outperforms self-supervised masked reconstruction pretraining on the majority of downstream datasets, while maintaining competitive performance even when the pretraining data is reduced to 25%. We hope that this work inspires the community to rethink the pretraining of EEG FMs and embrace the integration of cross-domain insights.

Acknowledgements

The work of Chenqi Li was supported by the Cyril and Phillis Long Scholarship at The Queen's College in partnership with the Clarendon Fund. The work of Tim Denison was funded by the Royal Academy of Engineering and supported by the NIHR Oxford Health Biomedical Research Centre (NIHR203316). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. The Royal Academy of Engineering supported the work of Tingting Zhu under the Research Fellowship scheme.

References

[1] Khald Ali I Aboalayon, Miad Faezipour, Wafaa S Almuhammadi, and Saeid Moslehpour. Sleep stage classification using eeg signal analysis: a comprehensive survey and new investigation. *Entropy*, 18(9):272, 2016.

[2] Mohammed Al-Maamari, Mehdi Ben Amor, Jelena Mitrović, and Michael Granitzer. Mixture of Modular Experts: Distilling Knowledge from a Multilingual Teacher into Specialized Modular Language Models. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, pages 945–852. 2025.

[3] Turkey N Alotaiby, Saleh A Alshebeili, Tariq Alshawi, Ishtiaq Ahmad, and Fathi E Abd El-Samie. Eeg seizure detection and prediction algorithms: a survey. *EURASIP Journal on Advances in Signal Processing*, 2014(1):183, 2014.

[4] Hamdi Altaheri, Ghulam Muhammad, Mansour Alsulaiman, Syed Umar Amin, Ghadir Ali Altuwaijri, Wadood Abdul, Mohamed A Bencherif, and Mohammed Faisal. Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: A review. *Neural Computing and Applications*, 35(20):14681–14722, 2023.

[5] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

[6] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, 2020.

[7] Richard B Berry, Rita Brooks, Charlene Gamaldo, Susan M Harding, Robin M Lloyd, Stuart F Quan, Matthew T Troester, and Bradley V Vaughn. Aasm scoring manual updates for 2017 (version 2.4), 2017.

[8] Yuqi Chen, Kan Ren, Kaitao Song, Yansen Wang, Yifan Wang, Dongsheng Li, and Lili Qiu. EEGFormer: Towards Transferable and Interpretable Large-Scale EEG Foundation Model, 2024.

[9] Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A. Joshi, and Richard M. Leahy. Neuro-GPT: Towards A Foundation Model for EEG, 2024.

[10] Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. LUNA: Efficient and Topology-Agnostic Foundation Model for EEG Signal Analysis, 2025.

[11] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[12] Zitao Fang, Chenxuan Li, Hongting Zhou, Shuyang Yu, Guodong Du, Ashwaq Qasem, Yang Lu, Jing Li, Junsong Zhang, and Sim Kuan Goh. NeurIPT: Foundation Model for Neural Interfaces, 2025.

[13] Robert S Fisher, J Helen Cross, Jacqueline A French, Norimichi Higurashi, Edouard Hirsch, Floor E Jansen, Lieven Lagae, Solomon L Moshé, Jukka Peltola, Eliane Roulet Perez, et al. Operational classification of seizure types by the international league against epilepsy: Position paper of the ilae commission for classification and terminology. *Epilepsia*, 58(4):522–530, 2017.

[14] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[15] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network, 2015.

[17] Nan Huang, Haishuai Wang, Zihuai He, Marinka Zitnik, and Xiang Zhang. Repurposing Foundation Model for Generalizable Medical Time Series Classification, 2024.

[18] Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI, 2024.

[19] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models, 2024.

[20] Gyeongman Kim, Gyouk Chu, and Eunho Yang. Every Expert Matters: Towards Effective Knowledge Distillation for Mixture-of-Experts Language Models, 2025.

[21] Jun Li, Che Liu, Sibo Cheng, Rossella Arcucci, and Shenda Hong. Frozen Language Model Helps ECG Zero-Shot Learning. In *Medical Imaging with Deep Learning*, pages 402–415, 2024.

[22] Linhao Li, Han Zang, Yu Wang, Yongfeng Dong, and Liang Yang. Multi-expert collaboration: Enhancing heterogeneous knowledge independence and alignment in knowledge distillation. 2024.

[23] Xiang Li, Yazhou Zhang, Prayag Tiwari, Dawei Song, Bin Hu, Meihong Yang, Zhigang Zhao, Neeraj Kumar, and Pekka Marttinen. Eeg based emotion recognition: A tutorial and review. *ACM Computing Surveys*, 55(4):1–57, 2022.

[24] Zekun Li, Shiyang Li, and Xifeng Yan. Time Series as Images: Vision Transformer for Irregularly Sampled Time Series, 2023.

