EEG-DLite: Dataset Distillation for Efficient Large EEG Model Training

Yuting Tang[1], Weibang Jiang[1,2], Shanglin Li[3], Yong Li[4], Chenyu Liu[1], Xinliang Zhou[1], Yi Ding[1*], Cuntai Guan[1*]

[1]College of Computing and Data Science, Nanyang Technological University
[2]Shanghai Jiao Tong University
[3]Advanced Telecommunications Research Institute International
[4]Southeast University
yuting.tang@ntu.edu.sg, 935963004@sjtu.edu.cn, shanglin@atr.jp, mysee1989@gmail.com, {chenyu003, xinliang001}@e.ntu.edu.sg, {ding.yi, ctguan}@ntu.edu.sg

Abstract

Large-scale EEG foundation models have shown strong generalization across a range of downstream tasks, but their training remains resource-intensive due to the volume and variable quality of EEG data. In this work, we introduce EEG-DLite, a data distillation framework that enables more efficient pre-training by selectively removing noisy and redundant samples from large EEG datasets. EEG-DLite begins by encoding EEG segments into compact latent representations using a self-supervised autoencoder, allowing sample selection to be performed efficiently and with reduced sensitivity to noise. Based on these representations, EEG-DLite filters out outliers and minimizes redundancy, resulting in a smaller yet informative subset that retains the diversity essential for effective foundation model training. Through extensive experiments, we demonstrate that training on only 5 percent of a 2,500-hour dataset curated with EEG-DLite yields performance comparable to, and in some cases better than, training on the full dataset across multiple downstream tasks. To our knowledge, this is the first systematic study of pre-training data distillation in the context of EEG foundation models. EEG-DLite provides a scalable and practical path toward more effective and efficient physiological foundation modeling.

Introduction

Brain-computer interfaces (BCIs) enable direct communication between the human brain and external devices, unlocking a wide range of applications in neurorehabilitation (Wang et al. 2023), assistive technologies (Tang et al. 2024), and cognitive monitoring (Zhou et al. 2025). Among the various modalities used in BCIs, electroencephalography (EEG) is the most widely adopted due to its non-invasiveness and high temporal resolution. However, EEG signals are inherently noisy, high-dimensional, and subject to significant inter-subject variability, which presents challenges for learning robust and generalizable representations (Wang-Nöth et al. 2025; Carzaniga et al. 2025).

To address these challenges, recent advances have introduced EEG foundation models, which are large-scale neural networks pre-trained on extensive unlabeled EEG datasets using self-supervised objectives (Zhou et al. 2025). These models, often built on convolutional backbones or Transformer-based architectures, can exceed 400 million parameters. Typically, they are first pre-trained on large-scale EEG data from diverse BCI tasks and later fine-tuned for specific downstream applications. These models demonstrate competitive performance across various downstream tasks while maintaining a unified architectural framework (Jiang, Zhao, and Lu 2024; Yue et al. 2024; Wang et al. 2025). While effective, this pre-training paradigm is computationally expensive, often requiring significant GPU time, storage, and energy (Jiang, Zhao, and Lu 2024). The high cost also limits the feasibility of further exploration into parameter optimization and neural architecture search (Yu, Liu, and Wang 2023).

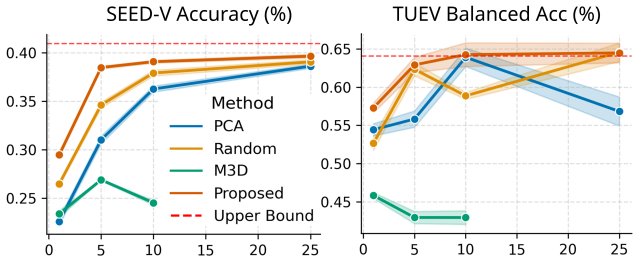

Figure 1: Performance trends of LaBraM on downstream tasks using pre-training datasets distilled by EEG-DLite at different ratios η.

Recent advances in computer vision have shown that it is possible to significantly reduce the computational burden of large-scale model training by distilling extensive datasets into compact, information-rich subsets (Zhang et al. 2024; Joshi, Ni, and Mirzasoleiman 2024; Killamsetty et al. 2021). Motivated by the broader goal of data-efficient pre-training, we turn our focus to EEG data, which presents unique and domain-specific challenges that demand tailored solutions. First, EEG signals exhibit a low signal-to-noise ratio (SNR), as neural activity is easily obscured by artifacts from eye movements, muscle activity, and external interference (Sadiya, Alhanai, and Ghassemi 2021). Even after preprocessing, residual noise can persist and adversely impact model learning. Second, EEG recordings often contain sig-

*Corresponding authors

nificant redundancy; temporally adjacent segments may capture overlapping or repetitive patterns, offering limited new information (Amin et al. 2016). These characteristics make EEG fundamentally different from vision data in structure and signal quality, necessitating data-efficient strategies that directly address its inherent noise and redundancy. Despite these challenges, little is known about how the volume and composition of pre-training EEG data influence the generalization ability of foundation models, leaving a critical gap in the development of scalable and effective large-scale EEG modeling.

To overcome these limitations, we propose **EEG-DLite**, a *data distillation framework* for data-efficient pre-training of large EEG foundation models. EEG-DLite systematically prunes large, unlabeled EEG datasets that are collected using varied channel montages into smaller, more representative subsets. Due to the high dimensionality and low signal-to-noise ratio of EEG signals, instead of selecting samples directly using EEG signals, our approach begins by encoding EEG segments into compact latent representations using a self-supervised autoencoder. Based on these representations, we apply two key techniques:

- A robust *outlier filtering mechanism* to remove noisy or corrupted EEG segments
- A *divergence-based redundancy reduction* method to reduce data redundancy while preserving informative diversity

These steps produce a significantly smaller dataset that still captures the diversity and structure necessary for effective model training. In our experiments, training on only **5% of a 2,500-hour EEG dataset** distilled using EEG-DLite achieves *comparable*, or *even superior*, performance on several downstream tasks compared to training on the full dataset, as shown in Figure 1. Additionally, this reduces GPU pre-training time from **30 hours to just 2 hours**, under the same hardware conditions.

