BrainMoE: Cognition Joint Embedding via Mixture-of-Expert Towards Robust Brain Foundation Model

Ziquan Wei **Tingting Dan** **Tianlong Chen** **Guorong Wu***
Departments of Computer Science and Psychiatry
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599
{ziquanw,tianlong}@cs.unc.edu;{Tingting_Dan,grwu}@med.unc.edu

Abstract

Given the large scale of public functional Magnetic Resonance Imaging (fMRI), e.g., UK Biobank (UKB) and Human Connectome Projects (HCP), brain foundation models are emerging. Although the amount of samples under rich environmental variables is unprecedented, existing brain foundation models learn from fMRI derived from a narrow range of cognitive states stimulated by similar environments, causing the limited robustness demonstrated in various applications and datasets acquired with different pipelines and limited sample size. By capitalizing on the variety of cognitive status as subjects performing explicit tasks, we present the mixture of brain experts, namely BrainMoE, pre-training on tasking fMRI with rich behavioral tasks in addition to resting fMRI for a robust brain foundation model. Brain experts are designed to produce embeddings for different behavioral tasks related to cognition. Afterward, these cognition embeddings are mixed by a cognition adapter via cross-attention so that BrainMoE can handle orthogonal embeddings and be robust on those boutique downstream datasets. We have pre-trained two existing self-regressive architectures and one new supervised architecture as brain experts on 68,251 fMRI scans among UKB and HCP, containing 12 different cognitive states. Then, BrainMoE is evaluated on a variety of applications, including sex, age prediction, human behavior recognition, disease early diagnosis of Autism, Parkinson's disease, Alzheimer's disease, and Schizophrenia, and fMRI-EEG multimodal applications, where promising results in eight datasets from three different pipelines indicate great potential to facilitate current neuroimaging applications in clinical routines.

1 Introduction

Like foundation models for other topics, brain foundation models aim to learn feature representation fundamentally from large-scale data of neuroimaging. Functional Magnetic Resonance Imaging (fMRI) of the brain, as it offers insight into the relationship between functional fluctuations and human behavior [1], is critical to discovering the enigma of human cognition and promoting clinical applications. Blood-Oxygen-Level Dependent (BOLD) signal in fMRI measures neuronal activity. Such raw signals are preprocessed as timeseries of regional mean or the functional connectivity (FC) by coefficient correlation for analysis with high Signal-to-Noise Ratio (SNR) [6]. While the exploration of brain foundation model has expanded to various masking strategies for either (latent) BOLD or FC reconstruction [23, 11, 34], previous works formulating this problem as transferring

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

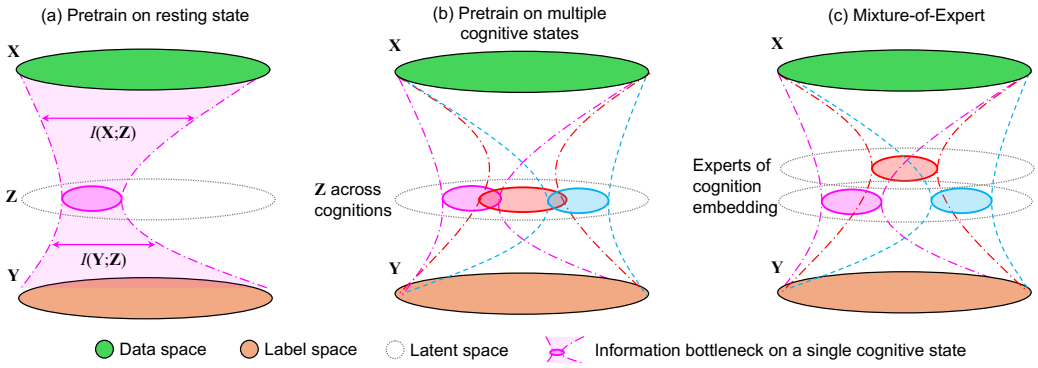

Figure 1: Motivation of BrainMoE through the lens of the information bottleneck theory. (a) Feature representation learning makes an information bottleneck between data and label space, where $\mathbf{X}$ denotes data, $\mathbf{Z}$ denotes latent feature representation, $\mathbf{Y}$ is the label of data, and $I(\cdot;\cdot)$ is the mutual information. (b) A model pre-trained on multiple cognitive states may compromise the underlying heterogeneity between different states, where $\mathbf{Z}$ cannot be optimal for all states. (c) The mixture of brain experts dedicated to diverse cognitive states leads to a joint cognition embedding so that the downstream applications can be advanced by stratified pre-training on rich behavioral tasks.

self-regressive methodologies from natural language and image to neuroimaging ignored the inter-correlation between non-imaging phenotypes [14]. Furthermore, they are restricted by a narrow range of one or two cognitive states, e.g., the resting state, causing samples with behaviors other than resting to be overlooked. Due to the lack of explicit designs to utilize the complete brain fMRI dataset with respect to neuroscience knowledge, a mixture of brain experts is proposed towards a robust brain foundation model in this work.

UK Biobank (UKB) [20] and two Human Connectome Projects (HCP) [29, 5] that contain healthy subjects 22 to 100 years of age are mainly used as pre-training datasets since they have a large scale. Previously, most subjects in resting state among UKB and HCP were included in BrainMass [34] and BrainJEPA [11]. While BrainLM [23] involved an additional tasking state in UKB, its performance has shown worse than others. Even though BrainMass has collected the most available published resting fMRI data on the OpenNeuro platform [22], ten available cognitive states in HCP datasets were ignored. It is intuitive to train a single model with all available data. However, as shown in Fig. 1 (a) and (b), simply pre-training with all cognitive states results in a single model being suboptimal to samples with different cognitive states, e.g., the red information bottleneck established by mutual information is suboptimal in the latent space where cognition related behaviors are variable. In fact, a single model is observed to compromise the underlying heterogeneity between cognitive states derived from diverse neural circuits stimulated by different behaviors [25]. This issue necessitates mixing experts specialized in different cognitive states, as shown in Fig. 1 (c), where each expert produces the cognition embedding, that is, a feature representation stratified by cognitive states.

On the other hand, tremendous efforts have been made to benchmark generally purposed models [8, 24, 10] and brain-dedicated architectures [19, 17, 4, 31] on brain fMRI data. A common observation on the results is that performance is diverse using BOLD or FC as the model input for different datasets, leading to related brain fMRI analysis works being categorized by types of input: (1) BOLD foundation models [23, 11] and (2) FC foundation models [15, 34]. Nevertheless, this reduces the adaptability of previous brain foundation models for datasets that fit better with a type of input differentiated from the pre-training stage. Mixture-of-experts (MoE) cooperating with router and adapter [35, 37] has demonstrated great potential for multimodal, referring to BOLD and FC in our data, and multitask, referring to multiple cognitive states. Therefore, a novel cognition adapter is proposed to facilitate BrainMoE as a robust brain foundation model learning from various cognitive states. Although adapters in MoE for language and vision fields are using small architecture like a multilayer perceptron (MLP) [18], a high scalability of the cognition adapter can ensure the transformation from cognition embeddings to objectives. Given that MLP is not scalable (see Appendix), a Transformer decoder is utilized for adapting BrainMoE to downstream applications.

To this end, this work has three contributions: (1) We propose BrainMoE, an MoE framework for brain fMRI data that towards high robustness for downstream tasks with different pipelines and limited sample size. (2) A cognition adapter is designed to adapt embeddings from experts pre-trained

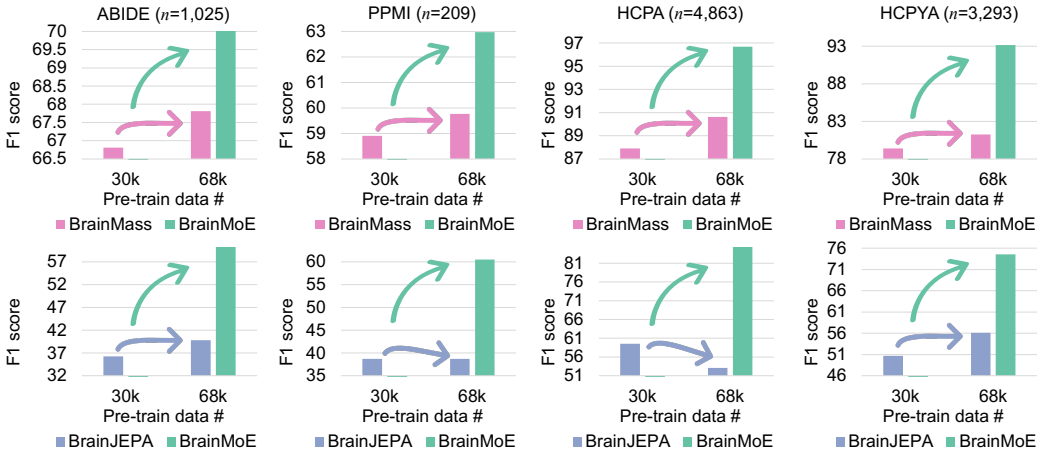

Figure 2: Increasing the scale of pre-training data of a brain foundation model leads to a marginal or negative performance boost due to 11 overlooked cognitive states. Four columns are four downstream applications, where ABIDE and PPMI are disease recognition, and HCPA and HCPYA are human behavior recognition. Two rows are two expert architectures, where BrainMass uses FC and BrainJEPA uses BOLD.

with various cognitive states, regardless of the input type, for finetuning. (3) Two existing brain foundation models pre-trained in self-supervised manners and one cognition classifier pre-trained in a supervised manner are both evaluated as experts in BrainMoE, where experts pre-train on 68,251 fMRI scans among UKB and HCPs and fine-tune on various applications, including sex prediction, human behavior recognition, and disease early diagnosis of Autism, Parkinson's disease, Alzheimer's disease, and Schizophrenia among seven datasets.