[25] Dingkun Liu, Yuheng Chen, Zhu Chen, Zhenyao Cui, Yaozhi Wen, Jiayu An, Jingwei Luo, and Dongrui Wu. Eeg foundation models: Progresses, benchmarking, and open problems. *arXiv preprint arXiv:2601.17883*, 2026.

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

[27] Yuang Liu, Wei Zhang, and Jun Wang. Adaptive Multi-Teacher Multi-level Knowledge Distillation. *Neurocomputing*, 415:106–113, 2020. doi: 10.1016/j.neucom.2020.07.048.

[28] Yunfei Luo, Yuliang Chen, Asif Salekin, and Tauhidur Rahman. Toward Foundation Model for Multivariate Wearable Sensing of Physiological Signals, 2025.

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

[30] Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A Foundation Model for EEG – Adapting to Any Setup with Large-Scale Pretraining on 25,000 Subjects, 2025.

[31] Cuong Pham, Tuan Hoang, and Thanh-Toan Do. Collaborative Multi-Teacher Knowledge Distillation for Learning Low Bit-width Deep Neural Networks, 2022.

[32] Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12490–12500, 2024.

[33] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025.

[34] Siuly Siuly, Yan Li, and Yanchun Zhang. Electroencephalogram (eeg) and its background. In *EEG signal analysis and classification: Techniques and applications*, pages 3–21. Springer, 2017.

[35] Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. Scale efficiently: Insights from pre-training and fine-tuning transformers. *arXiv preprint arXiv:2109.10686*, 2021.

[36] Anna Tegon, Thorir Mar Ingolfsson, Xiaying Wang, Luca Benini, and Yawei Li. FEMBA: Efficient and Scalable EEG Analysis with a Bidirectional Mamba Foundation Model, 2025.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

[39] Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A Criss-Cross Brain Foundation Model for EEG Decoding, 2025.

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

[41] Qinfan Xiao, Ziyun Cui, Chi Zhang, Siqi Chen, Wen Wu, Andrew Thwaites, Alexandra Woolgar, Bowen Zhou, and Chao Zhang. BrainOmni: A Brain Foundation Model for Unified EEG and MEG Signals, 2025.

[42] Zhitian Xie, Yinger Zhang, Chenyi Zhuang, Qitao Shi, Zhining Liu, Jinjie Gu, and Guannan Zhang. MoDE: A Mixture-of-Experts Model with Mutual Distillation among the Experts, 2024.

[43] Chaoqi Yang, M. Westover, and Jimeng Sun. BIOT: Biosignal Transformer for Cross-data Learning in the Wild. *Advances in Neural Information Processing Systems*, 36:78240–78260, 2023.

[44] Tianchen Yu and Guoxing Wang. Healthcare Event Prediction via Iterative in-Context Learning of LLM. In *2025 10th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, pages 205–210, 2025. doi: 10.1109/ICCCBDA64898.2025.11030505.

[45] Zhizhang Yuan, Daoze Zhang, Junru Chen, Gefei Gu, and Yang Yang. Brant-2: Foundation Model for Brain Signals, 2024.

[46] Zhizhang Yuan, Fanqi Shen, Meng Li, Yuguo Yu, Chenhao Tan, and Yang Yang. BrainWave: A Brain Signal Foundation Model for Clinical Applications, 2025.

[47] Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation Model for Intracranial Neural Signal. In *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023.

[48] Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. Time-VLM: Exploring Multimodal Vision-Language Models for Augmented Time Series Forecasting, 2025.

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

Appendix Table of Contents

A Further Discussion **14**

A.1 Repurposing Models for Time Series 14

A.2 Limitations 14

A.3 Broader Impact 14

B Experiment Setting and Additional Results **15**

B.1 Pretraining Dataset Preprocessing 15

B.2 Pretraining Hyperparameters 15

B.3 Training Loss Curves 15

B.4 Fine-tuning Hyperparameters 16

B.5 Extended Results 17

C Algorithm for Multi-Teacher Distillation Pretraining Framework **17**

A Further Discussion

A.1 Repurposing Models for Time Series

It is somewhat surprising that out-of-domain teachers transfer well to EEG. A potential explanation is that the teachers from other domains are pretrained on orders of magnitude more data and are orders of magnitude larger in terms of parameters compared to EEG FMs. They are robust pattern recognizers, which transfer well to time series like EEG.

