Brain-OF: An Omnifunctional Foundation Model for fMRI, EEG and MEG

Hanning Guo[1,2] **Hanwen Bi**[3,4] **Farah Abdellatif**[1,2] **Andrei Galbenus**[1,2]
Jon. N. Shah[1,5,6,7] **Abigail Morrison**[8,2] **Jürgen Dammers**[1,9*]

[1]INM-4, Forschungszentrum Jülich, Germany
[2]Department of Computer Science 3 - Software Engineering, RWTH Aachen University, Germany
[3]INM-7, Forschungszentrum Jülich, Germany
[4]Institute of Systems Neuroscience, Heinrich Heine University, Germany
[5]Department of Neurology, RWTH Aachen University, Germany
[6]JARA-BRAIN-Translational Medicine, Germany
[7]INM–11, JARA, Forschungszentrum Jülich, Germany
[8]IAS-6, Forschungszentrum Jülich, Germany
[9]Department of Psychiatry, Psychotherapy and Psychosomatics, RWTH Aachen University, Germany

{h.guo, j.dammers}@fz-juelich.de

Abstract

Brain foundation models have achieved remarkable advances across a wide range of neuroscience tasks. However, most existing models are limited to a single functional modality, restricting their ability to exploit complementary spatiotemporal dynamics and the collective data scale across different neuroimaging techniques. This limitation largely arises from severe **semantic heterogeneity** and **resolution discrepancies** among modalities. To address these challenges, we propose **Brain-OF**, an omnifunctional brain foundation model jointly pretrained on fMRI, EEG and MEG, capable of handling both unimodal and multimodal inputs within a unified framework. To reconcile heterogeneous spatiotemporal resolutions, we introduce the *Any-Resolution Neural Signal Sampler*, which projects diverse brain signals into a shared semantic space. To further manage semantic shifts, the Brain-OF backbone integrates DINT attention with a Sparse Mixture of Experts, where shared experts capture modality-invariant representations and routed experts specialize in modality-specific semantics. Furthermore, to explicitly internalize the characteristics of neural activity through self-supervised learning, we propose *Masked Temporal-Frequency Modeling*, a dual-domain pretraining objective that jointly reconstructs brain signals in both the time and frequency domains. Brain-OF is pretrained on a large-scale corpus comprising around 40 datasets and demonstrates superior performance across diverse downstream tasks, highlighting the benefits of joint multimodal integration and dual-domain pretraining. Our code and pretrained model weights are available on our project page `https://hanningguo.github.io/BrainOF`.

1 Introduction

A central lesson from the recent success of deep learning is that pretraining on large-scale, high-quality and diverse data facilitates the emergence of general-purpose models that consistently outperform task-specific ones [1]. Following this paradigm, brain foundation models have shown promising generalization across a wide range of neural signal processing tasks, including cognitive decoding, affective computing and neurological disease diagnosis [2–5]. Despite this progress, the scale of available neuroimaging data remains orders of magnitude smaller than the billion-scale corpora in natural language processing and computer vision. Although open-source large-scale neuroimaging

*Corresponding author

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

collections are growing, data from any single modality remain scarce for training large models due to high acquisition costs and privacy constraints. Meanwhile, most existing brain foundation models are pretrained on a single modality, restricting both data scale and representational diversity, and failing to exploit the complementary characteristics of different brain signals. To address these limitations, we integrate data from the three primary functional neuroimaging modalities (fMRI, EEG and MEG) to develop an *omnifunctional* brain foundation model.

A pivotal question, therefore, is how to effectively leverage such heterogeneous multi-modal data to pretrain an omnifunctional brain foundation model. The first challenge lies in the variability introduced by diverse scanning devices, sampling rates, acquisition protocols and recording durations across international sites, leading to significant **structural variability**. A deeper challenge stems from the fundamental differences in the physical imaging mechanisms of each modality. Specifically, fMRI detects hemodynamic changes, such as oxygenation, blood volume and flow, to measure the Blood-Oxygen-Level-Dependent (BOLD) signal, offering high spatial resolution but limited temporal fidelity [6]. In contrast, EEG captures electrical potentials generated by cortical neurons, providing excellent temporal resolution but poor spatial fidelity [7]. MEG records magnetic fields induced by neuronal electrical activity and shares EEG's high temporal resolution while providing improved spatial localization, though still inferior to that of fMRI [8]. These distinct signal origins result in severe differences in spatiotemporal resolution and **semantic heterogeneity**. However, this heterogeneity also presents a unique opportunity. By integrating these modalities, a unified model can exploit the high-frequency temporal dynamics of EEG and MEG, which are lost in the sluggish hemodynamic response of fMRI, while leveraging the superior spatial specificity of fMRI that is inherently lacking in electrophysiological recordings. Such complementary integration enables a holistic characterization of temporal, spectral, and spatial dynamics, ultimately fostering richer and more transferable representations. A third challenge is the inherently low signal-to-noise ratio of brain signals, which can mislead vanilla attention mechanisms to attend to irrelevant fluctuations rather than biologically meaningful patterns.

On the other hand, the success of large-scale pretraining is fundamentally rooted in the generative learning paradigm, which insists that reconstruction compels models to internalize the underlying data distribution [9, 10]. Brain signals are inherently multidimensional, characterized by complex spatial configurations, rich temporal dynamics and structured frequency compositions. Although recent brain foundation models have introduced dual-branch architectures to jointly process temporal and spectral features [3, 11], frequency domain information is typically treated as a static auxiliary input for passive feature extraction, rather than as a target for active reconstruction. Therefore, a specialized pretraining objective is essential to explicitly enforce joint reconstruction across both temporal and frequency domains. By coupling temporal and spectral reconstruction, the model is compelled to capture complementary dynamics and intrinsic semantics of neural signals across heterogeneous modalities.

To address these challenges, we propose **Brain-OF**, an omnifunctional brain foundation model designed to learn holistic representations from neural signals across the three primary functional modalities: fMRI, EEG and MEG. Our architecture integrates a Sparse Mixture of Experts (MoE) with DINT attention to effectively model heterogeneous semantic tokens generated by the Any-Resolution Neural Signal Sampler (ARNESS), which projects signals of varying spatiotemporal resolutions into a unified semantic space. The main contributions are summarized as follows:

- **Unified Omnifunctional Framework.** We introduce Brain-OF, a versatile framework capable of processing both unimodal and multimodal brain data within a single model. To support large-scale pretraining, we curate a corpus of around 40 public datasets spanning fMRI, EEG, and MEG. By cohesively integrating these three complementary neuroimaging modalities, Brain-OF enhances generalization across modalities, tasks, sites and subjects.
- **Heterogeneity Aligned Architecture.** We propose ARNESS to project heterogeneous neural signals into a shared semantic space. To accommodate semantic diversity, we employ a Sparse MoE architecture in which shared experts capture modality-invariant features while routed experts specialize in modality-specific semantics. Furthermore, we incorporate DINT attention to suppress irrelevant context and efficiently capture both local and global dependencies.
- **Joint Time–Frequency Pretraining.** We introduce Masked Temporal-Frequency Modeling (MTFM), a novel pretraining objective that enforces the joint reconstruction of original signals in both the time and frequency domains. This dual-domain paradigm enables the model to internalize complementary physical dynamics of brain activity, thereby enriching its unified representations.

2 Related Work

Brain Foundation Models. Foundation models have significantly advanced brain AI by learning generalizable representations from large-scale neuroimaging data. In fMRI, pioneering works such as BrainLM [12] and BrainMass [13] employ self-supervised learning to model BOLD time series. Subsequent approaches like Brain-JEPA [14] further advanced this direction by employing joint-embedding predictive architectures to improve spatiotemporal representation learning. Parallel advancements in EEG have led to models such as BENDR [15], MMM [16] and BrainBERT [17], which demonstrated the efficacy of contrastive and masked modeling objectives for capturing temporal dynamics in EEG signals. Recently, LaBraM [2] and CbraMod [3] have utilized massive data to enhance general-purpose EEG representation learning. Despite these advances, those foundation models remain confined to single modalities, failing to leverage the complementary characteristics across functional neuroimaging modalities.

Heterogeneous Foundation Models. More recently, a growing line of work has begun to address heterogeneity by jointly modeling multiple signal types within unified foundation frameworks. BIOT [18] pioneered this direction by pretraining on biosignals, specifically EEG and ECG. Within the electrophysiological domain, Brainwave [19] and PopT [20] learns joint representations across variable channel configurations for both intracranial and scalp EEG, while BrainOmni [21] extends to jointly model EEG and MEG signals. Beyond purely electrophysiological modalities, BrainHarmonix [4] bridges structural and functional neuroimaging by combining fMRI and MRI. Similarly, PhysioOmni [22] decouples homogeneous and heterogeneous features across multimodal physiological signals, including EEG, ECG, EOG and EMG. However, these approaches fall short of capturing the full spectrum of brain activity, missing the opportunity to harmonize the high spatial fidelity of hemodynamic signals with the high temporal resolution of electromagnetic fields. This gap motivates a unified framework for integrating heterogeneous brain signals across fundamentally different physical measurement principles.

3 Method

In this work, we propose **Brain-OF**, a heterogeneous pretraining framework (Figure 1) designed as a unified brain foundation model that supports both unimodal and multimodal data. The framework is explicitly designed to achieve broad generalization across four key dimensions: *cross-modality*, *cross-task*, *cross-site* and *cross-subject* settings. To address the intrinsic heterogeneity of brain signals, Brain-OF introduces an *Any-Resolution Neural Signal Sampler*, which projects inputs with diverse spatial, temporal, and spectral resolutions into a shared semantic space. To further model heterogeneous semantics while mitigating the allocation of attention to irrelevant context, the backbone integrates *DINT attention* with a *Sparse Mixture of Experts*. Finally, to learn informative representations, we propose *Masked Temporal-Frequency Modeling*, which jointly captures complementary knowledge in both the temporal and frequency domains.

3.1 Model Architecture

In this section, the set of modalities is defined as $\mathcal{M} = \{M_{\text{fMRI}}, M_{\text{EEG}}, M_{\text{MEG}}\}$. For any given modality $M_i \in \mathcal{M}$, the normalized brain signal is represented as a matrix $S \in \mathbb{R}^{N \times T}$, where N denotes the number of channels (for M/EEG) or regions of interest (ROIs; for fMRI), and T represents the temporal length of the signal.

Brain Signal Encoder. The primary challenge in handling our heterogeneous corpus, spanning diverse modalities, sites, tasks and devices, is the **structural variability** in channel/ROI counts N and temporal lengths T. To standardize the normalized brain signal S, we adopt a patching and flattening strategy using a fixed time window $\mathcal{T}$. Specifically, S is segmented into non-overlapping patches of length $\mathcal{T}$ along the temporal dimension for each channel/ROI, then flattened into a consistent sequence $X_c \in \mathbb{R}^{L \times \mathcal{T}}$, where $L = N \cdot \lceil \frac{T}{\mathcal{T}} \rceil +$ padding tokens. The sequence is padded to a maximum length of 2,048 to ensure uniform input across modalities. Each patch in X_c is subsequently processed by a temporal encoder comprising stacked 1D convolutional layers to capture local temporal dynamics, followed by RMSNorm and GELU activation. This yields a sequence of patch embeddings $X \in \mathbb{R}^{L \times D}$, where D is the embedding dimension:

$$X = \{x_{\text{patch}_1}, x_{\text{patch}_2}, \ldots, x_{\text{patch}_{N \cdot \lceil T/\mathcal{T} \rceil}}, \ldots, x_{\text{padding}_L}\}$$

An attention mask is generated to exclude zero-padded patches from attention computations, eliminating spurious interactions and ensuring training stability.

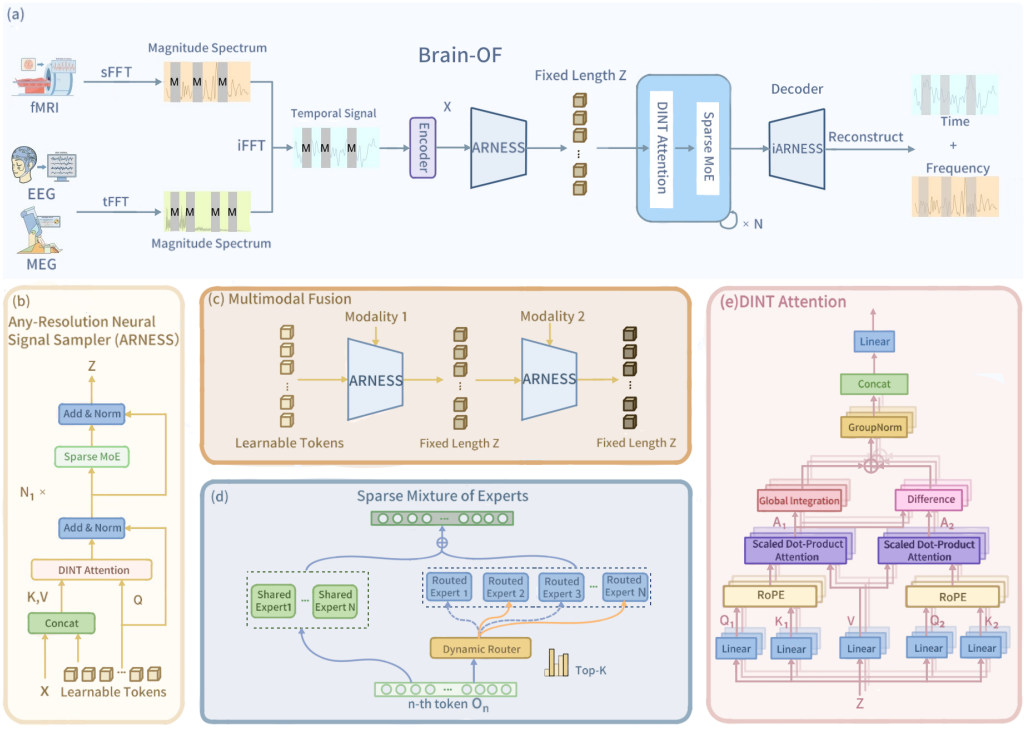


Figure 1: **Overview of the Brain-OF Architecture.** Brain-OF is an omnifunctional foundation model jointly pretrained on fMRI, EEG and MEG. **(a) Pretraining pipeline:** The original signals are masked in both the temporal and frequency domains to encourage the model to jointly learn coupled time–frequency representations. **(b) ARNESS** projects arbitrary resolution signals into a unified semantic space and also serves as a decoder to inverse-sample latent representations for reconstruction during pretraining. **(c) Multimodal Fusion** serially resamples multiple unimodal sequences via ARNESS without introducing modality-specific fusion branches. **(d) Sparse MoE** addresses semantic heterogeneity by dynamically routing tokens to specialized experts while extracting modality-invariant knowledge through shared experts. **(e) DINT Attention** mitigates attention allocation to irrelevant contextual information. Here, sFFT and tFFT denote Fast Fourier Transform along spatial and temporal dimensions; iFFT denotes the inverse Fast Fourier Transform; and M indicates the masking.

Any-Resolution Neural Signal Sampler. Despite the structural unification of the input into sequence X, intrinsic differences in temporal, spatial and spectral resolutions persist across modalities, resulting in significant **semantic heterogeneity**. To bridge this gap and enable a single model backbone to efficiently process diverse unimodal or multimodal signals, we introduce the **A**ny-**R**esolution **Ne**ural **S**ignal **S**ampler (ARNESS). As illustrated in Figure 1b, ARNESS employs Perceiver-style cross-attention [23] to project the heterogeneous sequence X into a shared semantic space via a fixed number of learnable latent tokens $Z \in \mathcal{R}^{C\times D}$ ($C = 128$ in our implementation). Specifically, X and Z are concatenated to form the Keys and Values, with Z serving as the Queries in cross-attention, thereby resampling essential semantic information from the high-dimensional X into a compact, fixed-length Z:

$$ARNESS(X, Z) = CrossAttn(Q = Z, KV = [X; Z], Attn_Mask)$$

This approach enables the subsequent backbone to handle all inputs uniformly, despite inherent heterogeneity. Lower-resolution modalities can potentially benefit by preserving the meaningful semantic space shaped by higher-resolution modalities. It also dramatically reduces the number of tokens for high-dimension brain signals, significantly lowering pretraining computational costs.

Moreover, ARNESS naturally supports multimodal fusion (Figure 1c) by sequentially resampling multiple unimodal sequences into a single fixed-length representation **without requiring modality-specific fusion branches**:

$$Z^{(i+1)} = ARNESS(X^{(i)}, Z^{(i)}), \quad i \in \mathcal{M}$$

DINT Attention Mechanism. The vanilla self-attention mechanism often overallocates attention to irrelevant context, resulting in "attention noise." This issue is particularly detrimental in neuroimaging, where residual noise in preprocessed brain signals can exacerbate this phenomenon, leading to coarse attention patterns and degraded performance [24]. To mitigate this, we introduce the **D**ifferential-**Int**egral (DINT) Attention mechanism [25], which extends Differential Attention [26] to capture precise local dependencies while amplifying global relevance and suppressing noise in the semantic space Z.