2 Preliminaries

Brain foundation models To the best of our knowledge, BrainLM [23] represents the first brain foundation model. It applied Masked Autoencoding (MAE) to BOLD signals reconstruction. However, densely filling the entire fMRI time series can impair the model's capacity to differentiate between noise and meaningful signals. Prior work [3] has demonstrated that masked pretraining in generative frameworks such as MAE often yields suboptimal results in off-the-shelf evaluations, such as linear probing. Similarly, BrainJEPA [11] introduces an alternative architecture employing a distinct JEPA-based masking strategy, addressing BrainLM's limitations by drawing on insights from I-JEPA [3]. Although BrainJEPA reports superior performance relative to linear probing, it does not explicitly incorporate pre-training with tasking fMRI. BrainMass [34] used a larger pre-training dataset (see Appendix) and a matching objective between pseudo FC matrices as a novel framework. Whilst, it used sololy the resting fMRI and overlooked more than 38k tasking fMRI in the dataset.

Resting- and tasking-state fMRI Neuroimaging data contains brain activity reflecting the interaction between functional fluctuations and human behavior. Studies controlled subjects in a resting or explicit tasking state during the data acquisition to offer distinct, complementary perspectives on brain function [36]. Large-scale studies [29, 5, 20] are observed to collect at least one tasking state in addition to the resting state, while brain foundation models mainly pre-train on the portion under the resting state, overlooking half or more data in the datasets. The intuitive method of mixing all data together may compromise the underlying heterogeneity between different states. Fig. 2 shows that the improvement gained from pre-training with 38k more data containing 11 overlooked cognitive states is marginal on four downstream classifications: ABIDE is 2-class classification for Autism diagnosis, PPMI is 4-class for staged Parkinson's disease, HCPA is 4-class, and HCPYA is 7-class for human behavior recognition. In contrast, BrainMoE can bring an impressive performance enhancement by stratifying and adapting cognitive states. Note that ABIDE and PPMI are preprocessed by the pipeline of [33], which is different from our pipeline (see Appendix) for HCPs and UKB datasets. Furthermore, BrainMass and BrainJEPA reconstruct FC and BOLD latent features, respectively.

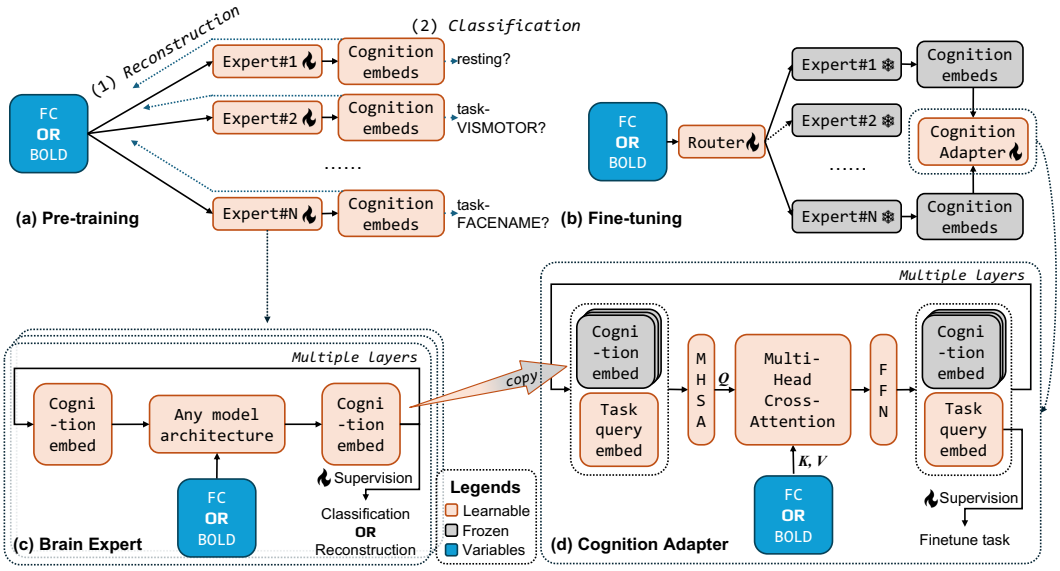

Figure 3: Framework of BrainMoE. **(a) Pre-training** has two options: (1) Previous models train with a reconstruction objective, where samples used for BrainMoE pre-training are stratified by cognitive states. (2) A new architecture of cognition classification can be set as the objective via cross-entropy between embeddings and cognitive states. **(b) Fine-tuning** a router for the expert selection and a cognition adapter for the combination of frozen experts. **(c) Brain expert** is adaptable to any model architecture, learning feature representation from either FC or BOLD signal, where the vector of latent feature before the final predictive layer is used as a cognition embedding. **(d) Cognition adapter** is a Transformer decoder, where MHSA stands for multi-head self-attention, and FFN is a feedforward network.

These primary results demonstrate empirical and clear evidence to support the motivation to stratify data according to cognitive states for multiple experts and learn joint cognition embeddings.

3 Methods

BrainMoE is designed to work with arbitrary brain foundation models as experts and to cooperate with a router and an adapter for fine-tuning on the downstream. Assume the input, BOLD or FC, is denoted by $\mathbf{X} \in \mathbb{R}^{M \times C_{in}}$ with M regions of the brain atlas and C_{in} channels of input vector. Target of brain experts is to produce cognition embeddings, $\mathbf{Z}$. For the router and adapter, it is to predict $\mathbf{Y}$ for downstream applications given pre-trained experts.

3.1 Framework

The framework of BrainMoE is separated into two stages as shown in Fig. 3 (a) pre-training and (b) fine-tuning, where N experts, denoted by $f(\cdot) : \mathbb{R}^{M \times C_{in}} \rightarrow \mathbb{R}^{C_{hid}}$ with C_{hid} the hidden channel number, learn from large-scale datasets containing subjects explicitly tasking on N cognition-related behaviors to produce a variety of cognition embeddings, and fine-tune a router for expert weights $\mathbf{P} \in \mathbb{R}^N$, for selecting top-$k$ ($k \in [1, N]$) experts, and a cognition adapter for predicting downstream tasks based on cognition embeddings. Following the observed performance that relies on input type, preprocessing pipeline, and model architecture, BrainMoE has a framework suitable to experts with no requirement for data type and architecture.

3.2 Expert pre-training

In Fig. 3 (a), two objectives can be used to pre-train the brain expert, the reconstruction and the classification. (1) The pre-training of existing brain foundation models is reconstructing latent feature of input, FC or BOLD, from its masked version via a bottleneck or transformer encoder architecture. We utilized BrainMass or BrainJEPA as candidates of brain experts. The latent feature is produced by

existing architectures as the cognition embedding, denoted by $\mathbf{Z} \in \mathbb{R}^{C_{hid}}$. Each expert is pre-trained with the data that has the same cognitive state.

$$\mathbf{Z} := f(\mathbf{X}) = \arg\min_{\mathbf{Z}} ||\mathbf{Z} - g(\mathbf{X})||^2, \tag{1}$$

where g is the target network. As shown in Fig. 3 (c), the brain expert can be any architecture that produces a latent feature representation, which is frozen and copied for the downstream. The pre-training objective is not restrict to the reconstruction or a new classification for cognitive states.

In Fig. 3 (a) (2), we propose a new pre-training objective, cognitive state classification, to explicitly learn from the cross-entropy between the latent feature and the cognitive state, $CELoss(\rho(\mathbf{Z}), \mathbf{Y}_{cog})$, where $\rho : \mathbb{R}^{C_{hid}} \rightarrow \mathbb{R}^1$ is a linear layer, and $\mathbf{Y}_{cog}$ is the binary label of a cognitive state. The architecture for this expert is the same as the cognition adapter introduced in the next section.