To our knowledge, this is the **first study to explore data distillation for physiological signals** like EEG in the context of large foundation models. We also perform systematic comparisons of *generative* and *selection-based* distillation approaches, providing new insights into what makes pre-training data effective in EEG modeling[1].

The main contributions of this work are as follows:

- We propose *the first data distillation framework tailored for large-scale EEG foundation model pre-training*, achieving comparable or even superior performance using only 5% of the original training data.
- We present the first comparison between synthetic data generation and selection-based distillation on EEG. The results show the challenges EEG poses on the generation approach and highlight the effectiveness of a core-set selection approach.
- We systematically analyze how the quantity of pre-training data affects model generalization, providing empirical evidence for the efficiency and robustness of our distilled EEG subsets.

[1]The code is available at https://github.com/t170815518/EEG-DLite

Algorithm 1: Overview of the EEG-DLite Framework

Require: EEG dataset $\mathcal{X} = \{X_i \in \mathbb{R}^{C\times T}\}_{i=1}^{N}$, encoder ξ, decoder $\mathcal{D}$, latent representation dimension d, outlier threshold τ, distillation ratio η

Ensure: Distilled EEG subset $\mathcal{S} \subset \mathcal{X}$

1: Compute spectral views $X_i^{(m)}, X_i^{(\phi)} \leftarrow \texttt{FFT}(X_i)$ and concatenate: $X_i^{\text{all}} = [X_i, X_i^{(m)}, X_i^{(\phi)}]$

2: Train $(\xi, \mathcal{D})$ to minimize:

$$\mathcal{L} = \mathcal{L}_{\text{Rec}} + \beta\mathcal{L}_{\text{IDC}}$$

3: Encode each X_i^{all} to $z_i = \xi(X_i^{\text{all}})$, collect $\mathcal{Z} = \{z_i\}_{i=1}^{N}$

4: Compute outlier scores: $\sum_{i=1}^{d} \log \frac{1}{p_i(x_i)+\alpha}$

5: Remove top $\tau\%$ OODs: $\mathcal{Z}' \subset \mathcal{Z}$, $\mathcal{X}' \subset \mathcal{X}$

6: Select $\eta\%$ with diversity sampling:

$$\min_{\boldsymbol{\mu}\subset\mathcal{Z}'} \max_{\mathbf{z}\in\mathcal{Z}'} \min_{k\in\mathcal{K}} ||\mathbf{z} - \boldsymbol{\mu}_k||_2^2.$$

7: Return subset $\mathcal{S} \subset \mathcal{X}'$

Related Work

EEG foundation models

EEG foundation models have been an emerging paradigm for decoding and modeling brain activity using deep neural networks pretrained on large-scale EEG recordings (Zhou et al. 2025; Lai et al. 2025). Typical architectures adopt self-supervised learning (SSL) objectives and employ CNNs, transformers, or hybrid architectures to extract generalizable neural representations. These models have shown strong performance on downstream tasks such as cognitive state decoding and clinical classification (Jiang, Zhao, and Lu 2024; Yue et al. 2024). However, training them requires massive datasets, often exceeding millions of EEG segments, which poses significant computational and storage burdens. More critically, existing efforts largely overlook the influence of training data composition and selection on model performance. While prior work has focused on architectural innovation and transfer learning strategies, few have explored how data quality, representativeness, and redundancy affect the foundation model's generalization ability. Our work addresses this gap by introducing a data-centric framework that selects diverse, informative EEG segments to reduce training costs without compromising generalization. By emphasizing principled data selection, we aim to enhance both the scalability and explainability of large EEG models, bridging the methodological gap between neuroscience and AI through efficient and standardized data curation.

Data distillation

Data distillation refers to the process of constructing a compact and informative subset of training data that allows a model to achieve comparable performance to that trained