The differential component reduces noise by computing the difference between two attention distributions. The query and key projections from Z are split into halves, (Q_1, Q_2) and (K_1, K_2), respectively. The differential attention matrix A_{diff} is defined as:

$$A_{\text{diff}} = \text{softmax}\left(\frac{Q_1 K_1^\top}{\sqrt{d}}\right) - \lambda \cdot \text{softmax}\left(\frac{Q_2 K_2^\top}{\sqrt{d}}\right)$$

where the learnable scalar λ is reparameterized as $\lambda = \exp(\lambda_{q_1} \cdot \lambda_{k_1}) - \exp(\lambda_{q_2} \cdot \lambda_{k_2}) + \lambda_{\text{init}}$. λ_{init} is a constant initialization value and $\lambda_{q,k}$ are learnable parameters. Following empirical settings [26], we set $\lambda_{\text{init}} = 0.8 - 0.6 \cdot \exp(-0.3 \cdot (l-1))$, where l denotes the layer depth.

To capture global dependencies, the integral component computes a global importance score by averaging the columns of the first attention map A_1 (from Q_1, K_1):

$$G = \frac{1}{C} \sum_{i=1}^{C} A_1[i, j]$$

yielding $G \in \mathbb{R}^{1 \times C}$ (C denotes the number of tokens, with $C = 128$ in the backbone). To align dimensions for combination, G is expanded along the query dimension to form $G_{\text{expanded}} \in \mathbb{R}^{C \times C}$. The final output integrates local (differential) and global (integral) semantics:

$$\text{DINTAttn}(Z) = (A_{\text{diff}} + \lambda \cdot G_{\text{expanded}})\, V$$

In addition, G_{expanded} ensures row sums of one, guaranteeing numerical stability. Rotary Position Embeddings (RoPE) [27] are used in our model to encode relative positional information within the sequence.

Sparse Mixture of Experts. Our preliminary experiments using dense feedforward layers revealed a *modality seesaw* effect, where improvements for one modality (e.g., fMRI) degraded performance for others (EEG/MEG). This arises from the diverse semantic features of fMRI, EEG and MEG within the sequence Z. To address this, we adopt a **S**parse **M**ixture **o**f **E**xperts (SMoE) architecture, which excels at modeling heterogeneous semantic spaces. Specifically, we employ a combination of shared experts for modality-invariant representations and routed experts for modality-specific information. Each expert is implemented via a Feed-Forward Network (FFN). For the n-th token O_n from the DINT attention output, the SMoE layer is defined as:

$$\text{SMoE}(O_n) = O_n + \sum_{i=1}^{N_s} \text{FFN}_i^{(s)}(O_n) + \sum_{i \in \mathcal{I}_n} g_{i,n} \text{FFN}_i^{(r)}(O_n)$$

where N_s is the number of shared experts, and $\mathcal{I}_n$ denotes the indices of the top-k routed experts selected for token O_n. The routing weights $g_{i,n}$ are normalized over the selected routed experts:

$$g_{i,n} = \frac{p_{i,n}}{\sum_{j \in \mathcal{I}_n} p_{j,n}}$$

where $p_{i,n}$ is the raw probability generated by a sigmoid-activated gating network. To prevent expert collapse, we use an auxiliary-loss-free load-balancing strategy [28]. A bias term b_i is added to $p_{i,n}$ prior to top-k selection (i.e., selection is based on $p_{i,n} + b_i$), while the actual routing weights for aggregating routed experts rely solely on $p_{i,n}$. At the end of each training step, b_i is dynamically adjusted by a bias update rate γ: decreased for overloaded experts and increased for underloaded ones. Through the dynamic adjustment, SMoE keeps balanced expert load without additional auxiliary losses.

3.2 Pretraining

Masked Temporal-Frequency Modeling (MTFM). The efficacy of foundation models stems largely from self-supervised pretraining, which facilitates the learning of rich representations by

Table 1: Overview of downstream tasks and evaluation datasets.

Downstream Tasks	Datasets	Modalities	#Chs/ROIs	Durations	#Subjects	#Samples	Labels
Emotion Recognition	SEED-V	EEG	62	1s	20	117,744	5-class
Seizure Detection	CHB-MIT	EEG	22	10s	22	326,993	2-class
Abnormal Detection	TUAB	EEG	19	10s	2,383	409,455	2-class
AD Diagnosis	ADNI	fMRI	246	∼300s	222	618	2-class
ADHD Identification	ADHD-200	fMRI	246	∼360s	926	1,369	2-class
	CamCAN	fMRI	246	520s	652	652	regression
Brain Age Prediction	CamCAN	MEG	306	2s	643	177,044	regression
	CamCAN	MEG + fMRI	-	-	621	171,057	regression
	LEMON	fMRI	246	448s	219	438	2-class
Aging Classification	LEMON	EEG	61	10s	203	17,235	2-class
	LEMON	EEG + fMRI	-	-	194	16,471	2-class

reconstructing masked inputs to capture underlying data distributions. In neuroimaging, brain signals inherently contain complementary information across the temporal, spatial and frequency domains. Consequently, effective representation learning requires a model to internalize coupled features from both spectral and temporal perspectives. To achieve this, we introduce **M**asked **T**emporal–**F**requency **M**odeling (MTFM) for learning informative representations across temporal and frequency domains in all modalities. Given a normalized timecourse of a brain signal S, we first transform it into its frequency domain representation F_S. Due to the distinct characteristics of different modalities, richer spatial information in fMRI and higher temporal resolution in EEG and MEG, we apply modality-specific frequency transformations:

$$F_S = \begin{cases} \mathcal{F}_{\text{spatial}}(S), & \text{if } S \in M_{\text{fMRI}} \\ \mathcal{F}_{\text{temporal}}(S), & \text{if } S \in \{M_{\text{EEG}}, M_{\text{MEG}}\} \end{cases}$$

where $\mathcal{F}$ denotes the Fourier transform applied along the spatial dimension for fMRI or the temporal dimension for EEG and MEG. Here, the spatial FFT over ROIs is not intended to model physical spatial frequencies, but serves as a fixed orthogonal transform to encourage global cross-region structure learning. Within the frequency domain, we apply a random frequency mask M_F, with a masking ratio sampled from a normal distribution. The masked frequency representation is then transformed back to the original signal domain via the inverse Fourier transform. To further enhance temporal modeling, a second masking operation M_T is applied to randomly occlude temporal patches in this spectrally perturbed signal. Following the backbone, a lightweight one-layer ARNESS decoder inverse-samples the latent representations to reconstruct the signal. The model is pretrained to jointly reconstruct both the masked temporal patches and the frequency components, thereby fostering the learning of coupled time–frequency representations.

Loss Function. The pretraining objective combines reconstruction losses in both the temporal and frequency domains:

$$\mathcal{L}_{\text{MTFM}} = (1-\alpha) \cdot \mathcal{L}_{\text{SL1}}(\hat{S}, S) + \alpha \cdot \mathcal{L}_{\text{SL1}}(|\hat{F}_S|, |F_S|)$$

where $\hat{S}$ denotes the reconstructed signal and $\hat{F}_S = \mathcal{F}(\hat{S})$ is its frequency domain representation; $|\cdot|$ indicates the magnitude; and $\alpha = 0.8$ is empirically chosen to balance the temporal and frequency reconstruction terms. We utilize the smooth L1 loss ($\mathcal{L}_{\text{SL1}}$) to ensure stable optimization and robustness to outliers.

4 Experiments

4.1 Datasets

Pre-training. We curate a heterogeneous collection of **37 public neuroimaging datasets** to jointly pretrain Brain-OF across fMRI, EEG and MEG modalities. In total, the pretraining dataset comprises **32,278 unique participants and 5,870,984 samples**. This diverse dataset integrates a wide range of M/EEG channel configurations and both 3T and 7T fMRI recordings from multiracial, multicultural international sites. Such heterogeneity in acquisition protocols, populations and modalities enables the construction of a general-purpose, omnifunctional brain foundation model. Further details on dataset descriptions and preprocessing are provided in Appendix A and Table 5.

Downstream Tasks. We evaluate Brain-OF on seven downstream tasks across 11 experiments spanning **fMRI, EEG and MEG** modalities in both unimodal and multimodal settings. To assess the generalizability of the model on EEG, we include emotion recognition (SEED-V), seizure detection

Table 2: Comparison results of different methods on downstream tasks.

Methods	SEED-V	CHB-MIT	TUAB	ADNI	ADHD-200	CamCAN	CamCAN	LEMON	LEMON
	EEG	EEG	EEG	fMRI	fMRI	fMRI	MEG	fMRI	EEG
	Kappa(%) ↑	BAC(%) ↑	BAC(%) ↑	BAC(%) ↑	BAC(%) ↑	MAE ↓	MAE ↓	BAC(%) ↑	BAC(%) ↑
EEGNet	10.06±1.43	56.58±1.06	76.42±0.36	60.22±2.01	57.66±2.03	15.57±1.79	11.60±0.47	59.12±2.10	78.84±2.89
EEGConformer	17.72±1.74	59.76±1.41	77.58±0.49	65.16±2.45	62.00±2.64	8.56±0.72	10.03±0.21	67.27±4.15	78.94±2.35
ContraWR	19.05±1.88	63.44±0.02	77.46±0.41	59.69±4.97	62.67±1.69	9.15±0.39	9.01±0.40	63.97±1.28	77.42±0.45
CNN-Transformer	20.72±1.83	63.89±0.67	77.77±0.22	62.27±4.87	61.94±0.43	16.00±0.10	15.69±0.09	64.83±1.54	78.68±0.77
FFCL	20.78±2.01	62.62±1.04	78.48±0.38	60.31±2.92	63.20±3.96	**8.50±1.24**	8.83±0.28	62.93±2.86	80.17±1.89
ST-Transformer	10.83±1.21	59.15±1.95	79.66±0.23	59.69±3.58	58.40±2.33	12.39±0.46	10.69±0.38	63.62±3.01	76.95±1.14
BrainNetTF	12.69±0.58	50.98±0.73	75.26±0.40	69.62±2.23	61.72±0.76	9.52±0.37	9.90±0.06	69.77±3.23	60.93±2.72
BrainNetCNN	11.06±1.04	51.87±1.85	74.26±0.87	67.20±1.58	60.55±1.95	12.29±1.29	14.67±0.14	**70.08±1.17**	62.01±0.81
LaBraM	23.86±2.09	70.75±3.58	81.40±0.19	60.26±2.84	56.58±3.22	16.10±0.04	13.66±1.46	56.43±2.40	75.59±2.60
CBraMod	25.69±1.43	73.98±2.84	78.91±0.30	61.01±4.57	60.09±5.66	15.80±0.14	15.71±0.21	53.84±4.40	75.29±6.29
BrainHarmonix-F	7.71±0.92	56.08±0.76	77.94±0.14	62.16±1.15	62.78±2.56	14.81±0.49	8.73±0.09	63.11±1.32	77.95±2.36
Brain-JEPA	3.19±0.21	71.66±1.22	74.25±0.59	61.51±2.10	60.58±2.43	12.28±0.70	11.40±0.28	63.79±2.86	72.87±1.11
Brain-OF Base	23.98±0.29	73.99±1.48	81.88±0.17	68.23±2.07	62.57±1.19	12.09±0.78	8.99±0.14	58.95±2.13	76.05±0.44
Brain-OF Large	24.68±0.37	74.12±1.13	81.90±0.15	69.01±2.62	63.30±0.47	11.71±0.67	8.86±0.21	62.85±4.29	78.04±1.59
Brain-OF Huge	**26.36*±0.17**	**74.87±0.84**	**82.87*±0.18**	**71.79±2.19**	**63.51±1.01**	9.51±0.51	**7.87*±0.22**	61.79±1.29	**80.39±0.59**

Performance is reported as mean ± standard deviation across five random seeds. The best result is highlighted in bold and the second-best result is underlined. * indicates a statistically significant improvement over the best baseline ($p < 0.05$).

(CHB-MIT) and abnormal detection (TUAB) as downstream tasks. For fMRI, two large multi-site datasets are used to evaluate diagnostic performance on neurodegenerative Alzheimer's disease (ADNI) and neurodevelopmental attention deficit hyperactivity disorder (ADHD-200), respectively. In addition, brain age prediction (CamCAN) serves as a regression benchmark to evaluate unimodal MEG, unimodal fMRI, and multimodal MEG+fMRI fusion. A second multimodal dataset (LEMON) is further employed to assess age-group classification in unimodal EEG, unimodal fMRI, and multimodal EEG+fMRI settings. Together, these downstream tasks evaluate Brain-OF's effectiveness on resting-state, task-state and ictal/interictal brain activity, while probing single-modality and multi-modality fusion. We adopt cross-subject/trial splits following [3] for EEG datasets and a stratified split following [29] for fMRI datasets (7:1.5:1.5), ensuring strict subject-wise separation of training, validation and test sets. Detailed dataset configurations are provided in Table 1 and Appendix A.3.

4.2 Experiment Setups

Pre-training Settings. Brain-OF is released in three model variants: a 12-layer Brain-OF Base (47.5M parameters, 21.5M active parameters), a 24-layer Brain-OF Large (331M parameters, 150M active parameters) and a 36-layer Brain-OF Huge (1.7B parameters, 500M active parameters). All models are implemented using Python 3.11.11 and PyTorch 2.4.1+CUDA 12.1. Model training adopts the AdamW optimizer in BF16 precision with a global batch size of 7,168. The optimization follows a cosine-annealing schedule with a peak learning rate of 3×10^{-3}, weight decay of 5×10^{-2}, and optimizer states $\beta = (0.9, 0.95)$ and $\epsilon = 10^{-8}$. Variable-length input signals are transformed and padded to a unified sequence length of 2,048. To stabilize routing, we employ an auxiliary-loss-free update for the router bias with coefficient $\gamma = 10^{-3}$. Pretraining is conducted on up to 56 compute nodes, each equipped with 4 NVIDIA A100 (40GB) GPUs (224 GPUs in total), for approximately six days. Comprehensive hyperparameter configurations are detailed in Appendix J.

Baselines. To provide a comprehensive comparison across modalities, we consider the following 12 baseline models: EEG task-specific models (EEGNet [30], EEGConformer [31], ContraWR [32], CNN-Transformer [33], FFCL [34] and ST-Transformer [35]), fMRI task-specific models (BrainNetTF [29] and BrainNetCNN [36]), EEG foundation models (LaBraM [2] and CBraMod [3]), and fMRI foundation models (BrainHarmonix-F [4] and Brain-JEPA [14]). Additional details on all baselines are provided in Appendix B.

Evaluation Metrics. We employ task-specific evaluation metrics for different downstream tasks. For binary classification tasks, we report Balanced Accuracy (BAC), which is also used as the primary metric for model checkpoint selection. In multi-class classification settings, Cohen's Kappa is adopted as the primary evaluation metric. For regression tasks, performance is assessed using Mean Absolute Error (MAE). All reported results are presented as the mean and standard deviation over five random seeds. For each run, the best model is selected based on the primary metric on the validation set, and final performance is reported on a held-out test set.

Table 3: Ablation studies on downstream tasks.