Existing literature has also reported similar observations that models from other domains can be repurposed or reprogrammed for time series tasks [24, 19, 48]. Early efforts investigated the feasibility of whether the structural understanding of vision transformers and the reasoning and sequential capabilities of LLMs could be transferred to time series applications. It was demonstrated that plotting irregularly sampled time series into line graph images allows pretrained vision transformers to outperform specialized baselines, offering a robust and simplified framework for extreme data scarcity [24]. Time-LLM [19] demonstrated that a reprogramming layer that aligns time series patches with text prototype embeddings enables LLMs to perform forecasting without changing the backbone parameters. Time-VLM [48] further extended the repurposing paradigm to vision-language models and established a new frontier for multimodal time series forecasting. Similar cross-modal repurposing has also been adopted for domain-specific applications, such as medical time series classification [17], zero-shot ECG learning [21], and prediction of healthcare events [44]. These studies provide further evidence and intuition on why out-of-domain teachers transfer surprisingly well to time series, but they also lack further analysis and theoretical support on why out-of-domain priors are beneficial for time series tasks.

A.2 Limitations

The use of DINOv3 and Chronos incurs additional data processing and inference costs compared to masked reconstruction pretraining. Although we minimize this cost by storing the teacher representations, it introduces additional hard disk and I/O burdens. Furthermore, while MTDP outperforms self-supervised pretraining on most datasets, it underperforms on certain tasks, such as CBraMod on TUAB. Future work would also benefit from further analysis and theoretical support on why out-of-domain teachers transfer well to EEG, beyond the observed performance gains.

A.3 Broader Impact

This paper advances EEG foundation models, which have potential for healthcare and assistive technology. The advancement reduces dependency on large, task-specific labeled datasets, improving accessibility in clinical and resource-constrained settings.

However, enhanced capabilities of foundation models raise potential privacy and equity concerns. EEG analysis could potentially reveal sensitive information, necessitating strict data protocols to prevent unauthorized usage. Furthermore, there is a risk of algorithmic bias, where the model may perform inconsistently across different populations if the training data lacks demographic diversity.

B Experiment Setting and Additional Results

B.1 Pretraining Dataset Preprocessing

We select 16 channels (Fp1, Fp2, F7, F3, F4, F8, T3, C3, C4, T4, T5, P3, P4, T6, O1, O2), apply band-pass (0.3-75 Hz) and notch (60 Hz) filter, resample to 200 Hz, segment into 30-second non-overlapping segments, discard samples with absolute amplitude exceeding 100 μV, and normalize to -1 to 1 range by setting unit to 100 μV. Unlike CBraMod, we apply an additional step to remove TUAB and TUEV samples to avoid downstream data contamination. This results in a total of 1,059,523 samples, which is around 8,829 hours in total.

B.2 Pretraining Hyperparameters

Table A1-A2 summarizes the hyperparameters used for pretraining.

Table A1: Pretraining hyperparameters (CBraMod)

Hyperparameters	Settings
Data	
Channels	16
Time points	6000
Patch dimension	200
Sequence length	30 s
CBraMod Model Architecture	
Patch Encoder (Time-Domain)	
Input dimension	{1, 25, 25}
Output dimension	{25, 25, 25}
Kernel size	{49, 3, 3}
Stride	{25, 1, 1}
Padding	{24, 1, 1}
Patch Encoder (Freq-Domain)	
FFT function	torch.fft.rfft
Fully-connected layer	(101, 200)
Positional Encoder	
Input dimension	200
Output dimension	200
Kernel size	(19, 7)
Stride	(1, 1)
Padding	(9, 3)
Criss-cross Transformer	
Layers	12
Hidden dimension	200
Heads	8
S-Attention heads	4
T-Attention heads	4
Feed-forward dimension	800

Hyperparameters	Settings
Stage 1 Teacher Representation Fusion	
Epochs	1
Batch size	256
Optimizer	AdamW
Learning rate	5e-4
Adam β	(0.9, 0.999)
Adam ϵ	1e-8
Weight decay	5e-2
Clipping gradient norm	1
Weights init	Kaiming normalization
Loss	MSELoss
Stage 2 Knowledge Distillation	
Epochs	40
Batch size	256
Optimizer	AdamW
Learning rate	5e-4
Adam β	(0.9, 0.999)
Adam ϵ	1e-8
Weight decay	5e-2
Scheduler	CosineAnnealingLR
Cosine cycle epochs	40
Minimal learning rate	1e-5
Clipping gradient norm	1
Weights init	Kaiming normalization
Loss	CosineEmbeddingLoss

B.3 Training Loss Curves

Figure A1 plots the loss curve of stage 1 and stage 2 pretraining for CBraMod and EEGMamba.