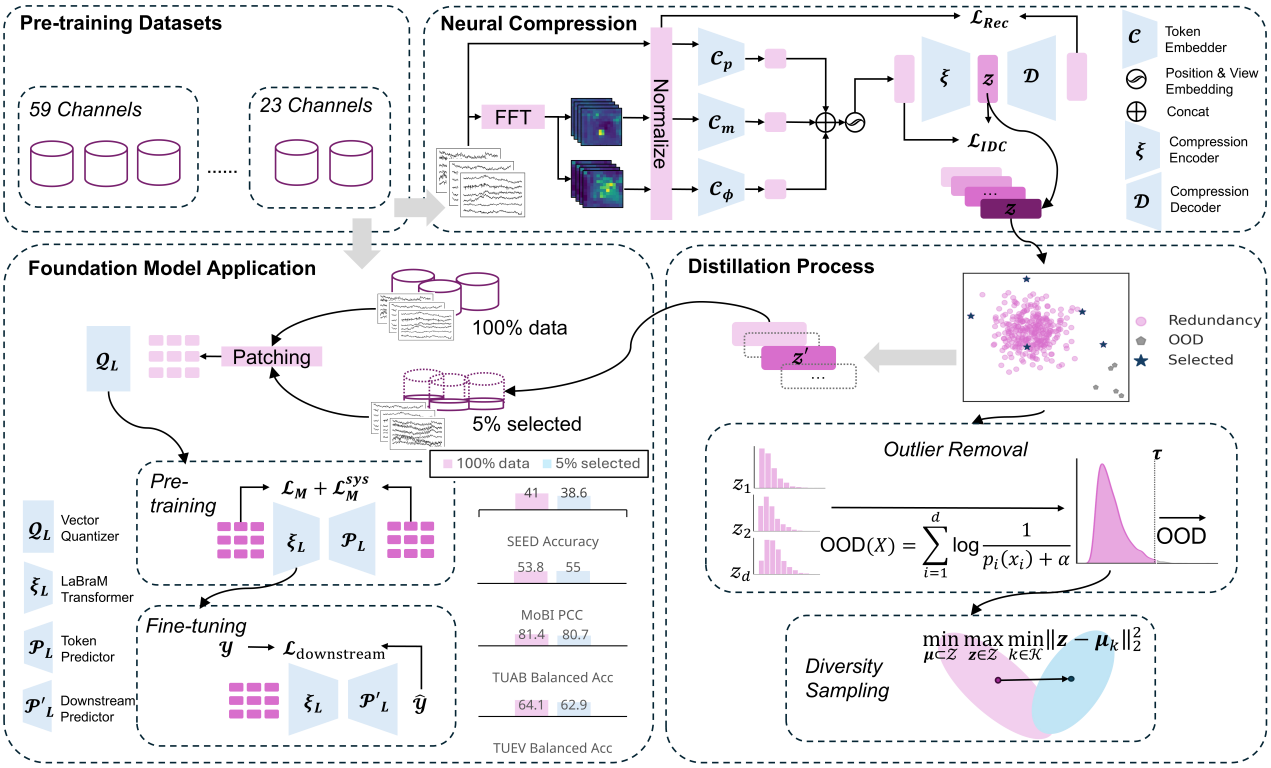

Figure 2: Overview of the proposed framework. In the first stage, an auto-encoder is trained with self-supervised learning to project each EEG segment X into a lower-dimensional latent representation $\mathbf{z}$. In the second stage, out-of-distribution scores are computed for all samples, and the top $\tau\%$ with the highest scores are excluded. Finally, $\eta\%$ of the datasets is selected via diversity sampling to retain only the most informative data.

with the full dataset, while significantly reducing computational costs. Existing distillation methods generally fall into two main categories: synthetic data generation and core-set selection (Joshi, Ni, and Mirzasoleiman 2024; Sener and Savarese 2018). Synthetic data generation focuses on generating samples that simulate the learning behavior or statistical properties of the original dataset. They have gained popularity in many computer vision tasks due to the flexibility and expressive capacity (Ding et al. 2024). These methods often use optimization objectives such as performance matching, parameter matching, or distribution matching (Yu, Liu, and Wang 2023; Zhang et al. 2024). However, they typically rely on bi-level optimization, which involves nested optimization loops, making them computationally intensive. Core-set selection, in contrast, selects a real and representative subset from the training data (Sener and Savarese 2018). They are more stable and interpretable since they work with real data, but their selection process becomes increasingly expensive as both data size and dimensionality grow. This is particularly problematic for EEG signals, which are high-dimensional, noisy, and vary across subjects. To address the computational burden, we adopt SSL to train a lightweight autoencoder that compresses EEG segments into low-dimensional latent representations, enabling efficient selection (Balestriero et al. 2023).

Methodology

This section elaborates on the details of the proposed framework, EEG-DLite, which is a model-agnostic EEG data distillation framework. It is designed to select a compact yet representative subset from massive unlabeled data for efficient pre-training. As illustrated in Figure 2 and Algorithm 1, the framework consists of three main steps: (1) a self-supervised multi-view autoencoder compresses high-dimensional EEG signals into low-dimensional latent representations; (2) an outlier detection module filters out anomalous or low-quality samples that may degrade diversity; and (3) a k-center selection algorithm identifies a diverse set of EEG segments that best represent the original dataset.

Multi-view Neural Compressor

To handle the high dimensionality of EEG signals, we train a lightweight autoencoder with SSL that maps raw EEG segments into compact latent representations without requiring labeled data (Zhao et al. 2024; Zhang et al. 2023). This architecture is intentionally decoupled from any specific foundation model to ensure flexibility and broad compatibility.

Model Architecture Each EEG segment $X \in \mathbb{R}^{C\times T}$, where C denotes the number of channels and T denotes the temporal length, is processed together with its spectral counterpart obtained via Fast Fourier Transform (FFT). Both the

raw and spectral views are partitioned into non-overlapping temporal patches of length T_W, denoted as $\boldsymbol{x}_i \in \mathbb{R}^{C \times T_W}$. These patches in the potential, magnitude, and phase domains are individually encoded by convolutional neural networks $\mathcal{C}_p$, $\mathcal{C}_m$, and $\mathcal{C}_\phi$ to extract localized spatiotemporal features and produce token embeddings. The resulting token sequences are enriched with positional and view-specific embeddings and passed to a transformer-based encoder, denoted by ξ, which captures global dependencies across the sequence. A segment-level representation $\mathbf{z}$ is then computed by averaging all token embeddings. Finally, a decoder $\mathcal{D}$, which consists of a shallow transformer block followed by multi-layer perceptron layers, reconstructs both the original signals and their spectral components from the compressed token representations.

Optimization Objective The SSL objective includes two components: a reconstruction loss $\mathcal{L}_{\text{Rec}}$ and an inter-instance discrimination loss $\mathcal{L}_{\text{IDC}}$. The reconstruction loss minimizes the mean squared error between input patches $\boldsymbol{x}_i$ with length L and reconstructions $\boldsymbol{x}'_i$, ensuring the encoder can accurately encode neural signal content. The loss of each sample X is defined as

$$\mathcal{L}_{\text{Rec}} = \sum_{i=1}^{L} (\boldsymbol{x}'_i - \boldsymbol{x}_i)^2 . \tag{1}$$

$\mathcal{L}_{\text{IDC}}$ penalizes excessive token similarity between samples in one batch, encouraging feature diversity (Chen, Lagadec, and Bremond 2021). Specifically, denoting $\mathbf{z}_i$ and $\mathbf{z}'_i$ the i-th token before and after the encoder, the network projects them into the same embedding space with two separate projectors g_1 and g_2 and penalizes the cosine similarity between an original token and the encoded tokens in other samples, as described below,

$$\mathcal{L}_{\text{IDC}} = \frac{\log \sum_{i=1}^{|B|} \sum_{j=1, i \neq j}^{|B|} \exp[\text{sim}(g_1(\mathbf{z}_i), g_2(\mathbf{z}'_j))]}{|B| \times (|B| - 1)},$$

where B represents a sample batch, $\text{sim}(\cdot)$ represents the cosine similarity. Therefore, the final optimization objective is defined as

$$\mathcal{L} = \mathcal{L}_{\text{Rec}} + \beta \cdot \mathcal{L}_{\text{IDC}}. \tag{2}$$

Outlier Sample Removal

At this stage, the goal is to remove isolated and unrepresentative samples within the dataset. The operation is conducted in the compressed representation space to ensure efficiency. Specifically, we adopt the Histogram-Based Outlier Score (HBOS) method, which identifies anomalies based on their probabilistic rarity in each feature dimension d.