Methods	SEED-V	CHB-MIT	TUAB	ADNI	ADHD-200	CamCAN	CamCAN	LEMON	LEMON
	EEG	EEG	EEG	fMRI	fMRI	fMRI	MEG	fMRI	EEG
	Kappa(%) ↑	BAC(%) ↑	BAC(%) ↑	BAC(%) ↑	BAC(%) ↑	MAE ↓	MAE ↓	BAC(%) ↑	BAC(%) ↑
Brain-OF (Base)	**23.98**$\pm$0.29	**73.99**$\pm$1.48	**81.88**$\pm$**0.17**	68.23$\pm$2.07	**62.57**$\pm$**1.19**	**12.09**$\pm$**0.78**	8.99$\pm$0.14	58.95$\pm$2.13	76.05$\pm$0.44
Objective Ablation									
- w/o MTFM	21.52$\pm$0.57	71.43$\pm$2.30	81.49$\pm$0.18	65.26$\pm$1.83	60.35$\pm$1.50	12.80$\pm$1.15	**8.96**$\pm$**0.33**	57.18$\pm$1.91	**76.90**$\pm$**0.51**
Modality Ablation									
- w/o fMRI	22.24$\pm$0.27	71.61$\pm$0.90	80.98$\pm$0.19	64.35$\pm$1.77	59.94$\pm$1.78	12.71$\pm$0.51	9.16$\pm$0.13	56.89$\pm$3.68	76.17$\pm$0.54
- w/o EEG	19.35$\pm$0.40	69.07$\pm$1.42	81.86$\pm$0.13	67.37$\pm$1.67	59.08$\pm$1.15	12.30$\pm$0.65	9.17$\pm$0.27	**60.29**$\pm$**2.82**	75.61$\pm$1.05
- w/o MEG	21.62$\pm$0.35	70.93$\pm$1.85	81.50$\pm$0.19	66.68$\pm$2.43	60.13$\pm$0.83	12.25$\pm$0.97	9.36$\pm$0.15	56.68$\pm$1.44	75.30$\pm$1.03
- fMRI-only	14.69$\pm$0.33	67.80$\pm$1.04	80.24$\pm$0.19	**70.10**$\pm$**1.64**	59.81$\pm$1.11	12.36$\pm$0.53	9.41$\pm$0.13	57.39$\pm$3.06	74.48$\pm$0.99
- EEG-only	22.15$\pm$0.53	70.15$\pm$1.18	79.86$\pm$0.29	63.70$\pm$1.79	57.01$\pm$1.35	12.88$\pm$0.23	9.34$\pm$0.09	56.37$\pm$1.70	75.79$\pm$1.24
- MEG-only	20.12$\pm$0.38	67.95$\pm$0.93	81.15$\pm$0.11	63.46$\pm$2.31	57.37$\pm$2.48	12.76$\pm$0.73	9.05$\pm$0.19	57.09$\pm$3.12	75.11$\pm$0.53
- No Pretraining	12.63$\pm$1.43	63.62$\pm$1.58	77.59$\pm$0.27	63.39$\pm$2.36	56.11$\pm$0.96	13.44$\pm$0.85	9.69$\pm$1.07	54.25$\pm$2.64	72.81$\pm$1.38

4.3 Comparison with State-of-the-Art Baselines

Brain-OF is evaluated on a diverse set of dowstream tasks including regression, binary and multi-class classification across EEG, fMRI, and MEG modalities. As detailed in Table 2, Brain-OF Huge achieves state-of-the-art performance, ranking first in 7 out of 9 tasks with an average rank of 2.2 across all evaluated settings. These results demonstrate its strong cross-modal generalization capability and robustness across heterogeneous neuroimaging modalities. We observe a consistent performance improvement as model capacity increases. Scaling from Brain-OF Base to Brain-OF Huge yields monotonic improvements across nearly all tasks. Notably, on the CamCAN (MEG) brain-age regression task, the MAE decreases from 8.99 (Base) to 7.87 (Huge), indicating that larger capacity effectively captures fine-grained, continuous biological variations. An important observation is that task-specific models (e.g., the EEG-based EEGConformer and fMRI-based BrainNetTF) occasionally outperform single-modality foundation models (e.g., the EEG-pretrained LaBraM and fMRI-pretrained BrainHarmonix-F) when tested on modalities different from their pretraining domain. This suggests that foundation models pretrained exclusively on a single modality may over-specialize in modality-specific signal characteristics, leading to limited transferability when applied to other modalities (e.g., EEG-pretrained models on fMRI or MEG tasks). In contrast, Brain-OF's joint multimodal pretraining mitigates this issue by learning shared neural representations while still accommodating modality-specific differences.

4.4 Ablation Studies

To rigorously assess the contributions of the proposed MTFM pretraining objective and the role of heterogeneous modality integration, we conduct comprehensive ablation studies across multiple downstream tasks, as summarized in Table 3.

Effectiveness of MTFM. To isolate the benefit of dual-domain pretraining, we compare Brain-OF (pretrained with MTFM) against a baseline pretrained with a standard Masked Autoencoder objective. As shown in Table 3, replacing MTFM with conventional temporal masking consistently degrades performance across most tasks. This confirms that jointly reconstructing signals in both the time and frequency domains forces the model to capture richer neural dynamics than temporal reconstruction alone. The performance gains are particularly pronounced for tasks requiring complex neural signal decoding, such as SEED-V and CHB-MIT. In contrast, the improvements are relatively modest for more static demographic prediction tasks (e.g., CamCAN and LEMON), where fine-grained frequency dynamics may play a less dominant role.

Cross-Modal Synergy in Joint Pretraining. We examine the contribution of each modality by selectively removing one or two modalities from the fMRI, EEG and MEG pretraining corpus. As shown in Table 3, the omnifunctional Brain-OF achieves stronger performance than its single- and dual-modality counterparts on most downstream tasks. An exception is observed on ADNI, where the fMRI-only variant slightly outperforms the omnifunctional model. This result reflects a potential generalist–specialist trade-off: a modality-specialized model can allocate its full capacity to modality-specific characteristics, whereas Brain-OF is optimized to learn representations that generalize across heterogeneous modalities and tasks. Despite this trade-off, Brain-OF remains competitive on ADNI while achieving more consistent performance across the broader benchmark. Notably, removing the native modality of a downstream task does not always cause a catastrophic performance drop. This suggests that Brain-OF learns robust modality-invariant representations, allowing knowledge transfer

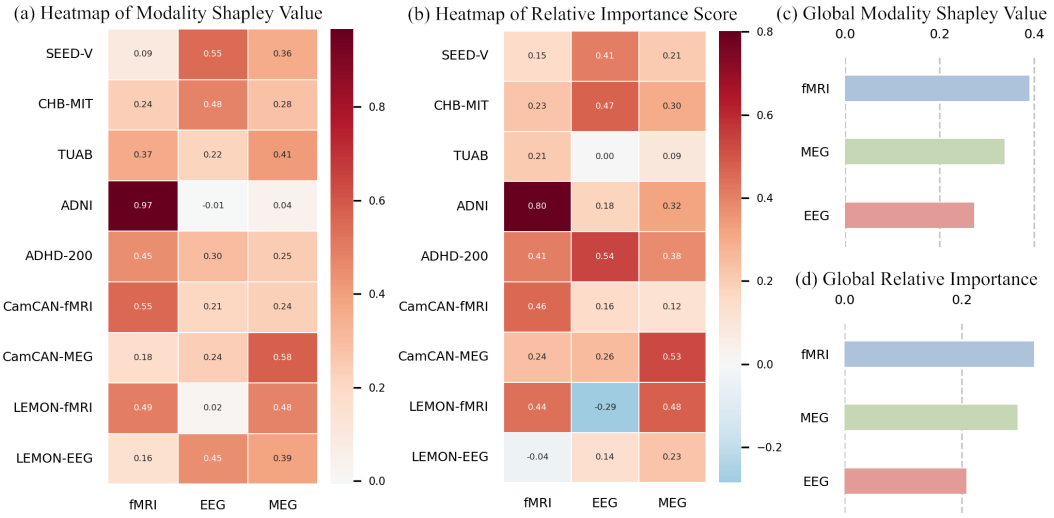


Figure 2: **Modality Importance Analysis.**

across modalities, while modality-specific data still provides additional benefits. Together, these results support the effectiveness of joint cross-modal pretraining beyond unimodal learning.

To further quantify how individual modalities contribute under joint pretraining, we employ two complementary Modality Importance Scores introduced in Appendix C, including the Modality Shapley Value and the Relative Importance Score. We compute both scores for each downstream dataset and further average them across tasks to obtain a global summary, as shown in Figure 2. The local heatmaps in Figure 2(a) and (b) reveal substantial cross-modal dependencies. While several tasks show strong contributions from their native modality, cross-modal contributions in some cases can be equally important or even larger. For example, on the TUAB EEG benchmark, the Shapley values of MEG (0.41) and fMRI (0.37) both exceed that of EEG (0.22), indicating that pretraining on non-EEG modalities provides transferable information for EEG abnormality detection. Because the Modality Shapley Value measures each modality's marginal contribution across all possible modality coalitions under cooperative game theory, it provides a principled view of such cross-modal synergy. More broadly, most off-modality entries in Figure 2(a) are positive, indicating that modalities generally contribute constructively to downstream tasks outside their own domain, although a few negative values suggest occasional cross-modal interference. Figure 2(c) and (d) further aggregate these task-level scores into global rankings. Under both measures, fMRI emerges as the most influential modality overall (global Modality Shapley Value: 0.39; global Relative Importance Score: 0.32), followed by MEG (0.34 and 0.29) and EEG (0.27 and 0.21). These results indicate that joint pretraining benefits from both modality-specific information and transferable cross-modal structure, with fMRI providing the strongest overall contribution across the evaluated tasks.

4.5 Performance Gains via Multimodal Fusion

Beyond unimodal evaluation, we assess multimodal fusion on two paired-modality benchmarks: brain-age prediction on CamCAN (fMRI + MEG) and age-group classification on LEMON (fMRI + EEG). Fusion is implemented via serial resampling with ARNESS, which integrates heterogeneous modalities into a unified fixed-length sequence without additional fusion modules or auxiliary alignment losses. Regarding sequence construction, we adopt an empirical ordering strategy in which the dominant modality (i.e., achieving higher unimodal performance) is placed at the end of the sequence to maximize its influence on the final representation. As shown in Table 4, multimodal fusion consistently outperforms unimodal baselines across nearly all model scales. For both tasks,

Table 4: **Performance benefits of multimodal fusion.**

Datasets	Metrics	Modality	Models		
			Base	Large	Huge
CamCAN	MAE ↓	fMRI-only	12.09$_{\pm 0.78}$	11.71$_{\pm 0.67}$	9.51$_{\pm 0.51}$
		MEG-only	8.99$_{\pm 0.14}$	8.86$_{\pm 0.21}$	**7.87**$_{\pm \mathbf{0.22}}$
		Multimodal Fusion	**8.87**$_{\pm \mathbf{0.09}}$	**8.78**$_{\pm \mathbf{0.07}}$	8.21$_{\pm 0.11}$
LEMON	BAC(%) ↑	fMRI-only	58.95$_{\pm 2.13}$	62.85$_{\pm 4.29}$	61.79$_{\pm 1.29}$
		EEG-only	76.05$_{\pm 0.44}$	78.04$_{\pm 1.59}$	80.39$_{\pm 0.59}$
		Multimodal Fusion	**76.77**$_{\pm \mathbf{0.60}}$	**78.32**$_{\pm \mathbf{1.64}}$	**81.10**$_{\pm \mathbf{0.54}}$

electrophysiological signals (MEG/EEG) achieve stronger unimodal performance than fMRI, reflecting the temporal sensitivity of these benchmarks. Importantly, incorporating fMRI further improves results by injecting complementary information. For example, with Brain-OF Large, fusion reduces the CamCAN MAE to 8.78 (vs. 8.86 with MEG-only) and increases LEMON BAC to 78.32% (vs. 78.04% with EEG-only). These results demonstrate that multimodal fusion in Brain-OF yields improved predictive performance compared to unimodal settings.

4.6 Visualization of Model Explanations

To interpret the representations learned by Brain-OF Base, we employ the occlusion method [37] to quantify the contribution of each channel/ROI to the model's predictive performance across three representative downstream tasks. For TUAB EEG abnormality detection (Figure 3(a)), the dataset contains heterogeneous pathological EEG patterns, including focal and generalized abnormalities, and therefore no single scalp topography is expected to characterize all abnormal recordings. Brain-OF assigns relatively high importance to frontal and central electrodes. This pattern is compatible with generalized or rhythmic slowing, which can be bilateral and anteriorly prominent [38]. Previous work on TUAB further showed that pathological EEG classification relies strongly on low-frequency delta and theta activity, particularly from temporal channels, while clinical reports also contain descriptions of both temporal and rhythmic frontal slowing [39]. For Alzheimer's disease (AD) classification on ADNI (Figure 3(b)), the top 10% most influential ROIs are primarily located in the fusiform gyrus, as well as the middle and inferior temporal gyri. These regions are relevant to visual and memory-related processing and are among the temporal neocortical regions reported to exhibit atrophy associated with progression toward Alzheimer's disease [40, 41]. Thus, the visualization suggests that Brain-OF uses disease-relevant temporal patterns for classification and we regard this correspondence as evidence of biological plausibility. Finally, the sensor-level contribution map for MEG-based age prediction on CamCAN (Figure 3(c)) shows that the most influential sensors are primarily located over temporal and parietal scalp areas. Prior MEG studies have shown that age can be predicted from electrophysiological dynamics that temporal and visual cortical signals contain particularly informative age-related features [42, 43]. The observed sensor distribution is therefore consistent with known age-related changes, further supporting the neurobiological plausibility for MEG. Taken together, these visualizations provide a complementary plausibility check, examining whether Brain-OF makes its predictions on neurobiologically meaningful representations across modalities rather than on obvious artifacts or dataset-specific shortcuts.

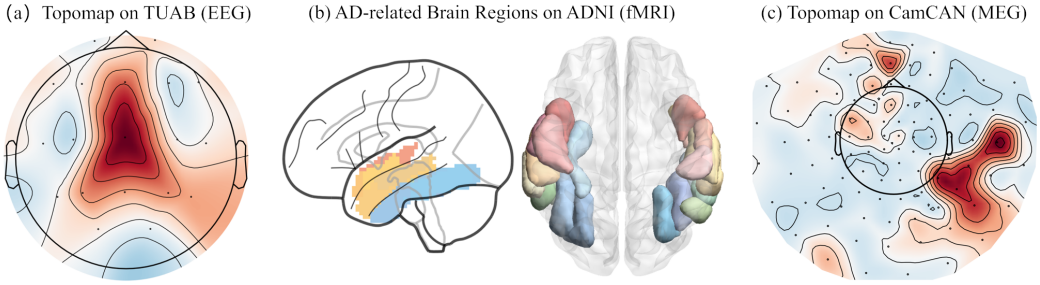


Figure 3: **Visualization of Brain-OF Interpretability.**

5 Conclusion

In this work, we introduce **Brain-OF**, an omnifunctional brain foundation model designed to harness the complementary spatiotemporal characteristics of fMRI, EEG and MEG. To address the heterogeneity of signal resolutions across modalities, we propose the *Any-Resolution Neural Signal Sampler*, which projects diverse inputs into a shared semantic space. Our architecture further addresses semantic complexity through a backbone integrating *DINT attention* and *Sparse Mixture of Experts*, where shared and routed experts capture modality-invariant knowledge and modality-specific representations, respectively. Crucially, to enforce deep understanding of brain dynamics, we propose *Masked Temporal-Frequency Modeling*, a pretraining objective that compels joint reconstruction of signals in both time and frequency domains. Extensive experiments across unimodal and multimodal tasks demonstrate that Brain-OF achieves superior performance, highlighting the benefits of large-scale joint pretraining across heterogeneous brain signals. We hope that Brain-OF paves the way toward general-purpose, scalable and interpretable brain foundation models for both unimodal and cross-modal neuroscience applications.

Acknowledgments and Disclosure of Funding

The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUWELS[44] at Jülich Supercomputing Centre (JSC). Data were provided (in part) by the Donders Institute for Brain, Cognition and Behaviour at Radboud University (Nijmegen, The Netherlands) and the Cambridge Centre for Ageing and Neuroscience (CamCAN).

References

[1] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[2] Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*, 2024.

[3] 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.

[4] Zijian Dong, Ruilin Li, Joanna Su Xian Chong, Niousha Dehestani, Yinghui Teng, Yi Lin, Zhizhou Li, Yichi Zhang, Yapei Xie, Leon Qi Rong Ooi, et al. Brain harmony: A multimodal foundation model unifying morphology and function into 1d tokens. *arXiv preprint arXiv:2509.24693*, 2025.

[5] Jarod Lévy, Hubert Banville, Jérémy Rapin, Jean-Remi King, Thomas Moreau, and Stéphane d'Ascoli. Dance: Detect and classify events in eeg. *arXiv preprint arXiv:2605.10688*, 2026.

[6] Nikos K Logothetis and Josef Pfeuffer. On the nature of the bold fmri contrast mechanism. *Magnetic resonance imaging*, 22(10):1517–1531, 2004.

[7] Bin He, Abbas Sohrabpour, Emery Brown, and Zhongming Liu. Electrophysiological source imaging: a noninvasive window to brain dynamics. *Annual review of biomedical engineering*, 20(1):171–196, 2018.

[8] James W Wheless, Eduardo Castillo, Vijay Maggio, Howard L Kim, Joshua I Breier, Panagiotis G Simos, and Andrew C Papanicolaou. Magnetoencephalography (meg) and magnetic source imaging (msi). *The neurologist*, 10(3):138–153, 2004.

[9] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[10] Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. Pre-trained models: Past, present and future. *Ai Open*, 2: 225–250, 2021.

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

[12] Josue Ortega Caro, Antonio H de O 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.