Table A2: Pretraining hyperparameters (EEGMamba)

Hyperparameters	Settings
Data	
Channels	16
Time points	6000
Patch dimension	200
Sequence length	30 s
EEGMamba Model Architecture	
Patch Encoder (Time-Domain)	
Input dimension	1
Output dimension	25
Kernel size	49
Stride	25
Padding	24
Patch Encoder (Freq-Domain)	
FFT function	torch.fft.rfft
Fully-connected layer	(101, 200)
Positional Encoder	
Input dimension	200
Output dimension	200
Kernel size	(7, 7)
Stride	(1, 1)
Padding	(3, 3)
Mamba Encoder	
Layers	12
Hidden dimension	200
State expansion factor	64
Head dimension	50
Local convolution width	4
Block expansion factor	2
Feed-forward dimension	936

Hyperparameters	Settings
Stage 1 Teacher Representation Fusion	
Epochs	1
Batch size	256
Optimizer	AdamW
Learning rate	5e-4
Adam β	(0.9, 0.999)
Adam ϵ	1e-8
Weight decay	5e-2
Clipping gradient norm	1
Weights init	Kaiming normalization
Loss	MSELoss
Stage 2 Knowledge Distillation	
Epochs	40
Batch size	256
Optimizer	AdamW
Learning rate	5e-4
Adam β	(0.9, 0.999)
Adam ϵ	1e-8
Weight decay	5e-2
Scheduler	CosineAnnealingLR
Cosine cycle epochs	40
Minimal learning rate	1e-5
Clipping gradient norm	1
Weights init	Kaiming normalization
Loss	CosineEmbeddingLoss

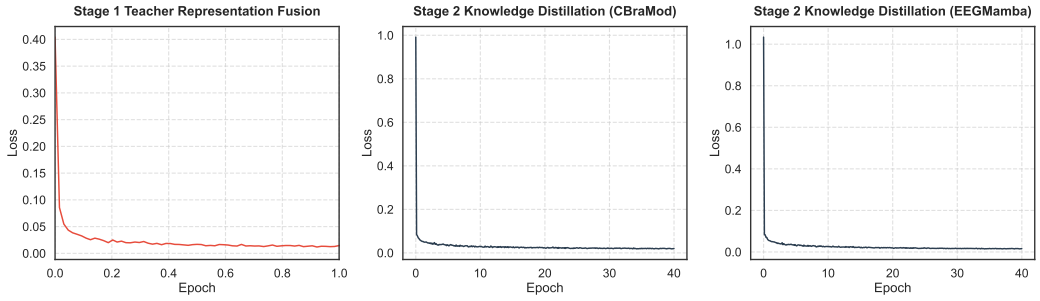

Figure A1: Loss curve of stage 1 and stage 2 pretraining

B.4 Fine-tuning Hyperparameters

We adopt the fine-tuning code from CBraMod and EEGMamba without any modification. We load our distilled checkpoint instead of the self-supervised checkpoint provided by CBraMod and EEGMamba. We perform a grid search over three different learning rates and a binary multi-LR setting. Multi-LR adjusts the classifier learning rate to be five times larger. Table A3 summarizes the fine-tuning hyperparameters for both CBraMod and EEGMamba.

Table A3: Fine-tuning hyperparameters

Hyperparameters	**Settings**
Epochs	50
Batch size	64
Dropout	0.1
Optimizer	AdamW
Bakbone Learning rate	{5e-5, 1e-4, 5e-4}
Multi-LR	{Yes, No}
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
Label smoothing (multi-class classification)	0.1

B.5 Extended Results

The main focus of this work is to understand the gain of multi-teacher distillation pretraining over self-supervised masked reconstruction pretraining for EEG FMs. For reference, we also include traditional EEG models and other state-of-the-art EEG FMs, which differ in the pretraining dataset used and downstream finetuning pipeline, in Table A4-A5.

C Algorithm for Multi-Teacher Distillation Pretraining Framework

We summarize the proposed multi-teacher distillation pretraining framework in Algorithm 1.

Table A4: Comparison of EEG foundation model performance on 12 downstream tasks. "25%" and "100%" indicate pretraining dataset size. Metrics acronyms: Balanced Accuracy (B. Acc.), Weighted F1 (F1-W), Cohen's Kappa (Kappa). (Part 1).