Each sample X can be assigned with an out-of-distribution (OOD) score, defined as Equation 3. The top τ percent of samples are excluded from the subsequent diverse selection step to prevent degrading the diversity and quality of the distilled subset.

$$\text{OOD}(X) = \sum_{i=1}^{d} \log \frac{1}{p_i(x_i) + \alpha} \tag{3}$$

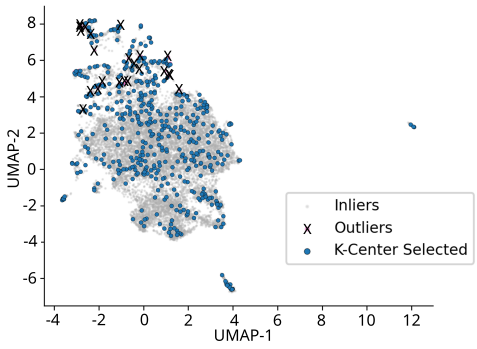

Figure 3: UMAP visualization of compressed EEG segments after OOD removal and diversity sampling on SEED-sleep (Li et al. 2025).

Diversity Sampling

The proposed framework utilizes the selection-based distillation method. We formalize the process of selecting the most representative EEG samples as the coreset construction problem. Formally, let $\mathcal{Z} \subset \mathbb{R}^{N \times d}$ be the EEG compressed pre-training dataset, where N is the sample size, d is the dimension of compressed representations. The objective is to select k points from $\mathcal{Z}$ such that every data point $\mathbf{z}$ is closest to one selected center, and the largest distance between any point to its closest center is minimized, which is described as

$$\min_{\boldsymbol{\mu} \subset \mathcal{Z}} \max_{\mathbf{z} \in \mathcal{Z}} \min_{k \in \mathcal{K}} ||\mathbf{z} - \boldsymbol{\mu}_k||_2^2. \tag{4}$$

This problem is NP-hard, so we utilize greedy approximation solver proposed in (Sener and Savarese 2018) to iteratively select points that maximize the minimum distance to previously chosen k-centers, of which the time complexity is $\mathcal{O}(k \times N \times d)$.

Experiments & Results

This section elaborates on the baselines, implementation details, the setup during pre-training and fine-tuning stages, followed by the results and in-depth analysis.

Baselines

We compare the proposed framework with three baselines under varying distillation ratios.

Random baseline uniformly samples data without considering structure, serving as a simple, unbiased lower-bound baseline.

PCA (Anuragi, Sisodia, and Pachori 2024; Fujiwara et al. 2020) reduces EEG dimensionality by projecting data onto top principal components. We use Incremental PCA to scale efficiently and assess how it compares with SSL-based embeddings in representing EEG diversity.

M3D (Zhang et al. 2024) generates synthetic data by minimizing embedding discrepancies using random, untrained networks. Although it is lightweight and computationally feasible at low distillation ratios, it still becomes impractical

Method	η (%)	SEED-V			MoBI		
		Accuracy (%)	κ (%)	F1 (%)	PCC	R^2	RMSE ($\times 10^2$)
Random	1	26.5 ± 0.2	7.4 ± 0.3	26.6 ± 0.2	0.468 ± 0.009	0.180 ± 0.007	13.97 ± 0.08
	5	34.6 ± 0.3	18.3 ± 0.4	34.9 ± 0.3	0.530 ± 0.015	0.260 ± 0.007	13.26 ± 0.06
	10	37.9 ± 0.3	22.4 ± 0.4	38.3 ± 0.3	0.540 ± 0.003	0.267 ± 0.004	13.22 ± 0.03
	25	39.1 ± 0.2	23.9 ± 0.3	39.4 ± 0.2	0.538 ± 0.003	0.263 ± 0.005	13.29 ± 0.05
M3D	1	23.4 ± 0.3	3.2 ± 0.4	22.7 ± 0.4	0.305 ± 0.003	0.016 ± 0.004	15.06 ± 0.04
	5	26.9 ± 0.1	7.9 ± 0.1	26.9 ± 0.1	0.465 ± 0.004	0.171 ± 0.004	14.04 ± 0.03
	10	24.5 ± 0.2	4.9 ± 0.3	24.3 ± 0.2	0.500 ± 0.009	0.209 ± 0.011	13.74 ± 0.08
	25	-	-	-	-	-	-
PCA + DS	1	22.6 ± 0.4	2.2 ± 0.5	22.2 ± 0.4	0.356 ± 0.004	0.046 ± 0.005	14.90 ± 0.03
	5	31.0 ± 0.4	13.4 ± 0.5	31.2 ± 0.4	0.534 ± 0.004	0.262 ± 0.000	13.32 ± 0.06
	10	36.3 ± 0.3	20.1 ± 0.4	36.5 ± 0.3	$\mathbf{0.541 \pm 0.002}$	0.265 ± 0.003	13.26 ± 0.02
	25	38.6 ± 0.2	23.5 ± 0.3	39.0 ± 0.2	$0.546 \pm 0.001^{\dagger}$	$0.276 \pm 0.004^{\dagger}$	$13.15 \pm 0.03^{\dagger}$
Proposed	1	$\mathbf{29.7 \pm 0.3}^{\dagger}$	$\mathbf{11.3 \pm 0.3}^{\dagger}$	$\mathbf{29.7 \pm 0.3}^{\dagger}$	$\mathbf{0.506 \pm 0.007}^{\dagger}$	$\mathbf{0.225 \pm 0.010}^{\dagger}$	$\mathbf{13.61 \pm 0.10}^{\dagger}$
	5	$\mathbf{38.6 \pm 0.2}^{\dagger}$	$\mathbf{23.1 \pm 0.2}^{\dagger}$	$\mathbf{38.9 \pm 0.2}^{\dagger}$	$\mathbf{0.550 \pm 0.001}^{\dagger}$	$\mathbf{0.283 \pm 0.001}^{\dagger}$	$\mathbf{13.15 \pm 0.07}^{\dagger}$
	10	$\mathbf{39.1 \pm 0.1}^{\dagger}$	$\mathbf{23.9 \pm 0.2}^{\dagger}$	$\mathbf{39.5 \pm 0.2}^{\dagger}$	$\mathbf{0.541 \pm 0.002}$	$\mathbf{0.277 \pm 0.014}$	$\mathbf{13.07 \pm 0.04}^{\dagger}$
	25	$\mathbf{39.7 \pm 0.2}^{\dagger}$	$\mathbf{24.7 \pm 0.2}^{\dagger}$	$\mathbf{40.1 \pm 0.2}^{\dagger}$	$\mathbf{0.550 \pm 0.003}^{\dagger}$	$\mathbf{0.281 \pm 0.005}^{\dagger}$	$\mathbf{13.14 \pm 0.06}^{\dagger}$
Full data	100	41.0 ± 0.6	26.1 ± 0.8	41.2 ± 0.6	0.538 ± 0.010	0.288 ± 0.003	12.25 ± 0.03