[13] 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*, 43(11):4004–4016, 2024.

[14] Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Chong, Fang Ji, Nathanael Tong, Christopher Chen, and Juan Helen Zhou. Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. *Advances in Neural Information Processing Systems*, 37:86048–86073, 2024.

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

[16] Ke Yi, Yansen Wang, Kan Ren, and Dongsheng Li. Learning topology-agnostic eeg representations with geometry-aware modeling. *Advances in Neural Information Processing Systems*, 36: 53875–53891, 2023.

[17] Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Brainbert: Self-supervised representation learning for intracranial recordings. *arXiv preprint arXiv:2302.14367*, 2023.

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

[19] Zhizhang Yuan, Fanqi Shen, Meng Li, Yuguo Yu, Chenhao Tan, and Yang Yang. Brainwave: A brain signal foundation model for clinical applications. *arXiv preprint arXiv:2402.10251*, 2024.

[20] Geeling Chau, Christopher Wang, Sabera Talukder, Vighnesh Subramaniam, Saraswati Soedarmadji, Yisong Yue, Boris Katz, and Andrei Barbu. Population transformer: Learning population-level representations of neural activity. *ArXiv*, pages arXiv–2406, 2025.

[21] Qinfan Xiao, Ziyun Cui, Chi Zhang, Siqi Chen, Wen Wu, Andrew Thwaites, Alexandra Woolgar, Bowen Zhou, and Chao Zhang. Brainomni: A brain foundation model for unified eeg and meg signals. *arXiv preprint arXiv:2505.18185*, 2025.

[22] Wei-Bang Jiang, Xi Fu, Yi Ding, and Cuntai Guan. Towards robust multimodal physiological foundation models: Handling arbitrary missing modalities. *arXiv preprint arXiv:2504.19596*, 2025.

[23] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736, 2022.

[24] Hanning Guo, Farah Abdellatif, Yu Fu, N Jon Shah, Abigail Morrison, and Jürgen Dammers. Xaiguiformer: explainable artificial intelligence guided transformer for brain disorder identification. In *The Thirteenth International Conference on Learning Representations*, 2025.

[25] Yueyang Cang, Yuhang Liu, Xiaoteng Zhang, Erlu Zhao, and Li Shi. Dint transformer. *arXiv preprint arXiv:2501.17486*, 2025.

[26] Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer. *arXiv preprint arXiv:2410.05258*, 2024.

[27] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[28] Lean Wang, Huazuo Gao, Chenggang Zhao, Xu Sun, and Damai Dai. Auxiliary-loss-free load balancing strategy for mixture-of-experts. *arXiv preprint arXiv:2408.15664*, 2024.

[29] 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.

[30] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

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

[32] Chaoqi Yang, Cao Xiao, M Brandon Westover, and Jimeng Sun. Self-supervised electroencephalogram representation learning for automatic sleep staging: model development and evaluation study. *JMIR AI*, 2(1):e46769, 2023.

[33] Wei Yan Peh, Yuanyuan Yao, and Justin Dauwels. Transformer convolutional neural networks for automated artifact detection in scalp eeg. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3599–3602. IEEE, 2022.

[34] Hongli Li, Man Ding, Ronghua Zhang, and Chunbo Xiu. Motor imagery eeg classification algorithm based on cnn-lstm feature fusion network. *Biomedical signal processing and control*, 72:103342, 2022.

[35] Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie. Transformer-based spatial-temporal feature learning for eeg decoding. *arXiv preprint arXiv:2106.11170*, 2021.

[36] Jeremy Kawahara, Colin J Brown, Steven P Miller, Brian G Booth, Vann Chau, Ruth E Grunau, Jill G Zwicker, and Ghassan Hamarneh. Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038–1049, 2017.

[37] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[38] Jeffrey W Britton, Lauren C Frey, Jennifer L Hopp, Pearce Korb, Mohamad Z Koubeissi, William E Lievens, Elia M Pestana-Knight, and Erk K St Louis. Electroencephalography (eeg): An introductory text and atlas of normal and abnormal findings in adults, children, and infants. 2016.

[39] R Tibor Schirrmeister, Lukas Gemein, Katharina Eggensperger, Frank Hutter, and Tonio Ball. Deep learning with convolutional neural networks for decoding and visualization of eeg pathology. *arXiv e-prints*, pages arXiv–1708, 2017.

[40] Niklas Vockert, Judith Machts, Luca Kleineidam, Aditya Nemali, Enise I Incesoy, Jose Bernal, Hartmut Schütze, Renat Yakupov, Oliver Peters, Daria Gref, et al. Cognitive reserve against alzheimer's pathology is linked to brain activity during memory formation. *Nature Communications*, 15(1):9815, 2024.

[41] A Convit, J De Asis, MJ De Leon, CY Tarshish, S De Santi, and H Rusinek. Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to alzheimer's disease. *Neurobiology of aging*, 21(1):19–26, 2000.

[42] Christina Stier, Elio Balestrieri, Jana Fehring, Niels K Focke, Andreas Wollbrink, Udo Dannlowski, and Joachim Gross. Temporal autocorrelation is predictive of age—an extensive meg time-series analysis. *Proceedings of the National Academy of Sciences*, 122(8): e2411098122, 2025.

[43] Jeyeon Lee, Brian J Burkett, Hoon-Ki Min, Matthew L Senjem, Emily S Lundt, Hugo Botha, Jonathan Graff-Radford, Leland R Barnard, Jeffrey L Gunter, Christopher G Schwarz, et al. Deep learning-based brain age prediction in normal aging and dementia. *Nature aging*, 2(5): 412–424, 2022.

[44] Jülich Supercomputing Centre. JUWELS Cluster and Booster: Exascale Pathfinder with Modular Supercomputing Architecture at Juelich Supercomputing Centre. *Journal of large-scale research facilities*, 7(A183), 2021. doi: 10.17815/jlsrf-7-183. URL `http://dx.doi.org/10.17815/jlsrf-7-183`.

[45] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Lauri Parkkonen, and Matti S Hämäläinen. Mne software for processing meg and eeg data. *NeuroImage*, 86:446–460, 2014.

[46] Oscar Esteban, Christopher J Markiewicz, Ross W Blair, Craig A Moodie, A Ilkay Isik, Asier Erramuzpe, James D Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, et al. fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, 16(1):111–116, 2019.

[47] Lingzhong Fan, Hai Li, Junjie Zhuo, Yu Zhang, Jiaojian Wang, Liangfu Chen, Zhengyi Yang, Congying Chu, Sangma Xie, Angela R Laird, et al. The human brainnetome atlas: a new brain atlas based on connectional architecture. *Cerebral cortex*, 26(8):3508–3526, 2016.

[48] Jingjing Chen, Xiaobin Wang, Chen Huang, Xin Hu, Xinke Shen, and Dan Zhang. A large finer-grained affective computing eeg dataset. *Scientific Data*, 10(1):740, 2023.

[49] Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Transactions on biomedical engineering*, 51(6):1034–1043, 2004.

[50] Nahian S. Chowdhury, Chuan Bi, Andrew J. Furman, Alan KI Chiang, Patrick Skippen, Emily Si, Samantha K Millard, Sarah M. Margerison, Darrah Spies, Michael L. Keaser, Joyce T. Da Silva, Shuo Chen, Siobhan M. Schabrun, and David A. Seminowicz. "predict". 2025. doi: doi:10.18112/openneuro.ds005486.v1.0.1.

[51] Nahian S Chowdhury, Chuan Bi, Andrew J Furman, Alan KI Chiang, Patrick Skippen, Emily Si, Samantha K Millard, Sarah M Margerison, Darrah Spies, Michael L Keaser, et al. Predicting individual pain sensitivity using a novel cortical biomarker signature. *JAMA neurology*, 82(3): 237–246, 2025.

[52] Stephan Getzmann, Patrick D Gajewski, Daniel Schneider, and Edmund Wascher. Resting-state eeg data before and after cognitive activity across the adult lifespan and a 5-year follow-up. *Scientific Data*, 11(1):988, 2024.

[53] Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015. doi: 10.1109/TAMD.2015.2431497.

[54] Wei-Long Zheng and Bao-Liang Lu. A multimodal approach to estimating vigilance using eeg and forehead eog. *Journal of Neural Engineering*, 14(2):026017, 2017. URL `http://stacks.iop.org/1741-2552/14/i=2/a=026017`.

[55] Wei-Bang Jiang, Xuan-Hao Liu, Wei-Long Zheng, and Bao-Liang Lu. Seed-vii: A multimodal dataset of six basic emotions with continuous labels for emotion recognition. *IEEE Transactions on Affective Computing*, 16(2):969–985, 2025. doi: 10.1109/TAFFC.2024.3485057.

[56] Yun Luo, Wei Liu, Hanqi Li, Yong Lu, and Bao-Liang Lu. A cross-scenario and cross-subject domain adaptation method for driving fatigue detection. *Journal of Neural Engineering*, 21(4): 046004, jul 2024. doi: 10.1088/1741-2552/ad546d. URL `https://dx.doi.org/10.1088/1741-2552/ad546d`.

[57] Xuan-Hao Liu, Yan-Kai Liu, Yansen Wang, Kan Ren, Hanwen Shi, Zilong Wang, Dongsheng Li, Bao-Liang Lu, and Wei-Long Zheng. EEG2Video: Towards decoding dynamic visual perception from EEG signals. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[58] Ziyi Li, Le-Yan Tao, Rui-Xiao Ma, Wei-Long Zheng, and Bao-Liang Lu. Investigating the effects of sleep conditions on emotion responses with eeg signals and eye movements. *IEEE Transactions on Affective Computing*, 2025. doi: 10.1109/TAFFC.2025.3572504.

[59] Hanneke Van Dijk, Guido Van Wingen, Damiaan Denys, Sebastian Olbrich, Rosalinde Van Ruth, and Martijn Arns. The two decades brainclinics research archive for insights in neurophysiology (tdbrain) database. *Scientific data*, 9(1):333, 2022.

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

[61] Banghua Yang, Fenqi Rong, Yunlong Xie, Du Li, Jiayang Zhang, Fu Li, Guangming Shi, and Xiaorong Gao. A multi-day and high-quality eeg dataset for motor imagery brain-computer interface. *Scientific Data*, 12(1):488, 2025.

[62] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667, 2014.

[63] Adriana Di Martino, David O'connor, Bosi Chen, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Joshua H Balsters, Leslie Baxter, Anita Beggiato, Sylvie Bernaerts, et al. Enhancing studies of the connectome in autism using the autism brain imaging data exchange ii. *Scientific data*, 4(1):1–15, 2017.

[64] M.G. Jansen, M.P. Zwiers, J.P. Marques, K. Chan, J.S. (Jitse) Amelink, M. Altgassen, J.M. Oosterman, and D.G. Norris. The Advanced BRain Imaging database on ageing and Memory (ABRIM) study: MRI collection, 2023. URL `https://doi.org/10.34973/7q0a-vj19`.

[65] Lukas Snoek, Maite M van der Miesen, Tinka Beemsterboer, Andries Van Der Leij, Annemarie Eigenhuis, and H Steven Scholte. The amsterdam open mri collection, a set of multimodal mri datasets for individual difference analyses. *Scientific data*, 8(1):85, 2021.

[66] Jianqiao Ge, Guoyuan Yang, Meizhen Han, Sizhong Zhou, Weiwei Men, Lang Qin, Bingjiang Lyu, Hai Li, Haobo Wang, Hengyi Rao, et al. Increasing diversity in connectomics with the chinese human connectome project. *Nature Neuroscience*, 26(1):163–172, 2023.

[67] Xi-Nian Zuo, Jeffrey S Anderson, Pierre Bellec, Rasmus M Birn, Bharat B Biswal, Janusch Blautzik, John Breitner, Randy L Buckner, Vince D Calhoun, F Xavier Castellanos, et al. An open science resource for establishing reliability and reproducibility in functional connectomics. *Scientific data*, 1(1):1–13, 2014.

[68] Denise C Park, Joseph P Hennessee, Evan T Smith, Micaela Y Chan, Xi Chen, Marianna Dakanali, Michelle E Farrell, Peiying Liu, Hanzhang Lu, Neil Rofsky, et al. The dallas lifespan brain study: A comprehensive adult lifespan data set of brain and cognitive aging. *Scientific Data*, 12(1):846, 2025.

[69] R Nathan Spreng, Roni Setton, Udi Alter, Benjamin N Cassidy, Bri Darboh, Elizabeth DuPre, Karin Kantarovich, Amber W Lockrow, Laetitia Mwilambwe-Tshilobo, Wen-Ming Luh, et al. Neurocognitive aging data release with behavioral, structural and multi-echo functional mri measures. *Scientific Data*, 9(1):119, 2022.

[70] Samuel A Nastase, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, Christopher J Honey, Yaara Yeshurun, Mor Regev, et al. The "narratives" fmri dataset for evaluating models of naturalistic language comprehension. *Scientific data*, 8(1):250, 2021.

[71] Russell H Tobe, Anna MacKay-Brandt, Ryan Lim, Melissa Kramer, Melissa M Breland, Lucia Tu, Yiwen Tian, Kristin Dietz Trautman, Caixia Hu, Raj Sangoi, et al. A longitudinal resource for studying connectome development and its psychiatric associations during childhood. *Scientific data*, 9(1):300, 2022.

[72] Rotem Botvinik-Nezer, Bogdan Petre, Marta Ceko, Martin A Lindquist, Naomi P Friedman, and Tor D Wager. Placebo treatment affects brain systems related to affective and cognitive processes, but not nociceptive pain. *Nature communications*, 15(1):6017, 2024.

[73] Gabriëlla AM Blokland, Katie L McMahon, Paul M Thompson, Nicholas G Martin, Greig I de Zubicaray, and Margaret J Wright. Heritability of working memory brain activation. *Journal of Neuroscience*, 31(30):10882–10890, 2011.

[74] Benjamin Sinclair, Narelle K Hansell, Gabriëlla AM Blokland, Nicholas G Martin, Paul M Thompson, Michael Breakspear, Greig I de Zubicaray, Margaret J Wright, and Katie L McMahon. Heritability of the network architecture of intrinsic brain functional connectivity. *NeuroImage*, 121:243–252, 2015.

[75] Xiao Chen, Bin Lu, Hui-Xian Li, Xue-Ying Li, Yu-Wei Wang, Francisco Xavier Castellanos, Li-Ping Cao, Ning-Xuan Chen, Wei Chen, Yu-Qi Cheng, et al. The direct consortium and the rest-meta-mdd project: towards neuroimaging biomarkers of major depressive disorder. *Psychoradiology*, 2(1):32–42, 2022.

[76] Dongtao Wei, Kaixiang Zhuang, Lei Ai, Qunlin Chen, Wenjing Yang, Wei Liu, Kangcheng Wang, Jiangzhou Sun, and Jiang Qiu. Structural and functional brain scans from the cross-sectional southwest university adult lifespan dataset. *Scientific data*, 5(1):1–10, 2018.

[77] Heejung Jung, Maryam Amini, Bethany J Hunt, Eilis I Murphy, Patrick Sadil, Yaroslav O Halchenko, Bogdan Petre, Zizhuang Miao, Philip A Kragel, Xiaochun Han, et al. Spacetop: A multimodal fmri dataset unifying naturalistic processes with a rich array of experimental tasks. *Scientific Data*, 12(1):1465, 2025.

[78] Russell A Poldrack, Eliza Congdon, William Triplett, KJ Gorgolewski, KH Karlsgodt, JA Mumford, FW Sabb, NB Freimer, ED London, TD Cannon, et al. A phenome-wide examination of neural and cognitive function. *Scientific data*, 3(1):1–12, 2016.

[79] Jason R Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A Shafto, Marie Dixon, Lorraine K Tyler, Richard N Henson, et al. The cambridge centre for ageing and neuroscience (cam-can) data repository: Structural and functional mri, meg, and cognitive data from a cross-sectional adult lifespan sample. *NeuroImage*, 144:262–269, 2017.

[80] 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.

[81] Anahit Babayan, Miray Erbey, Deniz Kumral, Janis D Reinelt, Andrea MF Reiter, Josefin Röbbig, H Lina Schaare, Marie Uhlig, Alfred Anwander, Pierre-Louis Bazin, et al. A mind-brain-body dataset of mri, eeg, cognition, emotion, and peripheral physiology in young and old adults. *Scientific data*, 6(1):1–21, 2019.

[82] Jan-Mathijs Schoffelen, Robert Oostenveld, Nietzsche HL Lam, Julia Uddén, Annika Hultén, and Peter Hagoort. A 204-subject multimodal neuroimaging dataset to study language processing. *Scientific data*, 6(1):17, 2019.