Methods	FACED			Physionet-MI			SHU-MI		
	B. Acc.	AUC-PR	AUROC	B. Acc.	AUC-PR	AUROC	B. Acc.	AUC-PR	AUROC
EEGNet	40.90±1.22	33.42±2.51	41.24±1.41	58.14±1.25	44.68±1.99	57.96±1.15	58.89±1.77	63.11±1.42	62.83±1.52
EEGConformer	45.59±1.25	38.58±1.86	45.14±1.07	60.49±1.04	47.36±1.71	60.62±0.95	59.00±1.07	63.70±0.93	63.51±1.01
SPaRCNet	46.73±1.55	39.78±2.89	47.29±1.33	59.32±1.52	45.64±2.34	59.37±1.47	59.78±0.97	65.10±0.62	64.31±0.82
ContraWR	48.87±0.78	42.31±1.51	48.84±0.74	58.92±1.33	45.27±2.48	59.18±1.16	58.73±1.28	63.15±1.05	62.73±1.13
CNN-Transformer	46.97±1.32	40.17±1.68	47.20±1.25	60.53±1.18	47.25±2.23	60.41±1.05	59.75±1.69	64.12±0.76	63.23±0.82
FFCL	46.73±1.58	39.87±3.83	46.99±1.45	57.26±0.92	43.23±1.82	57.01±0.79	56.92±2.52	59.43±1.72	60.14±1.68
ST-Transformer	48.10±0.79	41.37±1.33	47.95±0.96	60.35±0.81	47.12±1.99	60.53±0.75	59.92±2.06	63.94±1.22	64.31±1.11
BIOT [43]	51.18±1.18	44.76±2.54	51.36±1.12	61.53±1.54	48.75±2.72	61.58±1.97	61.79±1.83	67.70±1.19	66.09±1.27
EEGPT [38]	-	-	-	-	-	-	-	-	-
LaBraM [18]	52.73±1.07	46.98±1.88	52.88±1.02	61.73±1.22	49.12±1.92	61.77±1.41	61.66±1.92	67.61±0.83	66.04±0.91
NeurIPT [12]	-	-	-	-	-	-	-	-	-
CSBrain [49]	57.52±0.42	52.04±0.36	57.96±0.31	63.04±0.90	50.71±1.20	63.08±0.95	64.17±0.37	72.30±0.79	72.00±0.58
LUNA-Huge [10]	-	-	-	-	-	-	-	-	-
REVE [30]	56.46±1.64	50.80±1.91	56.59±1.72	64.80±1.40	53.06±1.87	64.84±1.70	-	-	-
EEGMamba 100% [40]	49.91±0.70	43.46±0.63	50.27±0.44	59.83±0.42	46.43±0.56	59.88±0.40	60.16±0.22	66.53±0.46	65.02±1.30
EEGMamba-MTDP 25%	48.83±0.51	42.09±0.75	48.63±0.79	60.62±0.55	47.49±0.73	60.93±0.63	61.06±0.67	68.49±0.88	66.93±1.41
EEGMamba-MTDP 100%	52.30±0.47	45.87±0.53	51.93±0.61	62.40±0.69	49.86±0.92	62.51±0.58	60.91±0.20	68.19±0.25	65.87±0.57
CBraMod 100% [39]	55.09±0.89	50.41±1.22	56.18±0.93	64.17±0.91	52.22±1.69	64.27±1.00	63.70±1.51	71.39±0.88	69.88±0.68
CBraMod-MTDP 25%	56.71±0.45	50.96±0.57	56.79±0.52	62.96±0.64	50.60±0.85	62.90±0.59	64.11±0.28	71.43±0.60	70.52±0.24
CBraMod-MTDP 100%	57.33±0.66	51.67±0.63	57.45±0.48	64.91±0.92	53.21±1.22	64.97±0.88	63.49±0.72	71.99±0.76	70.31±1.13