Table 1: Performance comparison across different distillation ratios (η) on SEED-V and MoBI datasets. κ refers to Cohen's kappa coefficient. Bold values indicate the best performance within each distillation ratio, excluding the full data case. DS refers to the diversity sampling. The † symbol indicates the result is significantly better ($p < 0.05$) than the corresponding Random baseline at the same η based on Mann–Whitney U test.

at 25 percent distillation ratio due to excessive GPU memory demands.

Implementation Details

The input signals from both the potential and spectral domains are independently normalized and segmented into 20 non-overlapping patches. Then, the neural compressor is trained for 50 epochs using the Adam optimizer, with the learning rate of 0.001 and gradient clipping applied at a maximum norm of 5.0. A scheduler is used to reduce the learning rate every 10 epochs with a decay factor of 0.5. The training objective, described in Equation 2, incorporates an instance discrimination loss term $\mathcal{L}_{\text{IDC}}$, weighted by $\beta = 0.0001$. The encoder consists of 6 self-attention layers with 8 heads per layer, while the decoder contains 2 transformer layers, each with 8 attention-heads as well. Each EEG segment X is projected into a 64-dimensional embedding space.

Experiment Setup

Dataset Construction We use the same pre-training datasets as in LaBraM (Jiang, Zhao, and Lu 2024), which includes 32 datasets over 2,500 hours of EEG data collected from multiple tasks (e.g., emotion recognition, motor imagery, etc.). These datasets have diverse channel montages. Each EEG sample is a segment into $X \in \mathbb{R}^{c \times T}$, where c is the number of channels and T is the temporal length (either 4 or 8 seconds), with a fixed stride of 4 seconds. All data are band-pass filtered within the range [0.1, 75] Hz and resampled to 200Hz. During the distillation process, each dataset is distilled independently.

Pre-training Configuration We adopt the LaBraM-base architecture, consisting of 12 encoder layers with a hidden size of 200, 10 attention heads, and a multi-layer-perceptron dimension of 800. The model is trained using the AdamW optimizer with $\beta_1 = 0.9, \beta_2 = 0.98$, and a cosine learning rate scheduler with the weight decay of 0.05. The pre-training objective follows the symmetric masking formulation $\mathcal{L} = \mathcal{L}_M + \mathcal{L}_M^{\text{sys}}$. All the pre-training and fine-tuning experiments are conducted on four NVIDIA RTX 4090 GPUs.

Downstream Evaluation To assess the effectiveness of each distillation method, we evaluate the pretrained foundation models on diverse types of downstream tasks. To perform rigorous comparison, each method distills the origi-

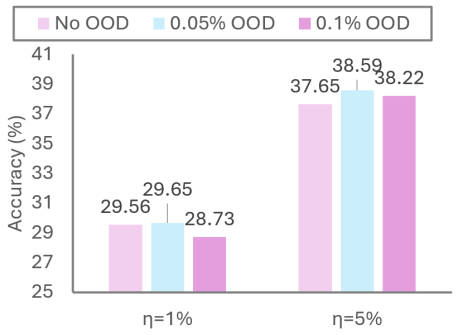

Figure 4: Impact of outlier removal on SEED-V accuracy at different distillation ratios η and OOD removal thresholds τ.