[83] Guiomar Niso, Christine Rogers, Jeremy T Moreau, Li-Yuan Chen, Cecile Madjar, Samir Das, Elizabeth Bock, François Tadel, Alan C Evans, Pierre Jolicoeur, et al. Omega: the open meg archive. *NeuroImage*, 124:1182–1187, 2016.

[84] Carolyn B McNabb, Ian D Driver, Vanessa Hyde, Garin Hughes, Hannah L Chandler, Hannah Thomas, Christopher Allen, Eirini Messaritaki, Carl J Hodgetts, Craig Hedge, et al. Wand: A multi-modal dataset integrating advanced mri, meg, and tms for multi-scale brain analysis. *Scientific Data*, 12(1):220, 2025.

[85] Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 2021.

[86] Ali Hossam Shoeb. *Application of machine learning to epileptic seizure onset detection and treatment*. PhD thesis, Massachusetts Institute of Technology, 2009.

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

[88] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.

[89] ADHD-200 consortium. The adhd-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in systems neuroscience*, 6:62, 2012.

[90] Ming Zhou, Zhengxin Gong, Yuxuan Dai, Yushan Wen, Youyi Liu, and Zonglei Zhen. A large-scale fmri dataset for human action recognition. *Scientific Data*, 10(1):415, 2023.

[91] Nadine Chang, John A Pyles, Austin Marcus, Abhinav Gupta, Michael J Tarr, and Elissa M Aminoff. Bold5000, a public fmri dataset while viewing 5000 visual images. *Scientific data*, 6 (1):49, 2019.

[92] Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

[93] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

A Details of Datasets

In this study, we integrate three heterogeneous functional brain imaging modalities (i.e., fMRI, EEG and MEG) to develop the omnifunctional brain foundation model (Brain-OF). These modalities differ substantially in their temporal, spatial, and spectral resolutions, giving rise to both semantic richness and considerable heterogeneity. Within each modality, further variability arises from heterogeneous acquisition protocols and scanners across international sites, reflected in diverse sampling rates (MEG/EEG), repetition times and field strengths (3 T vs. 7 T for fMRI), number of channels/sensors, recording durations and experimental paradigms (resting vs. task state).

Table 5: Summary of all pretraining datasets, covering 3 heterogeneous functional neuroimaging modalities (fMRI, EEG, MEG) across 37 public datasets, encompassing over 32K unique participants, approximately 5.9M samples, and 2.8 TB of preprocessed data.

Modalities	Datasets	Paradigms	SR/TR	#Chs/ROIs	Durations	#Participants	#Samples
EEG	FACED	task	250Hz	32	10s	123	10,332
	PhysioNet-MI	task	160Hz	64	4s	109	18,697
	PREDICT	rest	5,000Hz	63	10s	159	17,511
	REEG-BACA	rest	1,000Hz	64	10s	608	53,089
	SEED	task	1,000Hz	62	10s	15	14,895
	SEED-DV	task	1,000Hz	62	10s	20	7,644
	SEED-SD	task	1,000Hz	62	10s	40	38,120
	SEED-VIG	task	200Hz	17	8s	21	20,332
	SEED-VII	task	1,000Hz	62	10s	20	26,257
	SEED-VLA/VRW	task	300Hz	18	10s	20	17,287
	TDBRAIN	rest	500Hz	26	10s	1,274	26,422
	TUEG	rest	250Hz	19	10s	13,928	4,562,151
	WBCIC-MI	task	1,000Hz	58	4s	62	40,290
fMRI	ABIDE-I	rest	multi-site	246	-	1,112	1,152
	ABIDE-II	rest	multi-site	246	-	1,114	1,614
	ABRIM	rest & task	1,000ms	246	320s	282	562
	AOMIC-ID1000	task	2,200ms	246	660s	881	881
	AOMIC-PIOP1	rest & task	750ms	246	360s	216	1,236
	AOMIC-PIOP2	rest & task	2,000ms	246	480s	226	896
	CHCP	rest & task	710ms	246	227s	366	3,379
	CoRR	rest	multi-site	246	-	1,629	5,212
	DLBS	rest & task	2,000ms	246	374s	464	6,839
	NADR	rest	3,000ms	246	606s	301	596
	Narratives	task	1,500ms	246	480s	345	1,321
	NKI-RS	rest	1,400ms	246	448s	1,344	16,408
	Paingen placebo	task	460ms	246	132s	395	3,143
	QTIM	rest & task	2,100ms	246	315s	1,202	2,453
	REST-meta-MDD	rest	multi-site	116	-	2,380	2,380
	SALD	rest	2,000ms	246	480s	493	493
	SpaceTop	task	460ms	246	200s	101	11,660
	UCLA CNP	rest & task	2,000ms	246	304s	272	1,969
MEG / Multimodal	CamCAN-MEG	rest & task	1,000Hz	306	2s	456	279,965
	CamCAN-fMRI	rest & task	1,970ms	246	520s	465	1,763
	HCP-YA-MEG	rest & task	2,000Hz	248	2s	89	241,115
	HCP-YA-fMRI	rest & task	720ms	246	230s	1,113	32,281
	LEMON-EEG	rest	2,500Hz	61	10s	144	12,235
	LEMON-fMRI	rest	1,400ms	246	448s	160	320
	MOUS-MEG	rest & task	1,200Hz	275	2s	204	66,322
	MOUS-fMRI	rest & task	1,680ms	246	420s	204	678
	OMEGA	rest	2,400Hz	275	2s	644	164,834
	WAND-MEG	rest & task	1,200Hz	275	2s	170	155,474
	WAND-fMRI	rest & task	2,000ms	246	638s	170	776
Total	**fMRI, EEG, MEG**	**rest & task**	-	-	-	**32,278**	**5,870,984**

Some datasets include multiple sampling rates (SR) or repetition times (TR) due to differences in acquisition protocols across different phases or paradigms (e.g., resting-state vs. task-based scans). For conciseness, we report the primary SR/TR for each dataset in the "SR / TR" column. Similarly, the reported durations primarily correspond to resting-state recordings. For detailed information, please refer to the original publications. Chs: channels. ROIs: regions of interest.

This diversity further supports the construction of a pluralistic pretraining corpus spanning multiracial, multicultural populations, including both healthy individuals and clinical patients. In total, our pretraining dataset aggregates 37 publicly available datasets across fMRI, EEG and MEG, comprising 32,278 unique participants, 5,870,984 samples, and 2.8 TB of preprocessed data. Detailed statistics for each dataset can be found in Table 5.

A.1 Data Preprocessing

If the public datasets provide official preprocessed data, we use them directly. Otherwise, we apply minimal, modality-specific preprocessing pipelines to remove artifacts. Raw EEG and MEG signals are preprocessed using MNE-Python [45], and fMRI data are preprocessed with fMRIPrep [46]. After preprocessing, z-score normalization is applied to all signals to harmonize their value ranges across fMRI, EEG and MEG.

EEG. To enhance stationarity, the initial and final 10 seconds of EEG signals are discarded. Signals are then band-pass filtered (1–80 Hz, zero-phase Hamming-windowed sinc FIR, –6 dB cutoff) and notched at the local power-line frequency (50 or 60 Hz). Filtered signals are resampled to 200 Hz, segmented into non-overlapping 10-second samples for resting-state data, and re-referenced to the common average.

MEG. MEG preprocessing similarly undergoes band-pass filtering (1–80 Hz, zero-phase Hamming-windowed sinc FIR, –6 dB cutoff) and notching at the power-line frequency. Filtered signals are resampled to 200 Hz and segmented into non-overlapping 2-second samples for resting-state recordings.

fMRI. For each BOLD series, a reference volume and its skull-stripped version are estimated. Head-motion parameters (volume-to-reference transforms and corresponding rotations/translations) are computed, followed by slice-timing correction using 3dTshift (AFNI). Functional volumes are co-registered to the T1-weighted anatomical reference via boundary-based registration (9 degrees of freedom) to minimize distortions. Motion correction, susceptibility distortion correction, BOLD-to-T1w, and T1w-to-template (MNI152NLin2009cAsym) warps are concatenated and applied in one interpolation step using antsApplyTransforms (ANTs). Nuisance regressors, including global signal, mean signals from white matter and CSF, tCompCor, aCompCor, framewise displacement, six motion parameters, DVARS, and spike regressors, are regressed out. Residual signals are spatially smoothed (6 mm FWHM Gaussian kernel) and subjected to Wishart filtering. Finally, regional time series are extracted from 246 parcels using the Brainnetome atlas [47].

A.2 Pretraining Datasets

- **FACED** [48] is a large finer-grained affective computing EEG dataset comprising 32-channel recordings at sampling rate of 250 or 1,000 Hz from 123 subjects (ages 17-38). Each participant was instructed to watch 28 emotion-elicitation video clips spanning nine discrete emotional states: amusement, inspiration, joy, tenderness, anger, fear, disgust, sadness and neutral emotion.
- **PhysioNet-MI** [49] is a large-scale motor imagery EEG collections, comprising 64-channel recordings (160 Hz sampling rate) from 109 healthy participants performing left fist, right fist, both fists, and both feet imagery. For pretraining, we use 4-second trials after preprocessing.
- **PREDICT** [50, 51] is a longitudinal pain-related EEG dataset comprising 5-minute, eyes-closed, 63-channel resting-state recordings from 159 participants (ages 18–44) with no history of chronic pain, neurological disorders, or psychiatric conditions. EEG data were collected on Days 0, 2, and 5 of the experimental protocol, providing multiple sessions per participant for identifying individual pain sensitivity in healthy adults.
- **REEG-BACA** [52] is a large-scale longitudinal resting-state EEG dataset collected from 608 participants (ages 20-70), with 208 participants returning for a follow-up session approximately 5 years later. High-density rs-EEG was acquired using a 64-channel elastic cap (10–20 system) at 1,000 Hz sampling rate. Recordings were obtained both before and after a 2-hour battery of cognitive tasks.
- **SEED Series** is a family of EEG datasets designed for affective computing and related cognitive neuroscience research. It includes **SEED** [53], **SEED-VIG** [54], **SEED-VII** [55], **SEED-VLA/VRW** [56], **SEED-DV** [57] and **SEED-SD** [58]. All datasets were

recorded using a 62-channel ESI NeuroScan system at a 1,000 Hz sampling rate during emotion-related paradigms spanning positive, negative and neutral valence, as well as more fine-grained affective states (e.g., amusement, fear, sadness). In total, the SEED Series provides EEG recordings from 136 unique participants, predominantly young adults.

- **TDBRAIN** [59] is a large single-site lifespan EEG database (ages 5–89), comprising resting-state recordings from 1,274 psychiatric patients. The psychophysiological data include 26 channel EEG acquired at a sampling rate of 500 Hz.
- **TUEG** [60] is the largest publicly available collection of clinical EEG dataset, comprising 69,652 recordings from 14,987 subjects across 26,846 sessions collected at Temple University Hospital. The dataset exhibits substantial variability in channel configurations, sampling rates and recording durations, reflecting the heterogeneity of real-world clinical practice. To prevent information leakage and ensure reliable downstream evaluation, we exclude all subjects that appear in the TUAB test set.
- **WBCIC-MI** [61] is a high-quality motor imagery EEG dataset comprising recordings from 62 healthy participants (ages 17–30) across three sessions. For pretraining, we incorporate the officially released preprocessed signals: 4-second trials recorded with 58 EEG channels.
- **ABIDE** is an international collaborative initiative that aggregates resting-state fMRI, structural MRI and phenotypic data collected independently across more than 24 imaging centers worldwide. It contains two large-scale collections: **ABIDE-I** [62], which includes 1,112 participants (ages 7–64; 539 individuals with autism spectrum disorder and 573 controls), and **ABIDE-II** [63], which includes 1,114 participants (521 individuals with autism spectrum disorder and 593 controls).
- **ABRIM** [64] is a cross-sectional adult lifespan MRI collection comprising structural and functional MRI scans from 295 healthy participants (ages 18–80), along with cognitive assessments, proxy measures of cognitive reserve, and 7-day actigraphy recordings of sleep-wake rhythms. After removing corrupted and missing data, a total of 282 subjects are included in our pretraining dataset.
- **AOMIC** [65] is one of the largest and richest open-access 3 T neuroimaging resources focused on young adults. It comprises three complementary datasets: **ID1000** (N = 928, ages 19–26), **PIOP1** (N = 216, age 18–26) and **PIOP2** (N = 226, ages 18–26), together providing 1,370 unique participants. Each sub-cohort includes high-resolution T1-weighted structural MRI, diffusion-weighted imaging, multiple resting-state runs, and an exceptionally broad battery of task-based fMRI paradigms (naturalistic movie watching, emotion processing, working memory, face recognition, cognitive conflict and response inhibition).
- **CHCP** [66] is the Chinese Human Connectome Project, comprising neuroimaging and behavioral data from 366 healthy adults living in an Eastern culture. Its acquisition protocols closely mirror the original HCP-YA, including high-resolution T1w/T2w structural MRI, resting-state and task-based fMRI (seven tasks), and high-angular-resolution diffusion MRI, as well as behavioral data based on Chinese population.
- **CoRR** [67] is the largest publicly available test-retest resting-state fMRI consortium, aggregating data from 36 independent research groups across 18 international sites. The dataset comprises 5,093 resting-state fMRI scans from 1,629 unique participants (ages 6–88), with many individuals providing multiple sessions. This multi-site, multi-scanner design introduces natural acquisition variability highly beneficial for learning robust representations.
- **DLBS** [68] is a longitudinal adult lifespan neuroimaging dataset designed to characterize brain–cognition relationships across aging. A total of 464 participants (ages 21–89 at baseline) completed the first assessment, with 338 and 224 participants returning for second and third assessment epochs, respectively, at intervals of approximately 3.5–5 years. Each epoch includes an extensive neuropsychological battery, questionnaires assessing physical health, psychosocial factors and brain health, as well as multimodal MRI acquisitions: T1-weighted structural imaging, diffusion-weighted imaging, hypercapnia imaging, arterial spin labeling and four functional MRI scans (resting-state and task-based fMRI), totaling 6,893 fMRI sessions across all visits.
- **NADR** [69] is a multisite neuroimaging dataset designed to characterize age-related differences in brain function. It includes 301 healthy adults, comprising 181 young adults (ages 18–34) and 120 older adults (ages 60–89). Each participant underwent two 10-minute

multi-echo resting-state fMRI runs and a high-resolution T1-weighted MPRAGE structural scan acquired on 3T MRI scanners across two sites.

- **Narratives** [70] is a naturalistic paradigm dataset comprising 891 fMRI runs from 345 healthy adults (age 18–53 years) who listened to 27 different spoken stories. These stories vary widely in length, speaker, and content, together providing 4.6 hours of unique auditory narrative stimulation (43,000 words).
- **NKI-RS** [71] is a large-scale longitudinal lifespan neuroimaging initiative comprising 1,344 participants, designed to characterize developmental and aging trajectories across the human lifespan. The dataset provides comprehensive psychiatric, medical, behavioral and cognitive phenotyping, along with multimodal brain imaging across multiple follow-up visits (resting-state fMRI, diffusion MRI, structural morphometric MRI and arterial spin labeling), genetic data and actigraphy measures.
- **Paingen placebo** [72] is a large-scale, high-quality pain and placebo analgesia fMRI collection comprising 395 healthy adults (age 30–43 years). Each participant completed four fMRI runs during a well-controlled thermal pain and placebo manipulation paradigm, plus anatomical scans and field maps.
- **QTIM** [73, 74] is one of the largest twin neuroimaging cohorts worldwide. It comprises resting-state and task-based fMRI from 1,202 young adult twins and siblings (ages 12–30 at baseline), including a test-retest subsample (N = 78, 3.5-month interval) and longitudinal adolescent twin cohorts scanned at ages 12 and 16 (N = 88 each), with a subset rescanned four years later (N = 24 at both 12 and 16; N = 38 at 16 and 20).
- **REST-meta-MDD** [75] is a large-scale resting-state fMRI consortium for major depressive disorder (MDD), aggregating data from 25 research groups across 17 hospitals in China. It comprises 2,428 resting-state fMRI scans from 1,300 patients with MDD and 1,128 healthy controls, providing a diverse and clinically representative sample for studying MDD alterations in brain function.
- **SALD** [76] is a cross-sectional adult lifespan neuroimaging dataset comprising structural MRI and resting-state fMRI from 493 healthy adults (ages 19–80). All scans were acquired on a 3 T Siemens Trio scanner at the Brain Imaging Center of Southwest University, China.
- **SpaceTop** [77] is a landmark naturalistic and multidimensional neuroimaging dataset comprising 101 healthy young adults, each completing approximately 40 fMRI runs. It uniquely integrates a broad spectrum of experimental paradigms, including cognitive, affective, social and somatic/interoceptive tasks, together with structural T1-weighted imaging, multi-shell diffusion MRI and autonomic physiological recordings.
- **UCLA CNP** [78] is a publicly shared neuroimaging dataset comprising 272 adults (ages 21–50): 130 healthy controls and 142 patients (50 schizophrenia, 49 bipolar disorder, 43 ADHD). The dataset provides an extensive set of imaging modalities, including task-based fMRI across multiple cognitive domains, resting-state fMRI, structural MRI and high–angular-resolution diffusion MRI.
- **CamCAN** [79] is a large multimodal, cross-sectional adult lifespan dataset from 656 healthy adults (ages 18–87). It includes 3T resting-state and task fMRI (movie-watching and sensorimotor task) as well as high-density MEG (306-channel Elekta Neuromag, 1,000 Hz sampling rate) during both resting state and the same sensorimotor task. For pretraining, we incorporate all available fMRI and MEG data except subjects assigned to downstream validation and test sets. This strict subject-level exclusion prevents contamination while maximizing pretraining data volume.
- **HCP Young Adult** [80] includes multiple 3T and 7T fMRI runs as well as MEG recordings from 1,113 young healthy adult (ages 22-35). For pretraining, we incorporate all available functional modalities: 3T resting-state and task fMRI, 7T resting-state and task fMRI and both resting-state and task MEG.
- **LEMON** [81] is a multimodal neuroimaging dataset that includes both fMRI and EEG recordings. Participants underwent a 3T fMRI scan and a 61-channel resting-state EEG acquisition. Due to the scarcity of high-quality multimodal datasets, we also incorporate LEMON as a downstream evaluation task. To avoid contamination in downstream evaluation, subjects overlapping with the LEMON validation and test splits are excluded from the pretraining dataset.