Methods	BCIC2020-3			Mumtaz2016			MentalArithmetic		
	B. Acc.	AUC-PR	AUROC	B. Acc.	AUC-PR	AUROC	B. Acc.	AUC-PR	AUROC
EEGNet	44.13±0.96	30.16±1.23	44.13±1.02	92.32±1.04	96.26±0.95	96.39±0.93	67.70±1.16	57.63±1.02	73.21±1.08
EEGConformer	45.06±1.33	31.33±1.83	44.88±1.54	93.08±1.17	96.84±1.05	97.02±1.01	68.05±1.23	58.29±1.34	74.24±1.28
SPaRCNet	44.26±1.56	30.33±2.33	44.20±1.08	93.16±0.95	97.54±0.65	97.81±0.83	68.79±1.07	58.25±1.93	74.18±1.32
ContraWR	42.57±1.62	30.78±2.18	44.07±1.82	91.95±1.15	95.89±1.02	96.21±0.92	66.31±0.97	57.87±1.64	73.32±0.82
CNN-Transformer	45.33±0.92	31.66±1.18	45.06±1.27	93.05±0.68	97.57±0.74	97.42±0.59	67.79±2.68	57.77±2.85	72.58±3.36
FFCL	46.78±1.97	33.01±3.59	46.89±2.05	93.14±0.38	97.17±0.21	97.53±0.33	67.98±1.42	57.86±2.66	73.30±1.98
ST-Transformer	41.26±1.22	29.41±1.59	42.47±1.38	91.35±1.03	95.78±0.86	95.94±0.59	66.31±1.73	56.72±2.59	71.32±1.74
BIOT [43]	49.20±0.86	36.50±1.76	49.17±0.79	93.58±0.52	97.36±0.34	97.58±0.42	68.75±1.86	60.04±1.95	75.36±1.44
EEGPT [38]	-	-	-	-	-	-	-	-	-
LaBraM [18]	50.60±1.55	38.00±2.42	50.54±2.05	94.09±0.79	97.98±0.93	97.82±0.57	69.09±1.25	59.99±1.55	77.21±0.93
NeurIPT [12]	-	-	-	98.03±0.62	99.81±0.44	99.79±0.45	86.46±1.07	78.27±1.97	91.11±1.63
CSBrain [49]	60.04±1.87	50.06±2.33	60.03±1.92	96.43±1.55	99.36±0.34	99.56±0.20	75.58±1.06	66.96±2.21	84.78±2.97
LUNA-Huge [10]	-	-	-	-	-	-	-	-	-
REVE [30]	56.35±1.23	45.43±1.54	56.33±1.24	96.44±0.97	99.61±0.13	99.57±0.15	76.60±3.55	74.70±8.07	84.50±5.14
EEGMamba 100% [40]	46.27±0.81	32.83±1.01	46.26±0.83	89.17±0.18	96.41±0.11	95.39±0.19	54.98±3.23	38.80±1.86	66.51±1.59
EEGMamba-MTDP 25%	52.67±1.14	40.83±1.42	52.60±1.17	93.60±0.09	98.81±0.28	98.74±0.31	70.02±1.12	67.90±1.11	81.25±1.45
EEGMamba-MTDP 100%	53.47±1.67	41.83±2.08	53.43±1.76	94.13±0.30	98.86±0.06	98.79±0.07	74.54±1.97	72.52±1.85	84.70±1.10
CBraMod 100% [39]	53.73±1.08	42.16±1.63	53.83±0.96	95.60±0.56	99.23±0.32	99.21±0.25	72.56±1.32	62.67±0.99	79.05±0.73
CBraMod-MTDP 25%	60.13±0.83	50.17±1.04	60.16±0.83	93.20±0.72	98.67±0.43	98.60±0.47	71.88±2.41	67.40±3.86	82.57±0.07
CBraMod-MTDP 100%	62.44±0.80	53.06±1.00	62.49±0.81	95.71±0.13	99.44±0.16	99.43±0.17	75.46±2.09	73.39±2.30	87.42±0.53

Table A5: Comparison of EEG foundation model performance on 12 downstream tasks. "25%" and "100%" indicate pretraining dataset size. Metrics acronyms: Balanced Accuracy (B. Acc.), Weighted F1 (F1-W), Cohen's Kappa (Kappa). (Part 2).

Methods	**ISRUC**			**TUEV**			**CHB-MIT**		
	B. Acc.	AUC-PR	AUROC	B. Acc.	AUC-PR	AUROC	B. Acc.	AUC-PR	AUROC
EEGNet	71.54±1.21	70.40±1.73	75.13±1.24	38.76±1.43	35.77±1.55	65.39±1.20	56.58±1.06	19.14±1.82	80.48±1.36
EEGConformer	74.00±1.33	71.43±1.62	76.34±1.51	40.74±1.64	39.67±1.95	69.83±1.52	59.76±1.41	22.09±2.15	82.26±1.70
SPaRCNet	74.87±0.75	70.97±1.32	76.24±0.92	41.61±2.62	42.33±1.81	70.24±1.04	58.76±1.91	12.47±1.19	81.43±1.48
ContraWR	74.02±1.26	71.78±1.56	76.10±1.37	43.84±3.49	39.12±2.37	68.93±1.36	63.44±0.02	22.64±1.74	80.97±1.14
CNN-Transformer	73.63±0.87	71.29±1.21	77.19±1.05	40.87±1.61	38.15±1.34	68.54±2.93	63.89±0.67	24.79±2.27	86.62±0.82
FFCL	72.77±1.82	70.16±2.91	76.14±1.97	39.79±1.04	37.32±1.88	67.83±1.20	62.62±1.04	20.49±3.46	82.71±0.51
ST-Transformer	73.81±2.05	70.13±3.52	76.81±1.75	39.84±2.28	37.65±3.06	68.23±1.90	59.15±1.95	14.22±0.94	82.37±4.91
BIOT [43]	75.27±1.21	71.92±2.31	77.90±1.46	52.81±2.25	52.73±2.49	74.92±0.82	70.68±4.57	32.77±4.60	87.61±2.84
EEGPT [38]	-	-	-	62.32±1.14	63.51±1.34	81.87±0.63	-	-	-
LaBraM [18]	76.33±1.02	72.31±1.82	78.10±1.33	64.09±0.65	66.37±0.93	83.12±0.52	70.75±3.58	32.87±4.02	86.79±1.99
NeurIPT [12]	-	-	-	67.61±1.33	69.70±1.85	84.28±0.89	-	-	-
CSBrain [49]	79.25±0.30	74.06±1.02	79.90±0.91	69.03±0.59	68.33±0.47	83.33±0.57	72.62±1.15	51.64±4.49	89.15±3.21
LUNA-Huge [10]	-	-	-	-	-	-	-	-	-
REVE [30]	78.19±0.78	75.00±1.56	80.05±1.35	67.59±2.29	67.83±1.99	84.51±1.29	-	-	-
EEGMamba 100% [40]	75.84±0.61	72.35±0.56	78.32±0.50	60.65±0.73	58.59±2.27	78.05±1.16	56.76±2.05	35.50±0.89	89.99±1.44
EEGMamba-MTDP 25%	77.83±0.26	73.88±0.31	79.73±0.40	65.65±0.55	64.88±1.87	81.61±0.87	69.31±1.61	47.82±2.41	91.61±1.62
EEGMamba-MTDP 100%	79.03±0.47	75.48±0.86	80.74±0.55	68.17±1.14	70.17±0.79	84.33±0.27	73.11±3.51	46.65±5.09	92.62±1.25
CBraMod 100% [39]	78.65±1.10	74.42±1.52	80.11±0.99	66.59±1.24	67.44±1.21	83.31±0.71	73.98±2.84	36.89±3.82	88.92±1.54
CBraMod-MTDP 25%	79.15±0.27	75.23±0.43	80.55±0.42	66.41±0.95	69.97±0.66	84.49±0.48	71.32±8.09	60.30±2.66	94.20±0.84
CBraMod-MTDP 100%	79.25±0.12	75.82±0.80	81.06±0.58	67.18±2.20	72.25±1.48	85.51±0.86	71.78±3.24	65.50±1.73	94.03±0.29