Method	η (%)	TUEV			TUAB		
		Balanced Acc (%)	κ (%)	F1 (%)	Balanced Acc (%)	PR AUC (%)	AUROC (%)
Random	1	52.7 ± 0.9	46.8 ± 0.9	73.6 ± 0.6	79.5 ± 0.1	87.7 ± 0.1	88.2 ± 0.1
	5	62.3 ± 0.9	58.1 ± 0.9	79.3 ± 0.6	$\mathbf{80.7 \pm 0.2}$	89.9 ± 0.2	90.0 ± 0.1
	10	58.9 ± 0.5	56.7 ± 1.5	78.9 ± 0.7	81.2 ± 0.1	90.4 ± 0.2	90.6 ± 0.2
	25	64.4 ± 1.2	61.5 ± 1.5	81.2 ± 0.7	81.1 ± 0.2	90.2 ± 0.1	90.3 ± 0.1
M3D	1	45.8 ± 0.4	37.6 ± 0.6	68.1 ± 0.4	77.3 ± 0.1	83.4 ± 0.2	85.2 ± 0.1
	5	42.9 ± 0.8	42.4 ± 0.8	71.2 ± 0.4	79.5 ± 0.1	86.2 ± 0.1	88.1 ± 0.1
	10	42.9 ± 0.9	38.9 ± 1.1	69.1 ± 0.7	80.1 ± 0.1	86.9 ± 0.3	88.3 ± 0.2
	25	-	-	-	-	-	-
PCA + DS	1	$54.4 \pm 0.8^{\dagger}$	$48.0 \pm 0.5^{\dagger}$	73.9 ± 0.3	77.8 ± 0.1	85.0 ± 0.2	86.2 ± 0.1
	5	55.8 ± 1.1	51.0 ± 1.0	75.7 ± 0.6	$81.3 \pm 0.1^{\dagger}$	$\mathbf{89.9 \pm 0.1}$	$\mathbf{90.6 \pm 0.1}^{\dagger}$
	10	$63.9 \pm 1.2^{\dagger}$	$61.6 \pm 1.9^{\dagger}$	$81.2 \pm 1.0\dagger$	81.0 ± 0.1	90.1 ± 0.1	90.2 ± 0.1
	25	56.8 ± 1.9	52.1 ± 1.6	76.8 ± 0.6	$\mathbf{81.5 \pm 0.4}^{\dagger}$	$\mathbf{90.3 \pm 0.3}$	$\mathbf{90.4 \pm 0.3}$
Proposed	1	$\mathbf{57.3 \pm 0.6}^{\dagger}$	$\mathbf{52.6 \pm 0.3}^{\dagger}$	$\mathbf{76.7 \pm 0.4}^{\dagger}$	$\mathbf{80.0 \pm 0.3}^{\dagger}$	$\mathbf{87.9 \pm 0.1}^{\dagger}$	$\mathbf{88.7 \pm 0.1}^{\dagger}$
	5	$\mathbf{62.9 \pm 1.0}$	$\mathbf{61.0 \pm 0.9}^{\dagger}$	$\mathbf{80.7 \pm 0.6}^{\dagger}$	$\mathbf{80.7 \pm 0.0}$	89.5 ± 0.1	$90.3 \pm 0.0^{\dagger}$
	10	$\mathbf{64.3 \pm 1.5}^{\dagger}$	$\mathbf{63.0 \pm 0.3}^{\dagger}$	$\mathbf{82.2 \pm 0.1}^{\dagger}$	$\mathbf{81.5 \pm 0.1}^{\dagger}$	$\mathbf{90.6 \pm 0.1}^{\dagger}$	$\mathbf{90.8 \pm 0.1}^{\dagger}$
	25	$\mathbf{64.5 \pm 1.3}$	$\mathbf{63.4 \pm 1.2}$	$\mathbf{82.1 \pm 0.5}^{\dagger}$	81.3 ± 0.2	90.1 ± 0.2	90.3 ± 0.1
Full data	100	64.1 ± 0.7	66.4 ± 1.0	83.1 ± 0.5	81.4 ± 0.2	89.7 ± 0.2	90.2 ± 0.1

Table 2: Performance comparison across distillation ratios (η) on TUEV and TUAB datasets. $\boldsymbol{\kappa}$ refers to Cohen's kappa coefficient. Bold values indicate the best performance within each distillation ratio, excluding the full data case. DS refers to the diversity sampling. The † symbol indicates the result is significantly better ($p < 0.05$) than the corresponding Random baseline at the same η based on Mann–Whitney U test.

nal pre-training dataset at multiple compression ratios ($\eta = 1\%, 5\%, 10\%, 25\%$) first, and the resulting distilled subsets are used for pre-training. Finally, the pretrained models are fine-tuned separately on four representative downstream tasks that cover both classification and regression settings. We follow the identical evaluation protocol established in LaBraM to ensure direct comparability. All experiments use consistent hyperparameters during pre-training and fine-tuning. Results are averaged across five random seeds, and standard deviations and results' significance are reported in Table 1 and 2. Below are the details of downstream tasks:

- **TUEV** (Obeid and Picone 2016): Six-class classification of EEG events.
- **TUAB** (Obeid and Picone 2016): Binary classification of normal versus abnormal EEG signals.
- **SEED-V** (Liu et al. 2022): Five-class emotion recognition based on EEG.
- **MoBI** (He et al. 2018): Continuous regression of bilateral lower-limb joint angles from EEG during walking.

Distillation Performance

The foundation model retains comparable performance with 5% of the pre-training data. As the dataset percentage increases, the performance on all the downstream tasks demonstrates improvements, indicating enhanced model generalization with more training data, and the performance change is also aligned with the experiments reported in (Jiang, Zhao, and Lu 2024). Tables 1 and 2 indicate a consistent pattern of data redundancy in large-scale EEG pre-training datasets. Specifically, we observe that using only 5% of the pre-training data selected by EEG-DLite achieves performance close to the upper bound. In contrast, the random selection baseline requires around 25% of the data to reach a similar level. These results suggest that a substantial portion of the data has limited impact on model performance, highlighting the potential for more efficient training and deployment through targeted data selection.

EEG-DLite consistently outperforms across all downstream datasets and distillation ratios. Across all evaluated datasets and ratios, the proposed framework consistently outperforms the random sampling and M3D baselines, demonstrating its effectiveness in selecting informative and diverse subsets. These results highlight the importance of maintaining diversity during distillation to enhance generalization and model robustness, even with limited data. Furthermore, we find that SSL generates more stable and discriminative representations than PCA under the same output dimensionality. On datasets such as TUEV, MoBI, and TUAB, models trained on distilled data even surpass those trained on the full dataset, indicating that careful sample selection could be more effective than brute-force scaling. Although the EEG segments are unlabeled, the UMAP visualization in Figure 3 reveals distinct clustering patterns in the latent space, reflecting the inherent structure and diversity present in EEG signals as well.

OOD removal is helpful. Figure 4 presents an ablation study assessing the effect of varying the outlier removal

Method	η (%)	Acc.	F1	κ
Random	50	53.4	54.0	28.9
	25	52.8	52.1	28.0
PCA + DS	50	51.8	48.6	27.4
	25	51.6	49.9	26.2
Proposed ($\tau = 0$)	50	54.1	52.8	29.9
	25	54.6	55.1	29.1
Proposed ($\tau = 1\%$)	50	**56.6**	**56.7**	**33.2**
	25	**55.3**	**55.7**	**31.3**
Full Data	100	54.6	55.4	30.8

Table 3: Pilot study evaluating EEG-DLite on SEED using EEGNet in the cross-subject supervised setting at different distillation ratios η and OOD removal ratios τ.