3.3 Cognition adapter fine-tuning

The architecture of the cognition adapter is designed as a Transformer decoder shown in Fig. 3 (d). The purpose of this adapter is to adapt multiple experts from a stratified feature representation based on cognitive states to a downstream application, a classification task in this work.

Assume that the token vectors shown in dashed rectangle in Fig. 3 (d) is denoted by $\bar{\mathbf{Z}} \in \mathbb{R}^{(k+P)\times C_{hid}}$, where P is the class number in the downstream application. Note that $\bar{\mathbf{Z}}_{:k} := \mathbf{Z} \odot \mathbf{P}$ representing the top-k cognition embeddings from experts and $\bar{\mathbf{Z}}_{k:(k+P)}$ denotes randomly initialized task query embeddings. It is also a cognition classifier without $\bar{\mathbf{Z}}_{:k}$. Then, as demonstrated in the architecture, a layer of the adapter starts at a multi-head self-attention (MHSA), $\bar{\mathbf{Z}} = \texttt{Softmax}(QK^T/\sqrt{C_{hid}})V$, with following definitions

$$Q := \bar{\mathbf{Z}}\bar{\boldsymbol{\alpha}}_h, K := \bar{\mathbf{Z}}\bar{\boldsymbol{\beta}}_h, V := \bar{\mathbf{Z}}\bar{\boldsymbol{\gamma}}_h, \tag{2}$$

where $\bar{\boldsymbol{\alpha}}_h, \bar{\boldsymbol{\beta}}_h, \bar{\boldsymbol{\gamma}}_h \in \mathbb{R}^{C_{hid}\times C_{hid}}$ are learnable parameters, and h is the head index. Last, a multi-head cross-attention brings the information from the raw input to the task embeddings. Suppose $\mathbf{I} \in \mathbb{R}^{M\times M}$ is FC matrix. Cross-attention between $\bar{\mathbf{Z}}$ and $\mathbf{I}$ with alternative definitions

$$Q := \mathbf{I}\hat{\boldsymbol{\alpha}}_h, K := \bar{\mathbf{Z}}\hat{\boldsymbol{\beta}}_h, V := \mathbf{I}\hat{\boldsymbol{\gamma}}_h, \tag{3}$$

where $\hat{\boldsymbol{\alpha}}_h, \hat{\boldsymbol{\gamma}}_h \in \mathbb{R}^{M\times C_{hid}}, \hat{\boldsymbol{\beta}}_h \in \mathbb{R}^{C_{hid}\times C_{hid}}$ are learnable parameters. FFN denotes a feedforward network constructed by MLP. Note that the bias in linear layers is omitted in this section for clarity. Finally, after multiple layers of the cognition adapter, a linear layer, $\mathbb{R}^{P\times C_{hid}} \rightarrow \mathbb{R}^{P\times 1}$, takes only the task query and produces the logistic prediction to accomplish fine-tuning on a downstream application.

4 Experiments

We evaluate the proposed BrainMoE on 3 pre-training datasets, including UK Biobank (UKB), HCP Aging (HCPA), and HCP Young Adult (HCPYA), and 7 downstream datasets, including ADNI, ABIDE, PPMI, Taowu, SZ, HCPA, and HCPYA. UKB and two HCPs contain 68,251 scans of brain fMRI from 21,797 subjects under 12 different cognitive states on resting or tasking. Five disease-related datasets contain more than 1,500 subjects under the same resting state but various health status.

To comprehensively evaluate and showcase the performance, we conduct experiments on both randomly initialized and pre-trained models across tasks involving disease, sex, and brain state recognition. Specifically, our study aims to address two key research questions: (**RQ1**) To what extent does BrainMoE improve the prediction performance from the baseline using different expert architectures? (**RQ2**) Which pre-training objectives are the most robust across various downstream applications? Additionally, we provide ablation studies to further support our findings.

Datasets We preprocess fMRI and partition brain regions using the AAL atlas [28] for UKB, HCPA, HCPYA, and ADNI. Details can be found in the Appendix. Other datasets are preprocessed by [33]. SZ is an in-house data preprocessed by a third party.

UK Biobank (HCPA) dataset [20] is a large-scale dataset with MRI data. There are fMRI (n=51,780) involved in this work. It consists of one resting state and one tasking state that engages cognitive and sensory-motor [13].

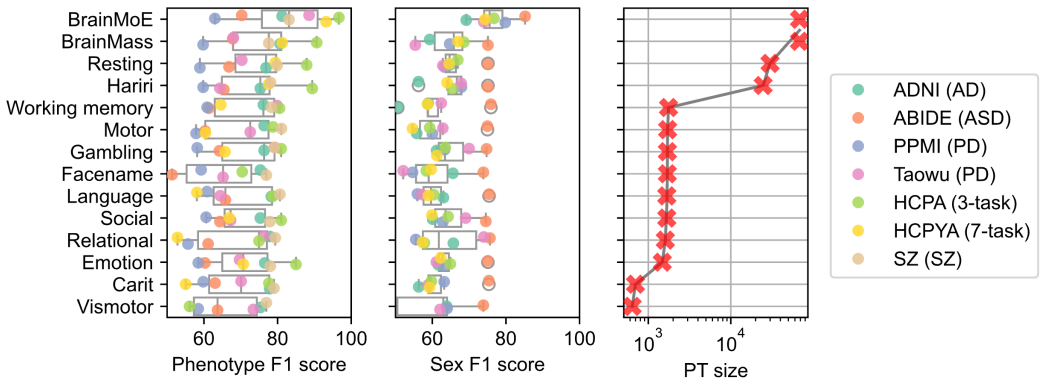

Figure 4: The performance of BrainMoE and BrainMass pre-trained with all samples, and n=12 pre-trained individual experts on phenotypic and sex classifications among 7 datasets, where scores lower than 50% are hidden for clarity, and the pre-training (PT) size ranges from 68,251 to 637.

The Lifespan Human Connectome Project Aging (HCPA) dataset [5] is instrumental in task recognition research, offering a comprehensive view of the aging process. It includes data from 717 subjects, encompassing fMRI records (n=4,863) with four human behaviors associated with memory, sensory-motor, and the resting state.

The Human Connectome Project Young Adult (HCPYA) dataset [29] has tackled key aspects of the neural pathways that underlie brain function and behavior via high-quality neuroimaging data in over 1,100 healthy young adults. It includes data from seven human behaviors associated with various cognitive tasks, e.g., language and working memory.

Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [32] serves as an invaluable resource, featuring a collection of pre-processed fMRI (n=138) and including clinical diagnostic labels. It encompasses a spectrum of cognitive states: Cognitive Normal (CN), Subjective Memory Complaints (SMC), Early-Stage Mild Cognitive Impairment (EMCI), Late-Stage Mild Cognitive Impairment (LMCI), and Alzheimer's Disease (AD). Considering the class unbalance issue, we simplified these categories into two broad groups based on disease severity: we combined CN, SMC, and EMCI into 'CN' group, while LMCI and AD were grouped as the 'AD' group.

Parkinson's Progression Markers Initiative (PPMI) dataset [33] presents a substantial collection of data from 209 subjects. It encompasses states of mental health: normal control, scans without evidence of dopaminergic deficit (SWEDD), prodromal, and Parkinson's disease (PD).

Taowu [33] is one of the earliest image datasets released for Parkinson's and contains 40 subjects.

Autism Brain Imaging Data Exchange (ABIDE) dataset [33] presents data from 1,025 young adults. The initiative aggregated fMRI data collected from laboratories around the world to support the research on Autism Spectrum Disorder (ASD).

Schizophrenia (SZ) is the in-house data that contains 189 subjects. There are 30 converted and 159 nonconverted.

Implementation Following previous works, our experiments are conducted with subject-level cross-validation (CV). The average score and the standard deviation are both listed. To make our results comparable with previous papers, HCPA, HCPYA, and ADNI use a 5-fold CV as same as [9, 31], while others use 10-fold as same as [33]. Since HCPs are used for both pre-training and fine-tuning, the training data in the two stages is always from the corresponding CV fold's training set to prevent data leakage. Hyperparameters, e.g., learning rate and hidden channels, can be found in the Appendix.

State-of-the-art (SOTA) brain foundation models, BrainMass [34] and BrainJEPA [11], are selected as expert architectures along with the new classifier architecture proposed in this work. Note that the original BrainMass fine-tuning utilizes the support vector machine (SVM) that has a lower scale of learnable parameters than others. Therefore, an enhancement of BrainMass is evaluated in this work by replacing SVM with a 2-layer MLP.

Table 1: MoE improvement on phenotypic classification F1 score compared to the baseline, where 30k is pre-trained on resting-state data (n=29,951), and 68k is pre-trained on all data (n=68,251). PT stands for pre-training. Colored text indicates the performance increase/decrease from using 68k.