Methods	**TUAB**			**SEED-V**			**BCIC-IV-2a**		
	B. Acc.	AUC-PR	AUROC	B. Acc.	AUC-PR	AUROC	B. Acc.	AUC-PR	AUROC
EEGNet	76.42±0.36	82.99±0.43	84.12±0.31	29.61±1.02	10.06±1.43	27.49±0.98	44.82±0.94	26.93±1.21	42.26±1.08
EEGConformer	77.58±0.49	84.27±0.54	84.45±0.38	35.37±1.12	17.72±1.74	34.87±1.36	46.96±1.06	29.24±1.41	45.33±1.28
SPaRCNet	78.96±0.18	84.14±0.18	86.76±0.12	29.49±0.78	11.21±1.39	29.79±0.83	46.35±1.17	28.47±1.47	44.32±1.26
ContraWR	77.46±0.41	84.21±1.04	84.56±0.74	35.46±1.05	19.05±1.88	35.44±1.21	46.78±1.25	29.05±1.60	44.13±1.42
CNN-Transformer	77.77±0.22	84.33±0.39	84.61±0.13	36.78±0.78	20.72±1.83	36.42±0.88	46.00±1.08	28.00±1.48	44.60±1.14
FFCL	78.48±0.38	84.48±0.65	85.69±0.51	36.41±0.92	20.78±2.01	36.45±1.32	44.70±1.43	26.27±1.76	42.38±1.39
ST-Transformer	79.66±0.23	85.21±0.26	87.07±0.19	30.52±0.72	10.83±1.21	28.33±1.05	45.75±1.45	27.33±1.98	44.71±1.42
BIOT [43]	79.59±0.57	87.92±0.23	88.15±0.43	38.37±1.87	22.61±2.62	38.56±2.03	47.48±0.93	29.97±1.39	46.07±1.25
EEGPT [38]	79.83±0.30	-	87.18±0.50	-	-	-	58.46±0.70	44.62±0.94	57.15±0.51
LaBraM [18]	81.40±0.19	89.65±0.16	90.22±0.09	39.76±1.38	23.86±2.09	39.74±1.11	48.69±0.85	31.59±1.54	47.58±1.03
NeurIPT [12]	82.93±0.16	90.40±0.22	89.49±0.21	41.04±0.21	26.29±0.39	41.58±0.37	55.04±0.72	40.05±1.21	53.76±0.86
CSBrain [49]	81.72±0.43	90.05±0.66	89.57±0.46	41.97±0.33	27.85±0.34	42.80±0.23	56.57±0.71	42.09±0.93	56.37±0.87
LUNA-Huge [10]	81.57±0.11	90.29±0.14	89.57±0.11	39.00±0.96	20.37±1.03	35.06±0.47	-	-	-
REVE [30]	83.15±0.14	92.81±0.09	92.45±0.13	-	-	-	63.96±0.95	51.94±1.26	63.39±1.10
EEGMamba 100% [40]	79.32±0.87	86.96±1.17	86.75±1.01	41.51±0.21	27.29±0.14	42.18±0.16	50.81±0.33	34.41±0.44	49.98±0.61
EEGMamba-MTDP 25%	79.43±0.39	87.62±0.51	86.43±0.37	41.29±0.21	26.79±0.29	41.83±0.22	53.76±0.27	38.35±0.35	52.75±0.58
EEGMamba-MTDP 100%	80.46±0.31	87.78±0.42	86.85±0.32	42.06±0.21	27.83±0.30	42.81±0.23	55.12±0.78	40.16±1.04	54.39±0.95
CBraMod 100% [39]	82.49±0.22	92.21±0.08	91.56±0.11	40.91±0.97	25.69±1.43	41.01±1.08	51.38±0.66	35.18±0.94	49.84±0.85
CBraMod-MTDP 25%	79.95±0.45	88.39±1.01	87.81±0.73	41.48±0.15	27.35±0.19	42.32±0.16	59.46±0.54	45.95±0.72	58.73±0.68
CBraMod-MTDP 100%	81.19±0.26	88.83±0.14	87.76±0.34	41.80±0.13	27.70±0.20	42.69±0.12	60.24±0.61	46.99±0.81	59.35±1.21