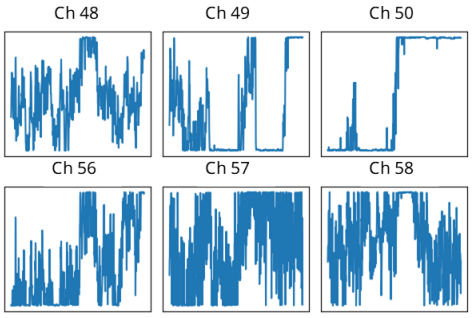

Figure 5: Example of generative EEG sample with M3D.

threshold τ. The results show that removing the top 0.05% samples with the highest OOD scores leads to improved downstream accuracy. Manual inspection reveals that these samples often contain noise or signal artifacts, which can degrade representation quality.

To further evaluate the effectiveness of EEG-DLite, we conduct a pilot study on SEED dataset (Zheng and Lu 2015a; Duan, Zhu, and Lu 2013) using a small network like EEGNet in the cross-subject supervised learning setting. The results in Table 3 show that EEG-DLite achieves the highest performance when trained on only 25% of the data, even outperforming models trained on the full dataset. We also observe the model is sensitive to the presence of OOD samples in the supervised setting.

Large-scale EEG datasets pose new challenges to the data synthesis approach. Although generative approaches have shown effectiveness in computer vision tasks (Zhang et al. 2024; Zhao and Bilen 2022), their effectiveness on physiological signals such as EEG remains largely unexplored. In our experiments, M3D, a lightweighted synthesizing framework, was selected considering the huge computational effort to generate large pre-training datasets. M3D consistently underperforms compared to all other baselines, including random sampling. As illustrated in Figure 5, the generated EEG segments exhibit unnatural characteristics, such as abrupt plateaus, flat transitions, and repetitive blocky patterns, all of which deviate from realistic

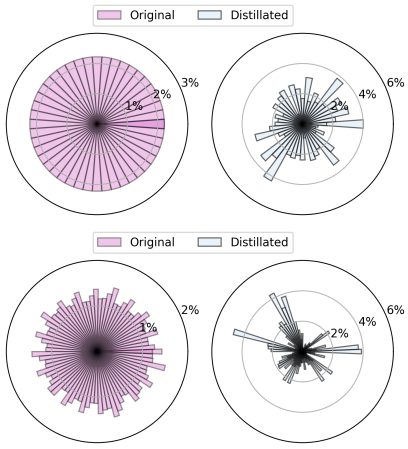

Figure 6: Sample distribution per subject in SEED (Zheng and Lu 2015b) (top) and SEED-VII (Jiang et al. 2025) (bottom). Each bin represents a subject, and its height indicates the proportion of samples contributed by that subject in the original or 5% distilled dataset.

brain signal dynamics. These artifacts reflect poor temporal and spectral fidelity. Furthermore, the generative process in M3D is computationally intensive, which poses scalability challenges for large-scale applications. Taken together, these findings highlight that for EEG and related physiological signals, both the synthetic quality and computational feasibility are important considerations.

Subject variance becomes significant after diversity sampling. BCI datasets typically involve recordings from multiple subjects, each contributing EEG segments with varying quality and characteristics. After applying diversity sampling, the percentage of samples per subject shows notable variation. As illustrated in Figure 6, some subjects contribute significantly more segments than others. This imbalance reflects inherent inter-subject variability in EEG signals, which can stem from differences in neural dynamics, recording conditions, or noise levels. Recognizing and understanding these subject-level patterns may inform several directions for future research. For instance, it may enable subject-aware pre-training strategies.

Conclusion

In this study, we propose a novel data distillation framework to condense large-scale unlabeled EEG datasets through three key steps: compression, outlier removal, and diversity sampling. The framework is evaluated across four downstream tasks at varying distillation ratios, demonstrating that only 5% of the original data is sufficient to train a foundation model with comparable performance. These results highlight that data quality and diversity contribute more significantly to the generalization ability of foundation models than data quantity. We believe EEG-DLite represents a practical step toward efficient foundation model pre-training and provides insights about diverse dataset design and selection for broader downstream applications.

Acknowledgments

This work is supported by the MOE Tier 2 Project (MOE-T2EP20124-0001).

References

Amin, H. U.; Malik, A. S.; Kamel, N.; and Hussain, M. 2016. A novel approach based on data redundancy for feature extraction of EEG signals. *Brain topography*, 29(2): 207–217.

Anuragi, A.; Sisodia, D. S.; and Pachori, R. B. 2024. Mitigating the curse of dimensionality using feature projection techniques on electroencephalography datasets: an empirical review. *Artificial Intelligence Review*, 57(3).

Balestriero, R.; Ibrahim, M.; Sobal, V.; Morcos, A.; Shekhar, S.; Goldstein, T.; Bordes, F.; Bardes, A.; Mialon, G.; Tian, Y.; Schwarzschild, A.; Wilson, A. G.; Geiping, J.; Garrido, Q.; Fernandez, P.; Bar, A.; Pirsiavash, H.; LeCun, Y.; and Goldblum, M. 2023. A Cookbook of Self-Supervised Learning. arXiv:2304.12210.

Carzaniga, F. S.; Hoppeler, G. T.; Hersche, M.; Schindler, K.; and Rahimi, A. 2025. The Case for Cleaner Biosignals: High-fidelity Neural Compressor Enables Transfer from Cleaner iEEG to Noisier EEG. In *The Thirteenth International Conference on Learning Representations*.

Chen, H.; Lagadec, B.; and Bremond, F. 2021. ICE: Inter-instance Contrastive Encoding for Unsupervised Person Re-identification. arXiv:2103.16364.