	BrainMass					BrainJEPA		
Predictor	SVM	SVM	MLP	MLP	BrainMoE	ViT	ViT	BrainMoE
PT #	30k	68k	30k	68k	68k	30k	68k	68k
ADNI	$75.32_{\pm 7.06}$	$75.32_{\pm 7.06}$	$76.86_{\pm 7.26}$	$80.70_{\pm 7.85}$	$81.23_{\pm 11.00}$	$74.16_{\pm 8.55}$	$74.16_{\pm 8.55}$	$77.11_{\pm 6.64}$
↳AD		**0.00**		**3.84** ↑	**4.37** ↑		**0.00**	**2.95** ↑
ABIDE	$62.31_{\pm 1.95}$	$64.12_{\pm 2.31}$	$66.81_{\pm 4.18}$	$67.81_{\pm 3.91}$	$70.26_{\pm 3.40}$	$36.25_{\pm 6.93}$	$39.82_{\pm 3.91}$	$54.55_{\pm 9.89}$
↳ASD		**1.81** ↑		**1.00** ↑	**3.45** ↑		**3.77** ↑	**18.30** ↑
PPMI	$54.87_{\pm 15.76}$	$56.52_{\pm 14.86}$	$58.90_{\pm 14.29}$	$59.77_{\pm 14.22}$	$62.97_{\pm 13.94}$	$38.69_{\pm 13.91}$	$38.69_{\pm 13.91}$	$60.49_{\pm 11.59}$
↳PD (staged)		**1.65** ↑		**0.87** ↑	**4.07** ↑		**0.00**	**21.80** ↑
Taowu	$58.33_{\pm 34.78}$	$65.67_{\pm 20.55}$	$70.29_{\pm 17.97}$	$68.00_{\pm 21.46}$	$88.57_{\pm 12.51}$	$36.08_{\pm 26.38}$	$36.94_{\pm 20.98}$	$79.86_{\pm 14.46}$
↳PD (binary)		**7.34** ↑		**2.29** ↓	**18.28** ↑		**0.86** ↑	**43.78** ↑
SZ	$76.95_{\pm 9.01}$	$76.95_{\pm 9.01}$	$79.85_{\pm 8.69}$	$77.63_{\pm 8.56}$	$83.10_{\pm 11.33}$	$76.98_{\pm 9.00}$	$78.97_{\pm 9.79}$	$82.86_{\pm 9.19}$
↳Schizophrenia		**0.00**		**2.22** ↓	**3.25** ↑		**1.99** ↑	**5.88** ↑
HCPA	$85.16_{\pm 0.41}$	$89.73_{\pm 0.58}$	$87.91_{\pm 0.48}$	$90.63_{\pm 0.74}$	$96.67_{\pm 0.77}$	$59.54_{\pm 15.47}$	$53.12_{\pm 14.19}$	$81.74_{\pm 0.51}$
↳3-task,rest		**4.57** ↑		**2.72** ↑	**8.76** ↑		**6.42** ↓	**22.20** ↑
HCPYA	$77.51_{\pm 2.42}$	$80.87_{\pm 1.77}$	$79.40_{\pm 1.78}$	$81.27_{\pm 1.27}$	$93.19_{\pm 0.72}$	$50.68_{\pm 25.20}$	$56.10_{\pm 29.16}$	$74.59_{\pm 3.79}$
↳7-task		**3.36** ↑		**1.87** ↑	**13.79** ↑		**5.42** ↑	**23.91** ↑

4.1 RQ1: MoE vs baselines

The average F1 scores of BrainMoE and BrainMass pretrained with all samples, and n=12 individual experts per cognitive state on phenotypic and sex classifications among 7 datasets are shown in Fig. 4. Previous studies have demonstrated good PT data scalability with resting-state fMRI data. However, according to Fig. 4, there are consistently existing task-specific experts (e.g., language for AD, working memory for PD) outperforming Resting experts, confirming that task-state fMRI contains valuable information for brain modeling. Conclusively, BrainMoE holds the best performance compared to all experts across 7 datasets. This supports that the utilization of cognitive embeddings from BrainMoE leads to more robustness of brain modeling than naively training a single task of fMRI.

In Table 1, we summarize the impact of BrainMoE on downstream phenotypic classification, reporting improvements in the F1 score relative to non–MoE baselines for disease and human behavior recognition. Across all tasks (ADNI, ABIDE, PPMI, Taowu, SZ, HCPA, HCPYA), intuitively expanding pre-training data from 30k to 68k yields modest gains, even negative gains, for both BrainMass with SVM and MLP, e.g. ABIDE BrainMass SVM: +1.81 F1 and SZ BrainMass MLP: -2.22 F1, and BrainJEPA with Vision Transformer (ViT), e.g., HCPA BrainJEPA: -6.42 F1. In contrast, introducing our BrainMoE with the proposed cognition adapter on top of the 68k pre-trained backbone amplifies these gains substantially: phenotypic F1 score uplifts range from +3.25 F1 (SZ Schizophrenia) to +43.78 F1 (Taowu PD), and it consistently brings a positive effect. Even BrainJEPA baselines that do not gain F1 improvement from more data on HCPA and HCPYA benefit from BrainMoE, albeit to a lesser extent (e.g., HCPYA: +3.04 F1). Notably, the largest relative benefit appears on smaller cohorts, e.g., Taowu (n=40), where BrainMoE achieves +18.28 and +43.76 F1 over 68k BrainMass and BrainJEPA baselines.

Table 2 reports analogous results for sex classification. Worse than phenotypic classification, increasing pre-training size delivers small, commonly no improvements for BrainMass and BrainJEPA using SVM, MLP, and ViT, where 11 out of 18 experiments have dropped F1 scores in red text. In contrast, BrainMoE recovers and exceeds prior performance. It consistently demonstrates F1 score gains, except for BrainJEPA on ABIDE. It is worth noting that the most dramatic uplift appears on the smallest dataset (Taowu, n=40), where BrainMoE increases F1 by +43.76 on BrainJEPA.

Overall, the above results demonstrate that (1) a large scale pre-training without stratifying cognitive states improves downstream performance modestly (sometimes negatively), and (2) the proposed BrainMoE framework produces substantial gains, especially on downstream applications with a limited sample size.

Table 2: MoE improvement on sex classification F1 score compared to the baseline, where the sample size of the downstream dataset is indicated. PT stands for pre-training.

	BrainMass					BrainJEPA		
Predictor	SVM	SVM	MLP	MLP	BrainMoE	ViT	ViT	BrainMoE
PT #	30k	68k	30k	68k	68k	30k	68k	68k
ADNI	$48.60_{\pm 6.55}$	$54.30_{\pm 12.48}$	$64.82_{\pm 4.30}$	$59.30_{\pm 13.05}$	$69.22_{\pm 5.26}$	$37.70_{\pm 8.73}$	$36.42_{\pm 8.22}$	$62.98_{\pm 7.76}$
↳n=138		**5.70** ↑		**5.52** ↓	**4.40** ↑		**1.28** ↓	**25.28** ↑
ABIDE	$73.84_{\pm 3.49}$	$73.84_{\pm 3.49}$	$75.12_{\pm 5.27}$	$75.12_{\pm 6.32}$	$85.21_{\pm 3.77}$	$78.08_{\pm 5.84}$	$78.08_{\pm 5.84}$	$78.08_{\pm 6.15}$
↳n=1025		**0.00**		**0.00**	**10.09** ↑		**0.00**	**0.00**
PPMI	$52.03_{\pm 14.16}$	$56.58_{\pm 12.28}$	$63.32_{\pm 14.96}$	$64.73_{\pm 14.12}$	$79.81_{\pm 8.46}$	$46.23_{\pm 9.00}$	$46.23_{\pm 9.00}$	$67.57_{\pm 7.24}$
↳n=209		**4.55** ↑		**1.41** ↑	**16.49** ↑		**0.00**	**21.34** ↑
Taowu	$46.24_{\pm 23.96}$	$46.24_{\pm 23.96}$	$62.86_{\pm 28.20}$	$55.38_{\pm 20.93}$	$74.00_{\pm 27.79}$	$46.24_{\pm 23.97}$	$51.67_{\pm 21.12}$	$90.00_{\pm 12.96}$
↳n=40		**0.00**		**7.48** ↓	**11.14** ↑		**5.43** ↑	**43.76** ↑
HCPA	$66.20_{\pm 1.58}$	$68.25_{\pm 1.70}$	$66.93_{\pm 0.63}$	$68.58_{\pm 0.99}$	$76.76_{\pm 0.93}$	$40.32_{\pm 4.07}$	$40.32_{\pm 4.07}$	$44.24_{\pm 7.01}$
↳n=4863		**2.05** ↑		**1.65** ↑	**9.83** ↑		**0.00**	**3.92** ↑
HCPYA	$63.33_{\pm 3.01}$	$65.47_{\pm 3.51}$	$64.57_{\pm 2.53}$	$66.98_{\pm 3.30}$	$74.36_{\pm 4.43}$	$40.20_{\pm 4.22}$	$40.20_{\pm 4.22}$	$43.24_{\pm 8.19}$
↳n=3293		**2.14** ↑		**2.41** ↑	**9.79** ↑		**0.00**	**3.04** ↑