- **MOUS** [82] is a multimodal neuroimaging dataset collected from 204 healthy adults (ages 18-33). The protocol includes whole-brain fMRI and high-density 275-channel MEG acquired during a Dutch language paradigm as well as during resting state. For pretraining, we incorporate all available MEG and fMRI sessions from the MOUS cohort.
- **OMEGA** [83] is a large, open-access collection of resting-state MEG recordings led by the McConnell Brain Imaging Centre (BIC) at the Montreal Neurological Institute (MNI), McGill University, in collaboration with the Université de Montréal. It is a continuously expanding resource; for this study, we include resting-state MEG from 644 participants (444 healthy controls and 200 patients with various psychiatric conditions).
- **WAND** [84] is a multiscale, multimodal neuroimaging dataset comprising in vivo brain data from 170 healthy adults (ages 18–63), including 3T and 7T structural and functional MRI as well as 275-channel MEG recorded at a 1,200 Hz sampling rate. We incorporate all available resting- and task-state fMRI (3 T and 7 T) and MEG data into the pretraining dataset.

A.3 Downstream Task Datasets

- **SEED-V** [85] is an emotion recognition EEG benchmark collected from 20 participants using a 62-channel ESI NeuroScan system at a 1,000 Hz sampling rate. It includes five emotion categories (happy, sad, fear, disgust and neutral) recorded across three sessions. Each session contains 15 trials, yielding a total of 117,744 1-second samples. Following the conventional strategy, we divide the fifteen trials of each session into three equal parts (5:5:5) as the training, validation and test sets, respectively.
- **CHB-MIT** [86, 87] is an intractable-seizure EEG database comprising recordings from 22 pediatric subjects (ages 1.5–22), acquired at 256 Hz with the international 10–20 system at Children's Hospital Boston. For our experiments, all recordings are resampled to 200 Hz and segmented into 326,993 10-second samples. Notably, seizure events constitute only 0.68% of all samples, resulting in a severe class imbalance. We follow the conventional subject-wise split: subjects 1 to 19 for training, subjects 20, 21 for validation, and subjects 22,23 for testing.
- **TUAB** [60] is a subset of the TUEG corpus, containing EEG recordings from 1,385 healthy controls (normal) and 998 subjects with a brain disorder (abnormal), totaling 2,993 sessions. After preprocessing, we obtain 409,455 10-second samples for binary classification task. The official evaluation split is used as our test set. To avoid potential data leakage effects, all subjects included in the TUAB test set are excluded from the TUEG dataset.
- **ADNI** [88] is a longitudinal, multisite neuroimaging initiative that includes structural, functional and molecular imaging, along with biofluid biomarkers, cognitive assessments, genetic data and demographic information from healthy older adults, individuals with mild cognitive impairment and patients with Alzheimer's disease (AD). For this downstream task, we incorporate resting-state fMRI data from 222 participants (110 patients with AD and 112 age- and gender-matched healthy controls), yielding a total of 508 scanning sessions. After preprocessing, a total of 618 samples are used.
- **ADHD-200** [89] is a large-scale, multi-site benchmark for neurodevelopmental classification, comprising resting-state fMRI, anatomical MRI and phenotypic data from 926 children and adolescents (573 typically developing controls, 353 individuals with ADHD) across eight international sites. In total, participants contributed 1,369 resting-state fMRI sessions. For the downstream binary ADHD vs. control classification task, all 1,369 sessions are used after mini-preprocessing.
- **CamCAN** [79] is a widely used benchmark for studying age-related changes in brain function, particularly in MEG. It contains both MEG and fMRI recordings from 656 healthy adults spanning the adult lifespan (ages 18–87). We use this dataset for chronological brain-age prediction under three evaluation settings: (i) unimodal resting-state fMRI (652 8min40s recordings), (ii) unimodal resting-state MEG (non-overlapping 177,044 2 second segments), and (iii) multimodal fusion of fMRI and MEG. To eliminate any risk of information leakage, all subjects included in the CamCAN validation and test splits are strictly excluded from the pretraining dataset.
- **LEMON** [81] consists of 227 healthy participants divided into a young cohort (N = 153, age 25.1 $\pm$ 3.1 years, range 20–35 years, 45 female) and an elderly cohort (N = 74, age

67.6 ± 4.7 years, range 59–77 years, 37 female). For downstream age-group classification (young vs. elderly), we evaluate on three settings: (i) unimodal fMRI using 438 7.5-minute resting-state scans, (ii) unimodal EEG using non-overlapping 17,235 10-second segments, and (iii) multimodal fusion of fMRI and EEG. To prevent information leakage and ensure reliable downstream evaluation, all subjects in the LEMON validation and test splits are excluded from the pretraining dataset.

- **HAD** [90] is a task-based fMRI dataset for human action recognition, consisting of fMRI responses to 21,600 video clips from 30 participants. The clips encompass 180 human action categories, covering a wide range of complex, real-world activities. For the downstream task, we perform a highly challenging 180-way event-related decoding classification of the evoked fMRI recordings. We employ a strict subject-wise split: subjects 1–24 for training (17,280 samples), 25–27 for validation (2,160 samples), and 28–30 for testing (2,160 samples).
- **BOLD5000** [91] is a slow event-related fMRI dataset collected from four participants across 16 sessions. During scanning, participants viewed nearly 5,000 distinct static images representing real-world scenes from the SUN, COCO, and ImageNet datasets. For the downstream evaluation, we conduct a 3-way cognitive decoding task to classify the source domain of each visual stimulus (COCO, ImageNet, or SUN). We adopt a strict subject-wise split: subjects 1 and 2 for training (10,508 samples), subject 3 for validation (5,254 samples), and subject 4 for testing (3,108 samples).

To explicitly characterize generalization to unseen data, sites, and subjects, as well as the partial overlap between pretraining and downstream datasets, we summarize the relationship between the pretraining corpus and each downstream dataset in Table 6. Our evaluation covers two major settings: *dataset-level generalization*, where the downstream dataset is entirely absent from pretraining, and *subject-level generalization*, where the same dataset or a related parent corpus contributes training subjects to pretraining, while all downstream validation and test subjects are strictly excluded.

Table 6: Relationship between the pretraining corpus and downstream evaluation datasets. ✓ denotes inclusion in pretraining, × denotes no inclusion or overlap, and △ denotes partial overlap through the same dataset or a related parent corpus. PT denotes pretraining. All downstream validation and test subjects are strictly excluded from pretraining.

Dataset	Pretraining	Relation	Train Subjects in PT	Val./Test Overlap	OOD Level
SEED-V	×	Independent	×	×	Dataset + Subject
CHB-MIT	×	Independent	×	×	Dataset + Site + Subject
TUAB	△	Partial	✓	×	Subject
ADNI	×	Independent	×	×	Dataset + Site + Subject
ADHD-200	×	Independent	×	×	Dataset + Site + Subject
CamCAN	△	Partial	✓	×	Subject
LEMON	△	Partial	✓	×	Subject
HAD	×	Independent	×	×	Dataset + Site + Subject
BOLD5000	×	Independent	×	×	Dataset + Site + Subject

B Baseline Models

To compare Brain-OF with representative baselines, we include both task-specific models and recent foundation models across different modalities, as summarized in Table 7.

Task-Specific Models:

- EEGNet [30] employs a compact CNN architecture with depthwise and separable convolutions to extract specialized features for EEG-based BCIs.
- EEGConformer [31] is a compact convolutional Transformer architecture designed to extract both local and global representations for EEG decoding.
- ContraWR [32] is a CNN-based encoder that employs contrastive learning on short-time Fourier transform (STFT) spectrograms for automated EEG sleep staging.
- CNN-Transformer [33] combines convolutional neural networks with transformer layers and employs a belief-matching loss to enable automated detection of EEG artifacts.

Table 7: **Overview of baseline models.** We categorize baselines into task-specific models (trained from scratch) and foundation models (pretrained). **Params** denotes parameter count in millions (M); **GFLOPs** measures computational cost per inference sample.

Models	Params (M)	GFLOPs	Architecture	Pretraining Modality
Task-Specific Models (Trained from scratch)				
EEGNet	0.004	0.13	CNN	✗
EEGConformer	0.81	0.58	Convolutional Transformer	✗
ContraWR	1.57	0.28	CNN	✗
CNN-Transformer	3.16	0.33	CNN + Transformer	✗
FFCL	2.47	0.96	CNN + LSTM	✗
ST-Transformer	3.43	0.22	Transformer	✗
BrainNetTF	5.56	0.74	Graph-structured Transformer	✗
BrainNetCNN	1.13	0.51	Connectivity-based CNN	✗
Foundation Models (Pretrained)				
LaBraM Base	5.82	2.70	Transformer	EEG
CBraMod	4.88	1.86	Criss-Cross Transformer	EEG
BrainHarmonix-F	85.25	1450.00	Vision Transformer	fMRI
Brain-JEPA	85.08	1271.79	Vision Transformer	fMRI
BrainOmni Base	32.71	11.43	Transformer	EEG + MEG

The high computational cost of BrainHarmonix-F (1,450 GFLOPs) and Brain-JEPA (1,272 GFLOPs) primarily derives from two factors: their support for long-range inputs exceeding 4k tokens and its large parameter count.

- FFCL [34] employs two independent branches that fuse multilevel spatial–temporal features, integrating CNN and LSTM architectures for EEG decoding.
- ST-Transformer [35] is a transformer-based architecture that leverages spatial–temporal attention mechanisms to capture both spatial dependencies and temporal dynamics in EEG signals for decoding tasks.
- BrainNetTF [29] is a transformer-based graph model for fMRI brain network analysis, which employs an orthonormal clustering readout to capture structured network representations.
- BrainNetCNN [36] employs specialized convolutional filters (edge-to-edge, edge-to-node, and node-to-graph) to capture the topological structure of brain networks for clinical outcome prediction.

Foundation Models:

- LaBraM [2] is a unified EEG foundation model that enables cross-dataset generalization by segmenting raw EEG signals into channel-wise patches for transformer-based representation learning.
- CBraMod [3] is an EEG foundation model that employs a criss-cross attention mechanism to jointly capture frequency domain and temporal features for self-supervised representation learning.
- BrainHarmonix-F [4] is the fMRI branch of the BrainHarmonix, a multimodal brain foundation model designed to process and integrate both structural and functional MRI data.
- Brain-JEPA [14] is a fMRI foundation model that utilizes the Joint-Embedding Predictive Architecture (JEPA) alongside Brain Gradient Positioning for representation learning.
- BrainOmni [21] is a brain foundation model pretrained on EEG and MEG recordings, incorporating learnable sensor position encoding and residual vector quantization.

To enable evaluation of these baselines on out-of-domain modalities, we adopt different adaptation strategies for task-specific models and pretrained foundation models. Task-specific models (EEGNet, EEGConformer, etc.) carry no pretrained weights and are trained from scratch on every downstream task, in every modality. Their input layers can accordingly be reconfigured freely to the target number of channels or ROIs and the target sequence length, and every such model receives the same input. This includes the connectivity-based models: for BrainNetCNN and BrainNetTF, the connectivity matrix is computed inside the model's own input pipeline, so they consume the same raw inputs as all other baselines. For pretrained brain foundation models (LaBraM, CBraMod, etc.), we instead

aim to preserve the pretrained architecture and parameters as completely as possible. Rather than reconfiguring and reinitializing the input layer, we leave the architecture untouched and instead crop or pad the input signal to a shape the pretrained model accepts. For example, the inputs to LaBraM and CBraMod are adjusted to the length that is a multiple of their 200-sample patch size. These results should therefore be read as the transfer of a fully pretrained model to an out-of-domain input, with no pretrained parameters discarded. To verify that this procedure preserves useful information from pretraining rather than merely reusing the architecture, we conduct an additional control experiment on LEMON-EEG using BrainHarmonix-F, comparing the pretrained model adapted as described above against the identical architecture trained from scratch under the same downstream protocol. Removing the pretrained weights decreases BAC from 77.95 to 73.13, indicating that the pretrained representation provides transferable information even under this out-of-domain adaptation.

C Modality Importance Scores

To quantify the contribution of each modality to overall model performance, we introduce two types of *Modality Importance Scores*. The first, termed the Modality Shapley Value, is grounded in cooperative game theory and measures the marginal contribution of each modality. In this framework, each modality represents a "player," and the model's downstream performance represents the outcome of the "game". The second metric, termed the Relative Importance Score, is derived directly from modality ablation results in Section 4.4, which measures the relative performance degradation on downstream tasks when a specific modality is removed during pretraining.

Since downstream tasks employ evaluation metrics with different units and scales, we first normalize performance gains to ensure comparability. Formally, the normalized value function of a modality subset S on downstream dataset D_j is defined as:

$$V_j(S) = \frac{P_j(S) - P_j(\emptyset)}{P_j(\mathcal{M}) - P_j(\emptyset)}$$

where $\mathcal{M} = \{M_{\text{fMRI}}, M_{\text{EEG}}, M_{\text{MEG}}\}$ is the set of all modalities used during joint pretraining, and $S \subseteq \mathcal{M}$ represents any subset (coalition) of these modalities. $P_j(S)$ denotes the downstream performance of the model pretrained on S, and $P_j(\emptyset)$ represents the baseline performance of a model trained from scratch (the empty coalition) on dataset D_j. By definition, $V_j(\emptyset) = 0$ indicates no pretraining gain, whereas $V_j(\mathcal{M}) = 1$ represents the total performance benefit derived from the omnifunctional joint pretraining. Higher values of $V_j(S)$ indicate greater gain derived from that specific subset of modalities. Notably, for lower-is-better metrics (e.g., MAE), the numerator and denominator are sign-flipped to preserve a consistent contribution direction.

Based on this normalized value function, the Modality Shapley Value for modality M_i on dataset D_j is computed as its expected marginal contribution to the overall performance, weighted and summed over all possible modality coalitions:

$$\phi_j(M_i) = \sum_{S \subseteq \mathcal{M} \setminus \{M_i\}} \frac{|S|!\,(|\mathcal{M}| - |S| - 1)!}{|\mathcal{M}|!} \left[V_j(S \cup \{M_i\}) - V_j(S)\right]$$

The Shapley value inherently satisfies the cooperative game theory properties of symmetry and additivity, ensuring that the total pretraining gain is fully decomposed across modalities:

$$\phi_j(M_{\text{fMRI}}) + \phi_j(M_{\text{EEG}}) + \phi_j(M_{\text{MEG}}) = 1$$

To complement this contribution analysis, the Relative Importance Score is introduced as a leave-one-out measure of modality dependence, quantifying the performance gain lost when modality M_i is explicitly withheld.

$$I_{M_i, D_j} = \frac{P_j(\mathcal{M}) - P_j(\mathcal{M} \setminus \{M_i\})}{P_j(\mathcal{M}) - P_j(\emptyset)} \times 100\%$$

This metric evaluates how strongly the full Brain-OF depends on a single modality M_i for task D_j. A positive score indicates that the modality contributes constructively to downstream performance. As above, lower-is-better metrics are adjusted to preserve consistent interpretation.