Ding, J.; Liu, Z.; Zheng, G.; Jin, H.; and Kong, L. 2024. CondTSF: One-line Plugin of Dataset Condensation for Time Series Forecasting. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Duan, R.-N.; Zhu, J.-Y.; and Lu, B.-L. 2013. Differential entropy feature for EEG-based emotion classification. In *6th International IEEE/EMBS Conference on Neural Engineering (NER)*, 81–84. IEEE.

Fujiwara, T.; Chou, J.-K.; Shilpika; Xu, P.; Ren, L.; and Ma, K.-L. 2020. An Incremental Dimensionality Reduction Method for Visualizing Streaming Multidimensional Data. *IEEE Transactions on Visualization and Computer Graphics*, 26(1): 418–428.

He, Y.; Luu, T. P.; Nathan, K.; Nakagome, S.; and Contreras-Vidal, J. L. 2018. A mobile brain-body imaging dataset recorded during treadmill walking with a brain-computer interface. *Scientific Data*, 5(1).

Jiang, W.-B.; Liu, X.-H.; Zheng, W.-L.; and Lu, B.-L. 2025. SEED-VII: A Multimodal Dataset of Six Basic Emotions With Continuous Labels for Emotion Recognition. *IEEE Transactions on Affective Computing*, 16(2): 969–985.

Jiang, W.-B.; Zhao, L.-M.; and Lu, B.-L. 2024. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI. arXiv:2405.18765.

Joshi, S.; Ni, J.; and Mirzasoleiman, B. 2024. Dataset Distillation via Knowledge Distillation: Towards Efficient Self-Supervised Pre-Training of Deep Networks. arXiv:2410.02116.

Killamsetty, K.; Sivasubramanian, D.; Ramakrishnan, G.; De, A.; and Iyer, R. 2021. GRAD-MATCH: Gradient Matching based Data Subset Selection for Efficient Deep Model Training. arXiv:2103.00123.

Lai, J.; Wei, J.; Yao, L.; and Wang, Y. 2025. A Simple Review of EEG Foundation Models: Datasets, Advancements and Future Perspectives. arXiv:2504.20069.

Li, Z.; Tao, L.-Y.; Ma, R.-X.; Zheng, W.-L.; and Lu, B.-L. 2025. Investigating the Effects of Sleep Conditions on Emotion Responses with EEG Signals and Eye Movements. *IEEE Transactions on Affective Computing*.

Liu, W.; Qiu, J.-L.; Zheng, W.-L.; and Lu, B.-L. 2022. Comparing Recognition Performance and Robustness of Multimodal Deep Learning Models for Multimodal Emotion Recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2): 715–729.

Obeid, I.; and Picone, J. 2016. The Temple University Hospital EEG Data Corpus. *Frontiers in Neuroscience*, Volume 10 - 2016.

Sadiya, S.; Alhanai, T.; and Ghassemi, M. M. 2021. Artifact Detection and Correction in EEG data: A Review. In *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*, 495–498.

Sener, O.; and Savarese, S. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *International Conference on Learning Representations*.

Tang, Y.; Robinson, N.; Fu, X.; Thomas, K. P.; Wai, A. A. P.; and Guan, C. 2024. Reconstruction of Continuous Hand Grasp Movement from EEG Using Deep Learning. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 1–4. IEEE.

Wang, J.; Zhao, S.; Luo, Z.; Zhou, Y.; Jiang, H.; Li, S.; Li, T.; and Pan, G. 2025. CBraMod: A Criss-Cross Brain Foundation Model for EEG Decoding. In *The Thirteenth International Conference on Learning Representations*.

Wang, P.; Cao, X.; Zhou, Y.; Gong, P.; Yousefnezhad, M.; Shao, W.; and Zhang, D. 2023. A comprehensive review on motion trajectory reconstruction for EEG-based brain-computer interface. *Frontiers in Neuroscience*, 17.

Wang-Nöth, L.; Heiler, P.; Huang, H.; Lichtenstern, D.; Reichenbach, A.; Flacke, L.; Maisch, L.; and Mayer, H. 2025. How much data is enough? Optimization of data collection for artifact detection in EEG recordings. *Journal of Neural Engineering*, 22(2): 026026.

Yu, R.; Liu, S.; and Wang, X. 2023. Dataset Distillation: A Comprehensive Review. arXiv:2301.07014.

Yue, T.; Xue, S.; Gao, X.; Tang, Y.; Guo, L.; Jiang, J.; and Liu, J. 2024. EEGPT: Unleashing the Potential of EEG Generalist Foundation Model by Autoregressive Pre-training. arXiv:2410.19779.

Zhang, H.; Li, S.; Wang, P.; Zeng, D.; and Ge, S. 2024. M3D: Dataset Condensation by Minimizing Maximum Mean Discrepancy. In *The 38th Annual AAAI Conference on Artificial Intelligence (AAAI)*.

Zhang, W.; Yang, L.; Geng, S.; and Hong, S. 2023. Self-Supervised Time Series Representation Learning via Cross Reconstruction Transformer. arXiv:2205.09928.

Zhao, B.; and Bilen, H. 2022. Dataset Condensation with Distribution Matching. arXiv:2110.04181.

Zhao, T.; Cui, Y.; Ji, T.; Luo, J.; Li, W.; Jiang, J.; Gao, Z.; Hu, W.; Yan, Y.; Jiang, Y.; and Hong, B. 2024. VAEEG: Variational auto-encoder for extracting EEG representation. *NeuroImage*, 304: 120946.

Zheng, W.-L.; and Lu, B.-L. 2015a. 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.

Zheng, W.-L.; and Lu, B.-L. 2015b. Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks. *IEEE Transactions on Autonomous Mental Development*, 7: 1–1.

Zhou, X.; Liu, C.; Chen, Z.; Wang, K.; Ding, Y.; Jia, Z.; and Wen, Q. 2025. Brain Foundation Models: A Survey on Advancements in Neural Signal Processing and Brain Discovery. arXiv:2503.00580.