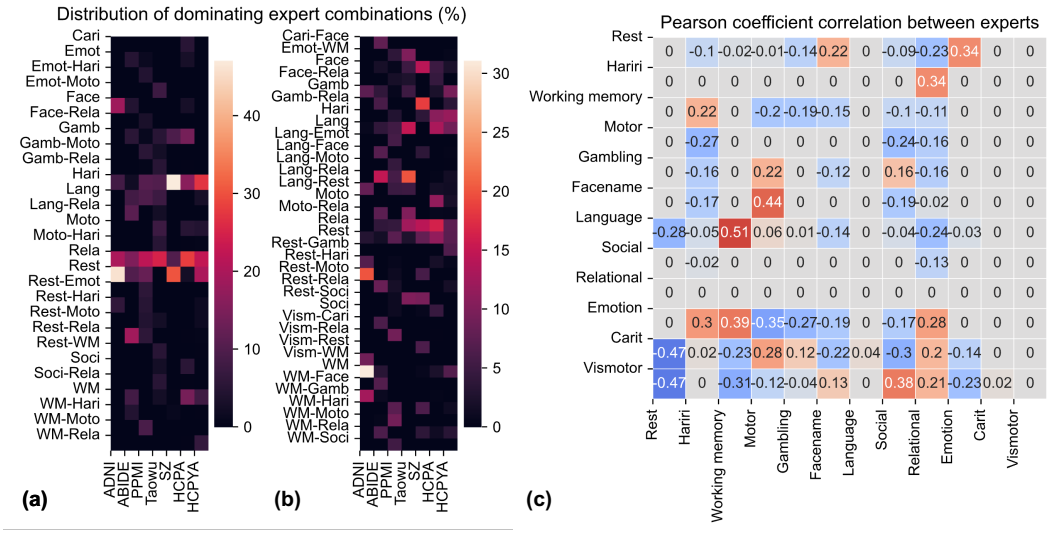

Figure 5: The distribution of dominating expert combinations of **(a)** late fusion MoE and **(b)** BrainMoE shows the router preference, where the first 4 letters of cognitive state are used as abbreviation. **(c)** The correlation between cognition embeddings of experts shows expert diversity.

4.2 RQ2: Pre-training objectives

As described in Sec 3, the BrainMoE framework has no requirement for data type and architecture. This property results in various pre-training objectives for the brain expert: FC reconstruction (FC recon.), BOLD reconstruction (BOLD recon.), and cognitive state classification (Cog. Classif.). To demonstrate which objectives are the most robust, we evaluated three objectives and an all-in-one BrainMoE on 7 downstream datasets in Table 3 and 4.

Briefly, FC reconstruction and all-in-one BrainMoE show the best robustness. They always rank in the first two places for phenotypic (Table 3) and sex classification (Table 4) across all downstream datasets, except for the smallest dataset Taowu (n=40). Unlike BrainMoE has the cognition adapter to implicitly utilize expert embeddings, the late fusion explicitly combines expert predictions, therefore cannot handle the unbalance issue. We can observe in Table 3 that FC reconstruction shows the most first rank, 4 out of 7 datasets, and all-in-one has the most first place, 4 out of 7, in Table 4. Although top-k selected experts in all-in-one BrainMoE contain experts in FC reconstruction, the self-attention in the cognition adapter mixes information between all types of pre-training objectives, yielding dropped and boosted performance for phenotypic and sex classification, respectively.

Table 3: MoE performance on phenotypic classification F1 score using three types of expert pre-trained with three objectives, along with an all-in-one MoE mixing all types of experts, where LF is Late Fusion, Ex. # denotes expert number. **Bold** is the first rank and underline is the second.

	Ex. #	ADNI	ABIDE	PPMI	Taowu	SZ	HCPA	HCPYA
Baseline								
BrainMass	1	$80.70_{\pm 7.85}$	$67.81_{\pm 3.91}$	$59.77_{\pm 14.22}$	$68.00_{\pm 21.46}$	$77.63_{\pm 8.56}$	$90.63_{\pm 0.74}$	$81.27_{\pm 1.27}$
BrainJEPA	1	$74.16_{\pm 8.55}$	$39.82_{\pm 3.91}$	$38.69_{\pm 13.91}$	$36.94_{\pm 20.98}$	$78.97_{\pm 9.79}$	$53.12_{\pm 14.19}$	$56.10_{\pm 29.16}$
LF-MoE	12	$73.33_{\pm 10.87}$	$\underline{69.89_{\pm 3.06}}$	$\underline{61.11_{\pm 15.29}}$	$\mathbf{91.24}_{\pm 11.72}$	$76.95_{\pm 9.50}$	$94.96_{\pm 2.64}$	$88.58_{\pm 3.39}$
BrainMoE								
FC recon.	12	$\mathbf{81.23}_{\pm 11.00}$	$\mathbf{70.26}_{\pm 3.40}$	$\mathbf{62.97}_{\pm 13.94}$	$88.57_{\pm 12.51}$	$83.10_{\pm 11.33}$	$\mathbf{96.67}_{\pm 0.77}$	$93.19_{\pm 0.72}$
BOLD recon.	12	$77.11_{\pm 6.64}$	$54.55_{\pm 9.89}$	$60.49_{\pm 11.59}$	$79.86_{\pm 14.46}$	$82.86_{\pm 9.19}$	$81.74_{\pm 0.51}$	$74.59_{\pm 3.79}$
Cog. classif.	12	$79.70_{\pm 10.28}$	$68.65_{\pm 3.81}$	$59.23_{\pm 14.65}$	$\underline{90.48_{\pm 14.64}}$	$\mathbf{83.36}_{\pm 10.08}$	$96.28_{\pm 0.70}$	$\underline{95.81_{\pm 0.48}}$
All-in-one	36	$\underline{79.73_{\pm 10.60}}$	$69.13_{\pm 4.08}$	$60.76_{\pm 14.85}$	$85.93_{\pm 18.32}$	$\underline{83.91_{\pm 8.07}}$	$\underline{96.66_{\pm 0.94}}$	$\mathbf{96.81}_{\pm 0.41}$

Table 4: MoE performance on sex classification F1 score using three types of expert pre-trained with three objectives, along with an all-in-one MoE mixing all types of experts. **Bold** is the first rank and underline is the second.

	Ex. #	ADNI	ABIDE	PPMI	Taowu	HCPA	HCPYA
Baseline							
BrainMass	1	$59.30_{\pm 13.05}$	$75.12_{\pm 6.32}$	$64.73_{\pm 14.12}$	$55.38_{\pm 20.93}$	$68.58_{\pm 0.99}$	$66.98_{\pm 3.30}$
BrainJEPA	1	$36.42_{\pm 8.22}$	$78.08_{\pm 5.84}$	$46.23_{\pm 9.00}$	$51.67_{\pm 21.12}$	$40.32_{\pm 4.07}$	$40.20_{\pm 4.22}$
BrainMoE							
FC recon.	12	$\underline{69.22_{\pm 5.26}}$	$\mathbf{85.21}_{\pm 3.77}$	$\underline{79.81_{\pm 8.46}}$	$74.00_{\pm 27.79}$	$\underline{76.76_{\pm 0.93}}$	$74.36_{\pm 4.43}$
BOLD recon.	12	$62.98_{\pm 7.76}$	$78.08_{\pm 6.15}$	$67.57_{\pm 7.24}$	$\mathbf{90.00}_{\pm 12.96}$	$44.24_{\pm 7.01}$	$43.24_{\pm 8.19}$
Cog. classif.	12	$65.75_{\pm 7.61}$	$82.82_{\pm 5.11}$	$79.07_{\pm 7.28}$	$72.22_{\pm 25.17}$	$75.65_{\pm 1.26}$	$\underline{75.18_{\pm 1.15}}$
All-in-one	36	$\mathbf{70.72}_{\pm 7.58}$	$\underline{82.85_{\pm 5.91}}$	$\mathbf{82.80}_{\pm 5.65}$	$\underline{75.29_{\pm 30.56}}$	$\mathbf{78.34}_{\pm 2.18}$	$\mathbf{77.67}_{\pm 1.54}$

4.3 Rounter and expert analysis

The preference of routers in late fusion MoE and BrainMoE is shown in Fig. 5 (a) and (b), respectively, where the percentage indicates how many samples have a combination of experts with dominating router logits ($\geqslant \frac{1}{N}$). Clearly, BrainMoE has diverse and similar dominating combinations for heterogeneous and homogeneous applications, respectively. The combinations are mainly dual, and datasets with the same task share similar pattern (i.e., cognitive state for HCPA and HCPYA, and PD for PPMI and Taowu). In contrast, late fusion consistently has single dominating expert due to data scale diversity, e.g., Rest (n=29,971), which implies that the routers trained by BrainMoE learned more neuroscientific knowledge than the late fusion. Furthermore, the investigation on expert embeddings is shown in Fig. 5 (c). The absolute value of correlation is mostly less than 0.5, indicating experts are not in conflict or redundant to each other.