D Ablation Study of Mask Ratios in the MTFM

To investigate the impact of the mask ratio in the MTFM, we evaluate ratios from 0.4 to 0.9 using the Brain-OF Base model on three representative downstream datasets: TUAB (EEG) for abnormal detection, ADNI (fMRI) for Alzheimer's diagnosis, and CamCAN (MEG) for brain age prediction. We adopt a symmetric masking strategy where the temporal and frequency mask ratios are set to the same value. The results are summarized in Table 8. Performance drops sharply at a low mask ratio of 0.4 across all modalities and metrics, so we do not explore ratios below this threshold. For TUAB and ADNI, a ratio of 0.7 yields the strongest overall results (e.g., 81.88% BAC on TUAB, 68.23% BAC on ADNI), though 0.8 performs comparably on ADNI (74.53% AUROC). In contrast, ratios of 0.6 and 0.5 are superior on CamCAN (e.g., 8.83 MAE at 0.6), but underperform on fMRI. Balancing these trade-offs, we select 0.7 as the final mask ratio for MTFM. Interestingly, different modalities exhibit preferences for distinct optimal ranges: fMRI favors higher ratios (0.7–0.8) whereas EEG and MEG benefit from moderate ones (0.5–0.7), revealing modality-specific sensitivities in self-supervised learning.

Table 8: Performance comparison across different mask ratios in the MTFM.

Mask Ratios	TUAB (EEG)			ADNI (fMRI)			CamCAN (MEG)		
	BAC(%) ↑	AUC-PR(%) ↑	AUROC(%) ↑	BAC(%) ↑	AUC-PR(%) ↑	AUROC(%) ↑	MAE ↓	R2(%) ↑	MAPE(%) ↓
$\mathcal{N}(0.4, 0.05^2)$	81.02±0.12	87.17±0.09	88.67±0.06	61.10±2.57	69.84±2.36	67.54±2.47	11.93±0.38	32.70±3.92	23.25±0.77
$\mathcal{N}(0.5, 0.05^2)$	81.74±0.21	87.59±0.13	89.29±0.19	65.01±2.91	75.24±2.28	71.42±2.84	8.92±0.18	**58.19±1.71**	17.00±0.31
$\mathcal{N}(0.6, 0.05^2)$	81.82±0.31	87.64±0.20	**89.45±0.27**	63.74±2.76	72.87±3.12	70.85±3.31	**8.83±0.19**	57.02±1.94	**16.50±0.48**
$\mathcal{N}(0.7, 0.05^2)$	**81.88±0.17**	**87.69±0.11**	89.26±0.09	**68.23±2.07**	76.23±1.87	73.27±0.71	8.99±0.14	56.19±1.84	16.92±0.30
$\mathcal{N}(0.8, 0.05^2)$	81.51±0.09	87.46±0.06	89.14±0.09	67.86±2.62	**76.77±3.17**	**74.53±4.71**	9.45±0.40	53.27±3.62	18.06±0.51
$\mathcal{N}(0.9, 0.05^2)$	81.26±0.25	87.31±0.16	88.76±0.35	64.76±3.78	71.61±3.31	69.25±2.23	10.39±0.32	44.38±2.75	19.94±0.66

The results presented above are obtained with the Brain-OF Base.

E Ablation Study of Router Bias Update Rate γ

To examine the sensitivity of Brain-OF to the router bias update rate γ during pretraining, we conduct an ablation study over a range of values from 10^{-5} to 10^{-2}. Each pretrained model is subsequently fully finetuned and evaluated on three representative downstream tasks: TUAB (EEG), ADNI (fMRI) and CamCAN (MEG).

As illustrated in Figure 4, a consistent optimal configuration emerges across all modalities. In particular, $\gamma = 10^{-3}$ achieves the highest Balanced Accuracy on EEG and fMRI tasks and the lowest Mean Absolute Error on MEG. Larger values (e.g., 10^{-2}) lead to noticeable performance degradation, likely due to overly aggressive bias updates that destabilize expert routing and impair specialization. In contrast, smaller values (10^{-5}) fail to sufficiently update the router bias, limiting the effectiveness of load balancing.

Interestingly, we observe a different preference during the finetuning stage. While $\gamma = 10^{-3}$ is crucial for effective pretraining, substantially smaller update rates (e.g., $\gamma = 10^{-5}$) generally yield improved stability and performance during downstream finetuning, preserving the specialized knowledge acquired during pretraining.

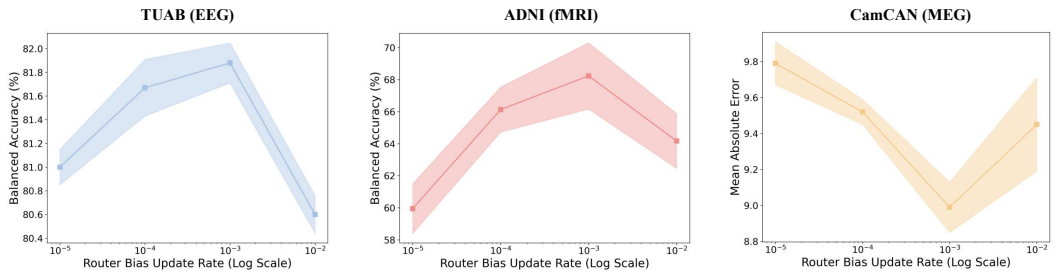


Figure 4: The impact of the router bias update rate γ during pretraining. The performances are evaluated on downstream tasks across three modalities.

F Task-State fMRI Downstream Evaluation

To further evaluate Brain-OF on fMRI modality, we extend downstream evaluation to two task-state fMRI benchmarks: HAD and BOLD5000. Both datasets involve visual stimuli tasks. Specifically, HAD is a complex 180-way classification of human actions based on evoked fMRI responses (evaluated via Top-10 Accuracy). BOLD5000 is a 3-way cognitive decoding task to identify the source domain of visual stimuli (images from COCO, ImageNet, or Scene databases, evaluated via Balanced Accuracy).

As shown in Table 9, Brain-OF achieves stronger performance than existing fMRI foundation models on both datasets. Notably, on the highly complex HAD benchmark, Brain-OF Base achieves 16.17% Top-10 Accuracy, surpassing Brain-JEPA by 3.86%. Similarly, it also achieves a better performance on BOLD5000 (40.49% BAC). This empirical evidence demonstrates that Brain-OF does not merely treat fMRI as a static clinical biomarker, but possesses robust, event-related functional decoding capabilities.

Table 9: Additional downstream evaluation on task-state fMRI.

Methods	**BrainHarmonix-F**	**Brain-JEPA**	**Brain-OF Base**
HAD (Top-10 Acc %) ↑	6.38±0.16	12.31±1.69	**16.17±0.44**
BOLD5000 (BAC %) ↑	38.94±1.66	39.79±0.41	**40.49±0.67**

G Visualization of Reconstruction

We visualize the reconstruction quality of Brain-OF Huge to assess the effectiveness of the proposed Masked Temporal-Frequency Modeling (MTFM) objective across fMRI, EEG, and MEG. Beyond simple waveform reconstruction, we examine whether the model preserves critical high-order features.

As shown in Figure 5, Brain-OF achieves high-fidelity reconstruction with effective noise suppression. For EEG, reconstructed signals closely match the original temporal dynamics while preserving dominant spectral bursts. In fMRI, functional connectivity derived from the original and reconstructed BOLD signals show that the complex inter-regional correlation structure is well preserved. For MEG, reconstructed topomaps accurately recover localized scalp-level magnetic field patterns despite signal sparsity. Overall, these results indicate that Brain-OF learns meaningful spatiotemporal and spectral generative structure rather than memorizing raw inputs.

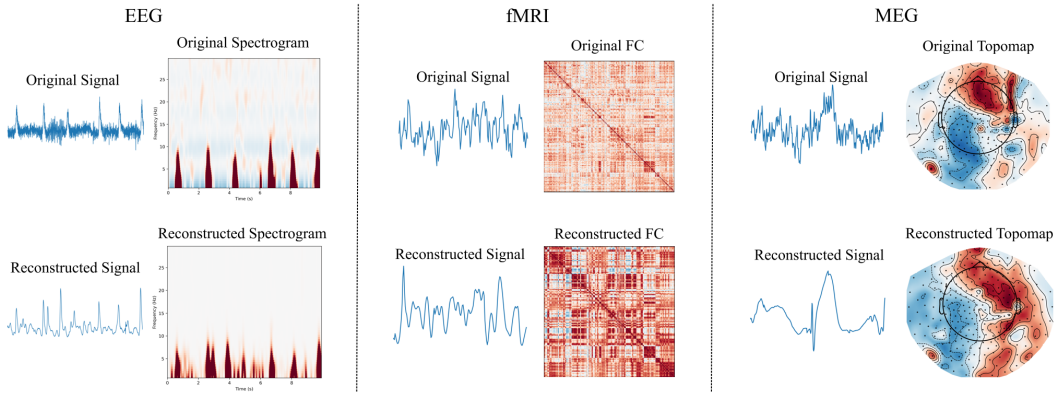


Figure 5: Qualitative visualization of signal reconstruction across fMRI, EEG and MEG.

H Additional Comparison with BrainOmni

We provide an additional comparison between Brain-OF Base and the recent foundation model BrainOmni Base, which is pretrained on EEG and MEG signals. BrainOmni is not included in the main comparison because its heavy reliance on rigid EMEG sensor layouts makes it structurally difficult to evaluate on fMRI datasets. In addition, its pretraining corpus includes the CamCAN

dataset, which gives it an unfair "seen-data" advantage on downstream tasks. Nevertheless, we include it here for completeness.

As shown in Table 10, Brain-OF achieves a +4.52% improvement on CHB-MIT and better performance on CamCAN-MEG (MAE 8.99 vs. 11.99). On TUAB, both models demonstrate highly comparable performance (81.88% vs. 81.90%). These results further support the robustness and generalization of Brain-OF across heterogeneous neuroimaging modalities.

Table 10: Additional comparison between Brain-OF and BrainOmni.

Methods	CHB-MIT	TUAB	CamCAN
	EEG	EEG	MEG
	BAC(%) ↑	BAC(%) ↑	MAE ↓
BrainOmni Base	69.47±2.19	**81.90±0.50**	11.99±0.84
Brain-OF Base	**73.99±1.48**	81.88±0.17	**8.99±0.14**

I Multimodal Fusion Comparison with Adapted BrainHarmonix

A direct comparison with existing multimodal brain foundation models is challenging because, to the best of our knowledge, no publicly available model provides a directly comparable pretrained pipeline for fusing fMRI with EEG or MEG on paired downstream datasets. Existing multimodal foundation models typically focus on different modality combinations. BrainHarmonix jointly models structural MRI (sMRI) and fMRI, BrainOmni jointly pretrains on EEG and MEG, while BIOT models heterogeneous biosignals such as EEG and ECG. These settings therefore do not directly address fusion between hemodynamic fMRI and electrophysiological EEG/MEG signals.

To provide a multimodal comparison, we adapt the BrainHarmonix fusion architecture, which originally consists of dual sMRI–fMRI branches, to our target setting. Specifically, we replace the original sMRI branch with a duplicated fMRI branch, such that one stream receives fMRI while the other receives either MEG or EEG. The complete two-branch model and fusion module are then finetuned end-to-end on the corresponding paired downstream training set. Because this configuration differs from the original sMRI–fMRI setting of BrainHarmonix, we regard it as an adapted architecture-level reference rather than the official BrainHarmonix configuration.

As shown in Table 11, the adapted BrainHarmonix fusion does not outperform its strongest unimodal branch on either dataset: its multimodal performance is worse than MEG alone on CamCAN and worse than EEG alone on LEMON. In contrast, Brain-OF Base and Large improve over their corresponding strongest unimodal settings on both benchmarks. Brain-OF also consistently outperforms the adapted BrainHarmonix baseline in the multimodal setting. These results suggest that a fusion mechanism originally designed for sMRI–fMRI integration does not directly transfer to heterogeneous fMRI–electrophysiology fusion, highlighting the distinct challenge addressed by Brain-OF.

Table 11: Comparison of multimodal fusion performance on CamCAN and LEMON. BrainHarmonix denotes an adapted architecture in which the original sMRI branch is replaced with a duplicated fMRI branch.

Methods	CamCAN			LEMON		
	fMRI	MEG	Multimodal	fMRI	EEG	Multimodal
	MAE ↓	MAE ↓	MAE ↓	BAC(%) ↑	BAC(%) ↑	BAC(%) ↑
BrainHarmonix	14.81 ± 0.49	8.73 ± 0.09	11.61 ± 0.75	63.11 ± 1.32	77.95 ± 2.36	74.43 ± 0.98
Brain-OF Base	12.09 ± 0.78	8.99 ± 0.10	**8.87 ± 0.09**	58.95 ± 2.13	76.05 ± 0.44	**76.77 ± 0.60**
Brain-OF Large	11.71 ± 0.67	8.86 ± 0.21	**8.78 ± 0.07**	62.85 ± 4.29	78.04 ± 1.59	**78.32 ± 1.64**
Brain-OF Huge	**9.51 ± 0.51**	**7.87 ± 0.22**	8.21 ± 0.11	61.79 ± 1.29	**80.39 ± 0.59**	**81.10 ± 0.54**

J Hyperparameter Configurations

In this section, we provide the detailed hyperparameter configurations in the Brain-OF pretraining, as well as the optimization settings for downstream tasks. All models were implemented in PyTorch and trained on NVIDIA A100 GPUs using automatic mixed precision.

J.1 Pretraining Settings

Table 12: Hyperparameters for pretraining Brain-OF.

Hyperparameters		Brain-OF Base	Brain-OF Large	Brain-OF Huge
Model Size		47.5M	331M	1.7B
Activated Parameters		21.5M	150M	500M
Patch size			64	
Max sequence length			2048	
Semantic space size			128	
Padding token			Full zero	
Brain Signal Encoder	Input dimension	{1, 4, 4}	{1, 8, 8}	{1, 12, 12}
	Output dimension	{4, 4, 4}	{8, 8, 8}	{12, 12, 12}
	Kernel size		{3, 3, 3}	
	Stride		{1, 1, 1}	
	Padding		{1, 1, 1}	
Brain-OF	ARNESS layer	4	4	6
	Backbone layer	12	24	36
	Attention head	4	8	12
	Hidden dimension	256	512	768
	Routed expert	8	16	28
	Activated expert	2	4	4
	Shared expert		1	
Pretraining settings	Epochs		150	
	Global batch size	7,168 (224×32 GPUs)	7,040 (80×88 GPUs)	7,168 (32×224 GPUs)
	Optimizer		AdamW	
	Adam β		(0.9, 0.95)	
	Peak learning rate		3e-3	
	Warmup learning rate		3e-5	
	Router bias update rate		1e-3	
	Learning rate scheduler		Cosine	
	Warmup epochs		8	
	Weight decay		0.05	
	Gradient clipping		5.0	
	Mask ratio		$\mathcal{N}(0.7, 0.05^2)$	
	Layer scale init	1e-4	1e-5	1e-6

J.2 Finetuning Settings on Downstream Tasks

For downstream tasks, Brain-OF is fully finetuned with a lightweight task-specific linear head following global average pooling. We employ cross-entropy loss for classification tasks and mean squared error loss for regression tasks. Optimization is performed using AdamW with $\beta = (0.9, 0.99)$ and a cosine learning rate scheduler. Detailed finetuning hyperparameter configurations for different model variants are provided in Table 13, Table 14 and Table 15.

Table 13: Hyperparameters for downstream finetuning (Brain-OF Base).

Datasets	Global bsz	Epochs	Warmup epochs	Lr	Warmup lr	Wd	Drop out	Drop path
SEED-V	1,024	50	5	1e-4	1e-6	5.0	0.0	0.0
CHB-MIT	1,024	50	5	3e-5	3e-7	0.1	0.0	0.1
TUAB	1,024	50	5	7e-6	7e-8	0.1	0.0	0.0
ADNI	32	50	5	1.5e-4	1.5e-6	0.1	0.1	0.0
ADHD-200	32	50	5	3e-5	3e-7	0.01	0.1	0.0
CamCAN-fMRI	32	100	5	7.5e-4	7.5e-6	0.1	0.1	0.0
CamCAN-MEG	1,024	50	5	1e-4	1e-6	5e-4	0.0	0.0
CamCAN-MM	512	50	5	2e-4	2e-6	0.1	0.0	0.2
LEMON-fMRI	32	100	5	2e-4	2e-6	0.1	0.1	0.1
LEMON-EEG	512	50	5	1.5e-5	1.5e-7	0.1	0.0	0.0
LEMON-MM	512	50	5	4e-5	4e-7	0.1	0.0	0.3

MM (Multimodal): Indicates finetuning Brain-OF on multimodal scenarios. Wd: Stands for weight decay. Specifically, CamCAN comprises paired fMRI and MEG, while LEMON comprises paired fMRI and EEG.