Algorithm 1 MTDP framework: two-stage multi-teacher distillation pretraining

Input: Pretraining dataset $\mathcal{D}$
Pretrained teachers $\{f_{\phi_k}\}_{k=1}^{K}$
Output: EEG foundation model f_θ
▷ **Stage 1: Teacher Representation Fusion**
Initialize gating network g_ψ
Initialize prediction heads $\{p_{\xi_k}\}_{k=1}^{K}$
for iteration $\leftarrow 1$ to N_{iter} **do**
  Sample EEG $\mathbf{x} \sim \mathcal{D}$
  Apply mask $\tilde{\mathbf{x}} \leftarrow \mathbf{x} \odot \mathbf{m}$
  for $k \leftarrow 1$ to K **do**
    Compute $\tilde{\mathbf{h}}_k = f_{\phi_k}(\tilde{\mathbf{x}})$ and $\mathbf{h}_k = f_{\phi_k}(\mathbf{x})$
  end for
  Compute $\mathbf{w} = [w_1, ..., w_K] = g_\psi([\tilde{\mathbf{h}}_1; \ldots; \tilde{\mathbf{h}}_K])$
  Compute $\tilde{\mathbf{h}}_{\text{fused}} = \sum_{k=1}^{K} w_k \cdot \tilde{\mathbf{h}}_k$
  Compute $\mathcal{L}_{\text{denoise}} = \sum_{k=1}^{K} \|p_{\xi_k}(\tilde{\mathbf{h}}_{\text{fused}}) - \mathbf{h}_k\|_2^2$
  Update ψ, ξ_k using $\nabla_\psi \mathcal{L}_{\text{denoise}}, \nabla_{\xi_k} \mathcal{L}_{\text{denoise}}$
end for
▷ **Stage 2: Knowledge Distillation**
Initialize EEG foundation model f_θ
Initialize projection heads $\{q_\zeta\}_{k=1}^{K}$
for iteration $\leftarrow 1$ to N_{iter} **do**
  Sample EEG $\mathbf{x} \sim \mathcal{D}$
  Compute $\mathbf{h}_{\text{fused}} = \sum_{k=1}^{K} w_k \cdot f_{\phi_k}(\mathbf{x})$
  Compute $\mathcal{L}_{\text{distill}} = 1 - \frac{\langle q_\zeta \circ f_\theta(x), \mathbf{h}_{\text{fused}} \rangle}{||q_\zeta \circ f_\theta(x)||_2 \cdot ||\mathbf{h}_{\text{fused}}||_2}$
  Update θ, ζ using $\nabla_\theta \mathcal{L}_{\text{distill}}, \nabla_\zeta \mathcal{L}_{\text{distill}}$
end for

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction are supported by experimental results presented in the subsequent sections.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in Section A.2.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 and 5 describe the proposed framework and experiment setups. Further implementation details are described in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All datasets used are publicly available. Code will be made publicly available upon publication.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experiment setups are described in Section 5. Experiment details are further described in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the mean and standard deviation over three seeds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute resources used are detailed in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Authors have reviewed and follow the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Discussions have been provided in Section A.3.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: All model architectures and datasets used in this study are already publicly available.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The owners of assets used in the paper have been properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code will be publicly available upon publication and is well-documented.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We use public datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: We use public datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.