4.4 More applications

The age regression across 6 datasets has been evaluated for the best baseline, 68k version of BrainMass, and BrainMoE, as listed in Table 5. We can observe the performance of BrainMoE is the best, where the improvement is especially significant for ABIDE (n=1,025, 6-58 yrs) with MSE 36.77 $\longrightarrow$ 4.86. This empirical evidence further supports the robustness of BrainMoE.

Table 5: Age regression performance compared to the baseline, where the sample size of the downstream dataset is indicated, and unit is year.

MSE	ADNI	ABIDE	PPMI	Taowu	HCPA	HCPYA
BrainMass	$36.28_{\pm 19.83}$	$36.77_{\pm 17.2}$	$33.09_{\pm 21.87}$	$38.10_{\pm 28.33}$	$22.66_{\pm 7.51}$	$5.46_{\pm 2.69}$
BrainMoE	$\mathbf{36.27}_{\pm 10.23}$	$\mathbf{4.86}_{\pm 2.67}$	$\mathbf{29.89}_{\pm 8.39}$	$\mathbf{30.72}_{\pm 12.25}$	$\mathbf{10.56}_{\pm 2.86}$	$\mathbf{3.45}_{\pm 1.08}$

Multimodal applications of cognitive state, sex classifications, and age regression in an fMRI-EEG dataset [27] are listed in Table 6. There are 388 fMRI-EEG pairs from 22 healthy subjects under 8 cognitive states (age$\in$[23,51], F:M=1:1). BrainMoE is evaluated here with n=1 pre-

Table 6: BrainMoE applies on a multimodal dataset, NATVIEW [27].

NATVIEW	8-task (F1)	Sex (F1)	Age (MSE)
BrainMass (fMRI)	$67.66_{\pm 5.74}$	$63.67_{\pm 5.16}$	$8.05_{\pm 5.58}$
CBraMod (EEG)	$68.71_{\pm 1.46}$	$65.39_{\pm 2.33}$	$8.26_{\pm 5.97}$
BrainMoE (13 Ex.)	$\mathbf{68.73}_{\pm 3.72}$	$\mathbf{65.47}_{\pm 5.38}$	$\mathbf{7.99}_{\pm 5.53}$

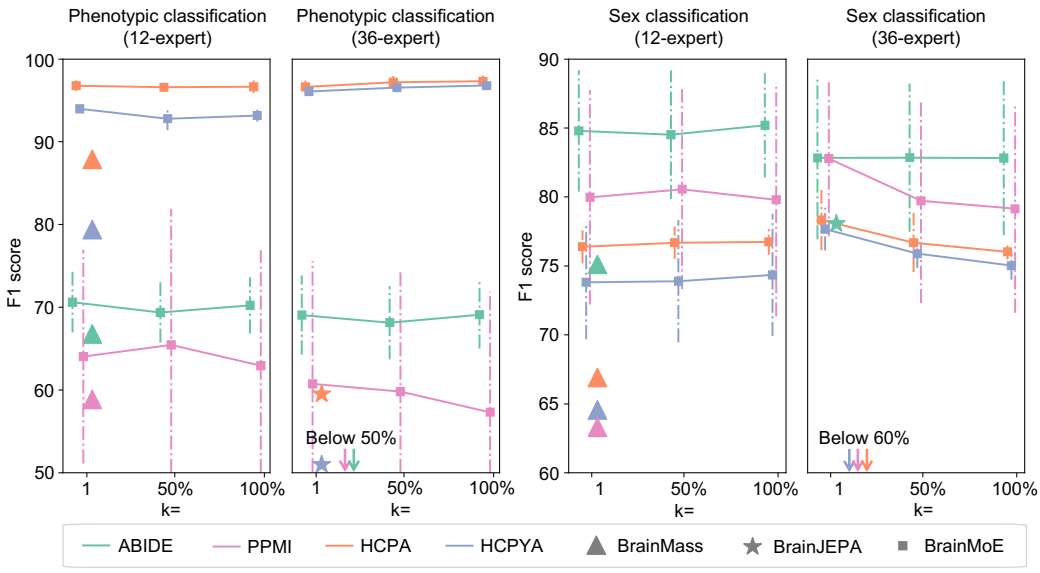

Figure 6: Impact of altering k on downstream classification performance. Mean F1 scores with standard deviation are reported for four downstream datasets under three expert-selection regimes (top-1, top 50%, and all experts). Colors represent different datasets.

trained CBraMod [30] expert plus n=12 cognition classifier experts (Shaefer400 version). Results from a 5-fold CV show the best performance by multimodal BrainMoE.

4.5 Ablations

Top-k We alter k in BrainMoE to strictly limit how many top experts can be selected for downstream applications. We evaluated $k = 1$, $k = 50\%$, and $k = 100\%$ with 12 and 36 experts on four datasets with relatively larger sample sizes in Fig. 6, where 12-expert BrainMoE uses BrainMass as the expert architecture. From the left two panels, it is obvious that increasing k has a slight difference for all phenotypic tasks, except for HCPA. In the right two panels, we can observe that sex classification benefits less from a larger k. Overall, expert scaling in BrainMoE yields clear benefits for complex, data-scarce phenotypic tasks, with most gains achieved by employing top half of the experts. Beyond this, adding experts delivers diminishing returns. In contrast, for the simpler binary sex classification task, especially with a large expert pool, additional experts do not meaningfully improve performance and may introduce redundancy or overfitting. Thus, tailoring the number of active experts to task complexity and dataset size is key for efficient MoE deployment.

5 Conclusion

In conclusion, we propose a new framework of the brain foundation model, BrainMoE, to pre-train with overlooked tasking-state fMRI for robust downstream applications. We observe that existing brain foundation models learn from fMRI derived from a narrow range of cognitive states, while there are 11 available cognitive states as subjects performing explicit tasks in large scale datasets. Furthermore, we showcase that (i) the straightforward utilization of data with rich human behavioral variables by pre-training with all data and (ii) the late fusion MoE both improve performance marginally. Aiming at these challenges, BrainMoE pre-trains each expert on a portion of the datasets with the same cognitive state among 12 different states for a robust brain foundation model. We design a scalable cognition adapter to mix brain experts for downstream fine-tuning so that BrainMoE can handle orthogonal cognition embeddings and be robust on the boutique downstream datasets. With sufficient 68,251 pre-training fMRI scans among UKB and HCP with 12 different cognitive states, BrainMoE has shown impressive performance boosting on a variety of applications, including sex, age prediction, human behavior recognition, multimodal applications, and early diagnosis of various brain diseases. The promising results demonstrated on eight datasets from three different pipelines indicate great potential to facilitate current neuroimaging applications in clinical routines.

Acknowledgement

This work was supported by the National Institutes of Health (AG091653, AG068399, AG084375) and the Foundation of Hope. Tianlong Chen was partially funded by the National Institutes of Health (NIH) under award 1R01EB037101-01. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH.

References

[1] Analucia A Alegria, Joaquim Radua, and Katya Rubia. Meta-analysis of fmri studies of disruptive behavior disorders. *American Journal of Psychiatry*, 173(11):1119–1130, 2016.

[2] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

[3] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.

[4] Hasan A Bedel, Irmak Sivgin, Onat Dalmaz, Salman UH Dar, and Tolga Çukur. Bolt: Fused window transformers for fmri time series analysis. *Medical Image Analysis*, 88:102841, 2023.

[5] Susan Y Bookheimer, David H Salat, Melissa Terpstra, Beau M Ances, Deanna M Barch, Randy L Buckner, Gregory C Burgess, Sandra W Curtiss, Mirella Diaz-Santos, Jennifer Stine Elam, et al. The lifespan human connectome project in aging: an overview. *Neuroimage*, 185:335–348, 2019.

[6] César Caballero-Gaudes and Richard C Reynolds. Methods for cleaning the bold fmri signal. *Neuroimage*, 154:128–149, 2017.

[7] Li Chen, Patrick Bedard, Mark Hallett, and Silvina G Horovitz. Dynamics of top-down control and motor networks in parkinson's disease. *Movement Disorders*, 36(4):916–926, 2021.

[8] Hejie Cui, Wei Dai, Yanqiao Zhu, Xuan Kan, Antonio Aodong Chen Gu, Joshua Lukemire, Liang Zhan, Lifang He, Ying Guo, and Carl Yang. Braingb: a benchmark for brain network analysis with graph neural networks. *IEEE transactions on medical imaging*, 42(2):493–506, 2022.

[9] Tingting Dan, Jiaqi Ding, Ziquan Wei, Shahar Kovalsky, Minjeong Kim, Won Hwa Kim, and Guorong Wu. Re-think and re-design graph neural networks in spaces of continuous graph diffusion functionals. *Advances in Neural Information Processing Systems*, 36:59375–59387, 2023.

[10] Jiaqi Ding, Tingting Dan, Ziquan Wei, Hyuna Cho, Paul J Laurienti, Won Hwa Kim, and Guorong Wu. Machine learning on dynamic functional connectivity: Promise, pitfalls, and interpretations. *arXiv preprint arXiv:2409.11377*, 2024.

[11] Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Su Xian Chong, Fang Ji, Nathanael Ren Jie Tong, Christopher Li Hsian Chen, and Juan Helen Zhou. Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. *arXiv preprint arXiv:2409.19407*, 2024.

[12] Michael D Fox, Abraham Z Snyder, Justin L Vincent, Maurizio Corbetta, David C Van Essen, and Marcus E Raichle. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences*, 102(27):9673–9678, 2005.

[13] Ahmad R Hariri, Alessandro Tessitore, Venkata S Mattay, Francesco Fera, and Daniel R Weinberger. The amygdala response to emotional stimuli: a comparison of faces and scenes. *Neuroimage*, 17(1):317–323, 2002.

[14] Tong He, Lijun An, Pansheng Chen, Jianzhong Chen, Jiashi Feng, Danilo Bzdok, Avram J Holmes, Simon B Eickhoff, and BT Thomas Yeo. Meta-matching as a simple framework to translate phenotypic predictive models from big to small data. *Nature neuroscience*, 25(6):795–804, 2022.

[15] Zhibin He, Wuyang Li, Yifan Liu, Xinyu Liu, Junwei Han, Tuo Zhang, and Yixuan Yuan. Fm-app: Foundation model for any phenotype prediction via fmri to smri knowledge transfer. *IEEE Transactions on Medical Imaging*, 2024.

[16] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012.

[17] Xuan Kan, Wei Dai, Hejie Cui, Zilong Zhang, Ying Guo, and Carl Yang. Brain network transformer. *Advances in Neural Information Processing Systems*, 35:25586–25599, 2022.

[18] Raymond Li, Gabriel Murray, and Giuseppe Carenini. Mixture-of-linguistic-experts adapters for improving and interpreting pre-trained language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9456–9469, Singapore, December 2023. Association for Computational Linguistics.

[19] Xiaoxiao Li, Yuan Zhou, Nicha Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H Staib, Pamela Ventola, and James S Duncan. Braingnn: Interpretable brain graph neural network for fmri analysis. *Medical Image Analysis*, 74:102233, 2021.

[20] Thomas J Littlejohns, Jo Holliday, Lorna M Gibson, Steve Garratt, Niels Oesingmann, Fidel Alfaro-Almagro, Jimmy D Bell, Chris Boultwood, Rory Collins, Megan C Conroy, et al. The uk biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nature communications*, 11(1):2624, 2020.

[21] Charles J Lynch, Lucina Q Uddin, Kaustubh Supekar, Amirah Khouzam, Jennifer Phillips, and Vinod Menon. Default mode network in childhood autism: posteromedial cortex heterogeneity and relationship with social deficits. *Biological psychiatry*, 74(3):212–219, 2013.

[22] Christopher J Markiewicz, Krzysztof J Gorgolewski, Franklin Feingold, Ross Blair, Yaroslav O Halchenko, Eric Miller, Nell Hardcastle, Joe Wexler, Oscar Esteban, Mathias Goncavles, et al. The openneuro resource for sharing of neuroscience data. *Elife*, 10:e71774, 2021.

[23] Josue Ortega Caro, Antonio Henrique Oliveira Fonseca, Christopher Averill, Syed A Rizvi, Matteo Rosati, James L Cross, Prateek Mittal, Emanuele Zappala, Daniel Levine, Rahul M Dhodapkar, et al. Brainlm: A foundation model for brain activity recordings. *bioRxiv*, pages 2023–09, 2023.

[24] Anwar Said, Roza Bayrak, Tyler Derr, Mudassir Shabbir, Daniel Moyer, Catie Chang, and Xenofon Koutsoukos. Neurograph: Benchmarks for graph machine learning in brain connectomics. *Advances in Neural Information Processing Systems*, 36:6509–6531, 2023.

[25] Hannah S Savage, Peter CR Mulders, Philip FP Van Eijndhoven, Jasper Van Oort, Indira Tendolkar, Janna N Vrijsen, Christian F Beckmann, and Andre F Marquand. Dissecting task-based fmri activity using normative modelling: an application to the emotional face matching task. *Communications Biology*, 7(1):888, 2024.

[26] Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. *arXiv preprint arXiv:2403.11207*, 2024.

[27] Qawi K Telesford, Eduardo Gonzalez-Moreira, Ting Xu, Yiwen Tian, Stanley J Colcombe, Jessica Cloud, Brian E Russ, Arnaud Falchier, Maximilian Nentwich, Jens Madsen, et al. An open-access dataset of naturalistic viewing using simultaneous eeg-fmri. *Scientific Data*, 10(1):554, 2023.

[28] Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Octave Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289, 2002.

[29] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

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

[31] Ziquan Wei, Tingting Dan, Jiaqi Ding, and Guorong Wu. Neuropath: A neural pathway transformer for joining the dots of human connectomes. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[32] Michael W Weiner, Dallas P Veitch, Paul S Aisen, Laurel A Beckett, Nigel J Cairns, Jesse Cedarbaum, Michael C Donohue, Robert C Green, Danielle Harvey, Clifford R Jack Jr, et al. Impact of the alzheimer's disease neuroimaging initiative, 2004 to 2014. *Alzheimer's & Dementia*, 11(7):865–884, 2015.

[33] Jiaxing Xu, Yunhan Yang, David Huang, Sophi Shilpa Gururajapathy, Yiping Ke, Miao Qiao, Alan Wang, Haribalan Kumar, Josh McGeown, and Eryn Kwon. Data-driven network neuroscience: On data collection and benchmark. *Advances in Neural Information Processing Systems*, 36:21841–21856, 2023.

[34] Yanwu Yang, Chenfei Ye, Guinan Su, Ziyao Zhang, Zhikai Chang, Hairui Chen, Piu Chan, Yue Yu, and Ting Ma. Brainmass: Advancing brain network analysis for diagnosis with large-scale self-supervised learning. *IEEE Transactions on Medical Imaging*, 2024.

[35] Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23219–23230, 2024.

[36] Shu Zhang, Xiang Li, Jinglei Lv, Xi Jiang, Lei Guo, and Tianming Liu. Characterizing and differentiating task-based and resting state fmri signals via two-stage sparse representations. *Brain imaging and behavior*, 10:21–32, 2016.

[37] Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context. *arXiv preprint arXiv:2404.13046*, 2024.

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: See Sec 1

   Guidelines:

   - The answer NA 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 NA 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: See Sec E

   Guidelines:

   - The answer NA 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: [NA] .

Justification: [NA] .

Guidelines:

- The answer NA 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: See Sec 4 and Appendix

   Guidelines:

   - The answer NA 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: See Appendix

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.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://nips.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, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Sec 4

   Guidelines:

   - The answer NA 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: See Sec 4

   Guidelines:

   - The answer NA 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: See Appendix

   Guidelines:

   - The answer NA 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: [Yes]

   Guidelines:

   - The answer NA 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: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA 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: See Sec 4

    Guidelines:

    - The answer NA 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: See Appendix

Guidelines:

- The answer NA 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: [NA]

Justification: [NA]

Guidelines:

- The answer NA 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: [NA]

Justification: [NA]

Guidelines:

- The answer NA 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 rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA 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 (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Accessibility

Public data is accessible via internet (UKB[2], HCPA[3], HCPYA[4], ADNI[5]. PPMI, ABIDE, and Taowu can be found here[6]). The licenses to obtain those data can also be accessed on the websites. The codes and data split settings can be acquired via this code repository[7].

B Data preprocessing

The neuroimage processing used for ADNI, UKB, HCPYA, and HCPA consists of the following major steps: (1) We segment the T1-weighted image into white matter, gray matter, and cerebral spinal fluid using FSL software [16]. (2) On top of the tissue segmentation in Fig. 7, we parcellate the cortical surface of fMRI into cortical regions according to the atlas as a regional signal of time-series in Fig. 7, where FC, in the end, is the Pearson correlation coefficient between regional time-series.

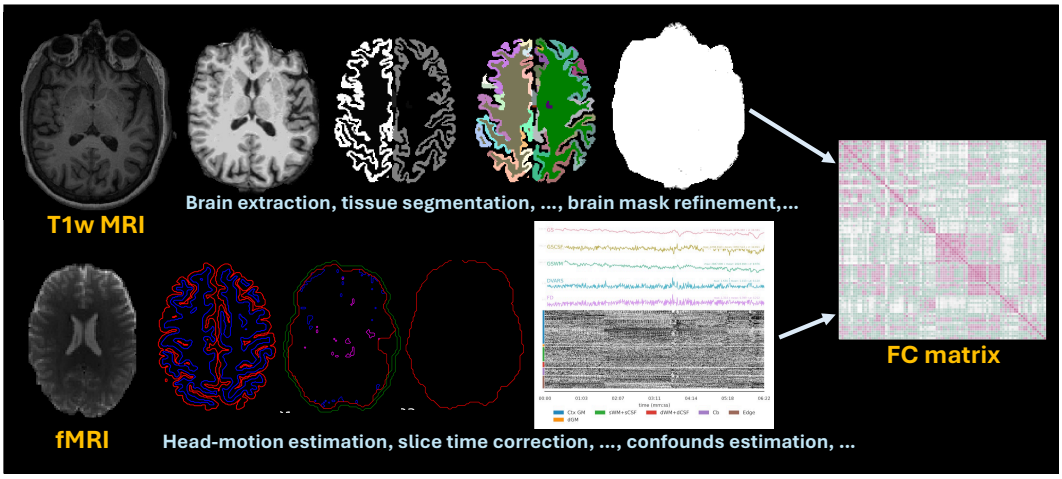

Figure 7: General workflows for processing T1-weighted image (T1w MRI) and functional MRI (fMRI). The output is shown at the right, including the brain network of FC.

C Computing environments and hyperparameters

The experiments are done on a Linux system with one NVIDIA RTX 6000 Ada. Batch size and learning rate are set as 128 and 1e-4, respectively. The maximum epoch is set as 200 and $C_{hid} = 2048$. Training will be early stopped if accuracy keeps dropping in 50 epochs.

D Comparison between previous works

We list the comparison of experimental datasets between previous works in Table 7.

E Computational complexity

The limitation of BrainMoE is more computational complexity, as listed in Table 8. BrainMoE with 12 experts spends $4\times$ more time than baselines. All-in-one with 36 experts nearly doubles the time cost.

[2] https://www.ukbiobank.ac.uk/
[3] https://www.humanconnectome.org/
[4] https://www.humanconnectome.org/study/hcp-young-adult/overview
[5] https://adni.loni.usc.edu/
[6] https://auckland.figshare.com/articles/dataset/NeurIPS_2022_Datasets/21397377
[7] https://github.com/Chrisa142857/brain_moe

Table 7: The comparison of experimental datasets between previous works.

	BrainLM (2024) [23]	BrainMass (2024) [34]	BrainJEPA (2024) [11]	BrainMoE (Ours)
Brain atlas	AAL424	C200	Schaefer400	AAL116
Cognitive state	resting, task-hariri	resting	resting	resting, 11 types of tasking
Pre-train dataset	UKB, HCP	UKB, HCP, OpenNeuron	UKB, HCP	UKB, HCP
Pre-train data #	61,038	64,584	40,162	68,251
Fine-tune dataset	UKB, HCP	ASD, ADHD, AD, PD, MDD	UKB, HCP, ADNI	HCP, ASD, AD, PD, SZ
Parameter amount	650M	34M	307M	709M

Table 8: Computational time cost of BrainMoE inference with two existing architectures and the all-in-one BrainMoE on the ABIDE dataset.

Test time (ms/sample)	BrainMass	BrainJEPA
Single model	37.08	28.13
BrainMoE	157.60	133.26
All-in-one	287.21	

F Visual decoding potential

Visual decoding task for a new dataset NSD [2] has also been evaluated for MindEye2 [26] as the baseline and BrainMoE. We pre-trained two MindEye2s as the specific experts for long-term (novel trials) and short-term memory (easy/hard trials), respectively. Since visual decoding is a generative task, output contains much higher dimensions (256×1664 vs. class number 2 to 7) than downstream tasks focused in the main text. Therefore, we skipped our cognition adapter by weighted summing the diffusion prior of two experts with the BrainMoE routing probabilities. Both baseline and BrainMoE are pretrained with subjects 2-7 and finetuned with subject 1 on the entire 40 sessions. The final train and test losses, cosine similarity, and Mean Squared Error (MSE) during finetuning are listed in Table 9. Given the evidence that the performance of BrainMoE is better than the single expert MindEye2, there is potential for BrainMoE to expand to visual decoding.

Table 9: Visual decoding performance.

	MindEye2	BrainMoE
Train loss	9.639	**7.994**
Test loss	11.142	**9.405**
Cos. Sim.	0.778	**0.840**
MSE	0.301	**0.261**

G Scalability analysis

MLP as a universal predictive head is used in related works for the MoE adapter. Fig. 8 is the comparison between MLP and the proposed cognition adapter with different amounts of learnable parameters.

H Visualization

The attention weights conducted by BrainMoE with different k is visualized in Fig. 9. We can observe: (1) Advanced by the cognition adaptor, BrainMoE agrees with current neuroscience knowledge since it mainly attends to DAN and DMN for ASD [12, 21], SMN and FPN for PD [7]. (2) Differences are slight across enlarged k, indicating that the router produces consistent expert weights.

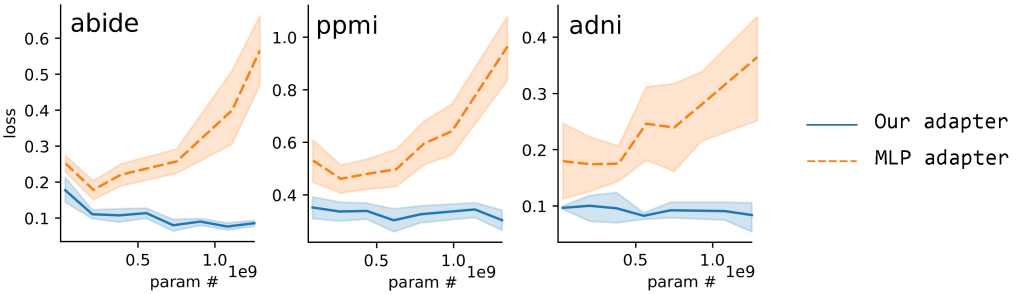

Figure 8: The scalability of the MLP adapter and our adapter on disease prediction, where the y-axis is the fine-tuning loss.

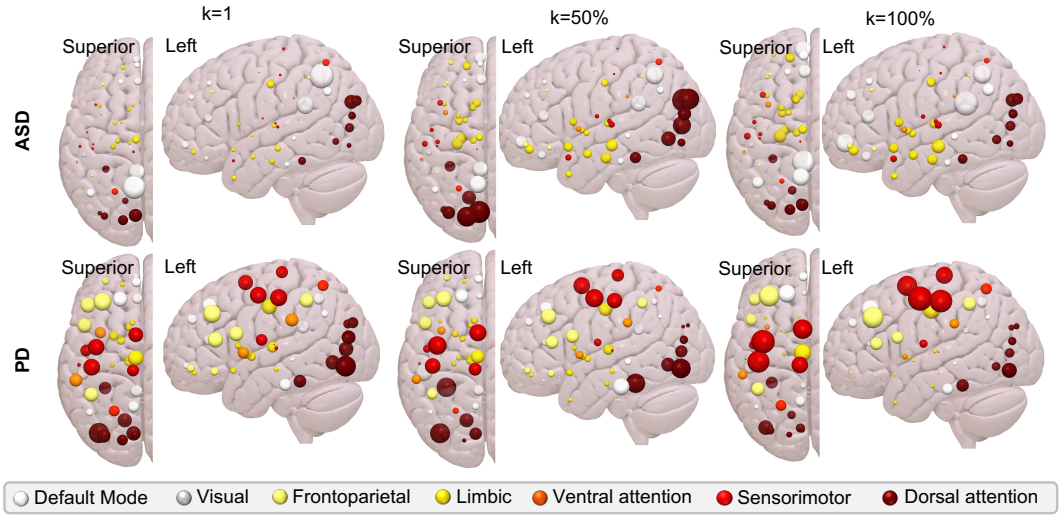

Figure 9: Visualization of attention weights by FC reconstruction BrainMoE.