Table 14: Hyperparameters for downstream finetuning (Brain-OF Large).

Datasets	Global bsz	Epochs	Warmup epochs	Lr	Warmup lr	Wd	Drop out	Drop path
SEED-V	1,024	50	5	4e-4	4e-6	5.0	0.0	0.1
CHB-MIT	1,024	50	5	3e-5	3e-7	0.5	0.0	0.1
TUAB	1,024	50	5	3e-5	3e-7	0.1	0.0	0.3
ADNI	32	50	5	1e-4	1e-6	0.1	0.0	0.3
ADHD-200	32	50	5	4e-5	4e-7	0.01	0.1	0.2
CamCAN-fMRI	32	100	5	3e-3	3e-5	0.1	0.1	0.1
CamCAN-MEG	1,024	50	5	6e-4	6e-6	0.1	0.0	0.3
CamCAN-MM	512	50	5	8e-5	8e-7	0.05	0.0	0.2
LEMON-fMRI	32	100	5	3e-4	3e-6	0.1	0.0	0.1
LEMON-EEG	512	50	5	6e-4	6e-6	0.1	0.0	0.1
LEMON-MM	512	50	5	4e-4	4e-6	0.1	0.0	0.1

MM (Multimodal): Indicates finetuning Brain-OF on multimodal scenarios. Wd: Stands for weight decay. Specifically, CamCAN comprises paired fMRI and MEG, while LEMON comprises paired fMRI and EEG.

Table 15: Hyperparameters for downstream finetuning (Brain-OF Huge).

Datasets	Global bsz	Epochs	Warmup epochs	Lr	Warmup lr	Wd	Drop out	Drop path
SEED-V	1,024	50	5	1e-4	1e-6	5.0	0.0	0.1
CHB-MIT	1,024	50	5	1e-4	1e-6	0.1	0.0	0.1
TUAB	1,024	50	5	7e-6	7e-8	0.1	0.0	0.1
ADNI	32	50	5	3e-4	3e-6	0.1	0.0	0.1
ADHD-200	32	50	5	1e-4	1e-6	0.1	0.0	0.3
CamCAN-fMRI	32	100	5	5e-3	5e-5	0.1	0.0	0.1
CamCAN-MEG	1,024	50	5	4e-4	4e-6	0.1	0.0	0.3
CamCAN-MM	512	50	5	1e-3	1e-4	0.1	0.0	0.3
LEMON-fMRI	32	100	5	5e-4	5e-6	0.1	0.001	0.1
LEMON-EEG	512	50	5	1.5e-4	1.5e-6	0.1	0.0	0.1
LEMON-MM	512	50	5	4e-4	4e-6	0.1	0.0	0.1

MM (Multimodal): Indicates finetuning Brain-OF on multimodal scenarios. Wd: Stands for weight decay. Specifically, CamCAN comprises paired fMRI and MEG, while LEMON comprises paired fMRI and EEG.

K Efficiency Analysis

We evaluate the computational efficiency of Brain-OF across three model scales: Base, Large and Huge. Table 16 summarizes the parameter counts, training resources and inference latency.

Model Scaling and Training Cost. Brain-OF employs a Sparse Mixture of Experts architecture, which decouples the total parameter count from the computational cost per forward pass. Although Brain-OF Huge contains 1.7B parameters, only approximately 500M parameters (29%) are activated per token. This sparsity enables efficient scaling: while the total parameter count increases by $35\times$ from Base (47.5M) to Huge (1.7B), the number of active parameters grows by only $\sim 23\times$. Pretraining the Huge variant required 153 hours on 224 NVIDIA A100 GPUs.

Inference Latency and Throughput. Inference latency is measured on a single NVIDIA A100 GPU using BF16 precision, with a sequence length of 2,048 tokens and a batch size of 16. The average *per-sample* latency remains low: 5.5 ms for Base, 11.1 ms for Large, and 18.6 ms for Huge. This corresponds to a throughput of 181.8 samples/s for the Base model, indicating that Brain-OF remains suitable for real-time despite its large-scale architecture. Notably, the efficiency is further enhanced by ARNESS, which compresses high-dimensional brain signals into a compact semantic representation, significantly reducing the effective token length and computational burden.

Table 16: Efficiency analysis across model variants. **Active Params** denotes the number of parameters used per forward pass (due to sparse MoE). **Inference Latency** is reported for a batch size of 16 with sequence length 2,048 on a single A100 GPU.

Model Variants	Total Params	Active Params	Comp. GFLOPs	Pretraining Cost		Inference (Batch=16)	
				Time	# GPUs	Latency	Throughput
Brain-OF Base	47.5M	21.5M	11.29	71h 50m	32	88 ms	181.8 samples/s
Brain-OF Large	331M	150M	61.01	82h 06m	88	178 ms	89.9 samples/s
Brain-OF Huge	1.7B	500M	178.19	153h 03m	224	297 ms	53.8 samples/s

L Representation Visualizations on Downstream Datasets

To qualitatively assess the learned representations, we employ principal component analysis (PCA) [92] to visualize the feature embeddings generated by finetuned Brain-OF across three downstream datasets. Figure 6 illustrates the evolution of representations from raw input features (Column 1) to embeddings learned by Brain-OF Base, Large and Huge (Columns 2–4).

The first two rows correspond to classification tasks on TUAB (EEG) and ADNI (fMRI). In the raw feature space, samples from different classes (e.g., Normal vs. Abnormal, Healthy Control vs. Alzheimer's Disease) are heavily entangled. As model capacity increases from Base to Huge, a clear trend emerges: class-specific clusters become progressively more compact and better separated. On TUAB, the first principal component of Brain-OF Base explains 73.0% of the embedding variance and is strongly associated with the binary class label (Pearson $r = 0.701$; Spearman $\rho = 0.678$), achieving an AUC of 0.892 using this component alone. This correlation becomes slightly stronger for Brain-OF Large ($r = 0.707$, $\rho = 0.687$, AUC $= 0.898$) and Brain-OF Huge ($r = 0.723$, $\rho = 0.700$, AUC $= 0.905$). On ADNI, the class-related structure is weaker but remains evident. The corresponding statistics improve from Brain-OF Base ($r = 0.380$, $\rho = 0.388$, AUC $= 0.724$) to Large ($r = 0.478$, $\rho = 0.438$, AUC $= 0.753$) and Huge ($r = 0.484$, $\rho = 0.496$, AUC $= 0.786$). These results indicate that scaling the model capacity directly enhances its ability to extract discriminative, class-specific representations. The third row visualizes the regression task on CamCAN (MEG), where color gradients indicate chronological age. For Brain-OF Base, the first principal component exhibits a strong association with age ($r = 0.746$; $\rho = 0.735$). The age association further increases for Brain-OF Large ($r = 0.784$, $\rho = 0.762$) and Brain-OF Huge ($r = 0.805$, $\rho = 0.772$). Rather than forming discrete clusters, the CamCAN embeddings exhibit a continuous trajectory consistent with the continuous nature of chronological age. This structure suggests that Brain-OF captures meaningful age-related biological variation in the MEG signals. Overall, the results further show that scaling Brain-OF generally improves class separability in classification tasks and strengthens the organization of continuous age-related variation in regression, supporting the quality and scalability of the learned representations.

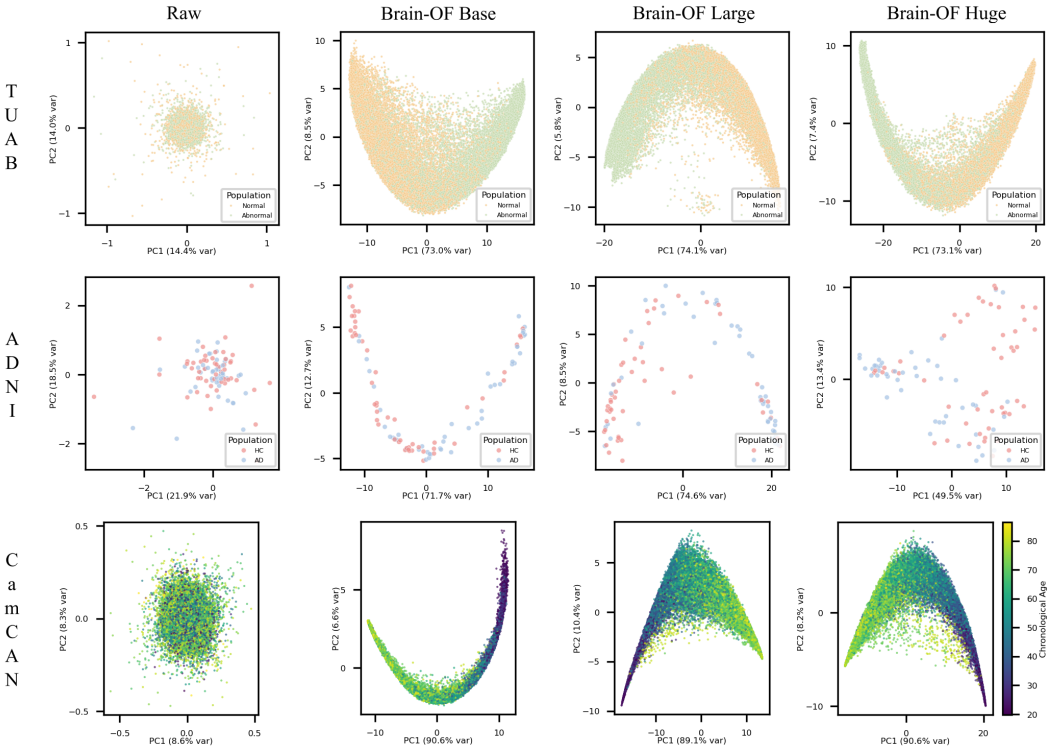


Figure 6: **PCA visualization of learned representations across model scales.** Comparison of raw input features (Left) with representations from Brain-OF Base, Large and Huge (Right). Rows correspond to different modalities and tasks: **Top:** TUAB (EEG, binary classification); **Middle:** ADNI (fMRI, binary classification); **Bottom:** CamCAN (MEG, age regression). As model size increases, the representations exhibit clearer class separation (Top/Middle) and a more coherent continuous age-related trajectory for regression (Bottom), demonstrating the benefits of scaling.

M Linear Probing and Pretraining Controls

Beyond the representation space visualization in Appendix L, we further evaluate the quality of the learned representations through frozen-backbone linear probing and same-architecture no-pretraining controls. We conduct these analyses on three representative downstream tasks spanning EEG (TUAB), fMRI (ADNI), and MEG (CamCAN).

We freeze the entire pretrained backbone and train only a prediction head. As shown in Table 17, on both classification tasks, linear probing outperforms full finetuning of the same architecture from random initialization. Specifically, on TUAB, linear probing achieves 79.14% vs. 77.59% (Base) and 81.92% vs. 80.82% (Huge). On ADNI, the corresponding results are 65.22% vs. 63.39% and 70.76% vs. 64.85%. Since the from-scratch models have identical architectures and strictly more trainable parameters, these results demonstrate that the pretrained representations themselves contain substantial task-relevant information beyond the effect of model capacity.

The no-pretraining control uses an identical architecture trained from scratch, thereby holding model capacity fixed while varying only the availability of pretrained weights. Pretrained Brain-OF Base outperforms the larger Brain-OF Huge trained from scratch on both TUAB (81.88 vs. 80.82) and ADNI (68.23 vs. 64.85). At matched model scale, removing pretraining decreases BAC by 4.29% and 4.84% on TUAB and ADNI for Brain-OF Base, and by 2.05% and 6.94% for Brain-OF Huge. The pretraining gains on TUAB and ADNI are also statistically significant at both model scales (two-sided Welch's t-test, corrected $p < 0.01$). These results indicate that simply increasing model capacity does not recover the gains provided by pretraining. On CamCAN-MEG, however, the effect is less straightforward. Full finetuning from pretrained checkpoints still yields numerically lower MAE than training from scratch (8.99 vs. 9.69 for Base; 7.87 vs. 8.59 for Huge), but these differences are not statistically significant, and frozen linear probing performs worse than full training from scratch. This suggests that the pretrained MEG representations remain useful for downstream adaptation, but brain-age regression benefits more strongly from nonlinear backbone adaptation than from a linear readout alone.

Table 17: Comparison of full finetuning, frozen-backbone linear probing, and training from scratch for Brain-OF Base and Huge.

Models	**Training Setting**	**TUAB**	**ADNI**	**CamCAN**
		EEG	**fMRI**	**MEG**
		BAC (%) ↑	BAC (%) ↑	MAE ↓
Brain-OF Base	Pretrained + Full FineTuning	81.88 ± 0.17	68.23 ± 2.07	8.99 ± 0.14
	Pretrained + Linear Probing	79.14 ± 0.24	65.22 ± 1.47	10.92 ± 0.22
	No Pretraining + Full FineTuning	77.59 ± 0.27	63.39 ± 2.36	9.69 ± 1.07
Brain-OF Huge	Pretrained + Full FineTuning	82.87 ± 0.18	71.79 ± 2.19	7.87 ± 0.22
	Pretrained + Linear Probing	81.92 ± 0.21	70.76 ± 1.22	9.02 ± 0.11
	No Pretraining + Full FineTuning	80.82 ± 0.18	64.85 ± 1.03	8.59 ± 0.94

N Ablation Study on Key Components

To verify the contribution of key architectural components, we conduct ablation studies on the Sparse MoE and DINT attention mechanism. Specifically, we pretrain two variants: (i) replacing Sparse MoE with a dense Feed-Forward Network (FFN), and (ii) substituting DINT attention with vanilla self-attention. As shown in Table 18, replacing Sparse MoE with a dense FFN results in the largest performance degradation across all tasks, with particularly severe drops on fMRI tasks. This suggests that a shared dense parameter space struggles to disentangle heterogeneous modality semantics, as the larger EEG/MEG corpora may dominate gradient updates in the dense model, hindering learning of subtle features from the smaller fMRI corpus. In contrast, Sparse MoE enables expert specialization, mitigating negative transfer and preserving performance even on under-represented modalities. Replacing DINT attention with vanilla self-attention also leads to consistent but smaller performance declines (e.g., SEED-V Cohen's κ drops from 23.98% to 22.82%). This indicates that standard attention is susceptible to the low signal-to-noise ratio of neural recordings, attending to irrelevant fluctuations rather than meaningful patterns. DINT alleviates this issue by suppressing irrelevant context and emphasizing salient signals. Overall, these ablations demonstrate that Sparse MoE and DINT attention are crucial for handling semantic heterogeneity and noise in brain signals.

Table 18: Ablation analysis of key components.

Methods	SEED-V	TUAB	ADNI	ADHD-200	CamCAN
	EEG	EEG	fMRI	fMRI	MEG
	Kappa(%) ↑	BAC(%) ↑	BAC(%) ↑	BAC(%) ↑	MAE ↓
Brain-OF (Base)	**23.98±0.29**	**81.88 ±0.17**	**68.23±2.07**	62.57±1.19	**8.99±0.14**
w/ Dense FFN	20.86±0.28	81.86±0.08	64.60±3.32	59.10±1.80	9.78±0.27
w/ Vanilla Attn.	22.82±0.38	81.85±0.09	64.32±2.82	**62.82±1.76**	9.50±0.53

O Robustness of DINT Attention under Low SNR

To explicitly validate the theoretical capability of DINT attention to handle low signal-to-noise ratio (SNR) conditions, we conducted a controlled SNR stress test on the TUAB dataset using Brain-OF (Base). We systematically injected varying levels of noise (from a high SNR of 20 dB down to a severe -5 dB) into the evaluation signals. As shown in Table 19, under high SNR (20 dB), the performance difference between DINT and vanilla attention is modest (+1.25%). However, as the noise level becomes extreme (-5 dB), vanilla attention suffers a catastrophic representation collapse (dropping to 63.96%). In contrast, DINT attention demonstrates strong robustness, maintaining 69.29% (+5.33%). These results provide empirical evidence that the differential/integral mechanism improves robustness under low-SNR conditions by stabilizing attention against noise.

Table 19: SNR stress test on the TUAB dataset comparing DINT and vanilla attention.

Methods	20 dB	10 dB	5 dB	0 dB	-5 dB
DINT Attention	**80.18±0.33**	**75.27±0.88**	**75.02±0.28**	**73.47±1.11**	**69.29±2.34**
Vanilla Attention	78.93±0.40	74.20±1.06	72.71±1.21	70.53±1.86	63.96±3.47

P Parameter-Efficient Fine-Tuning

While our primary downstream evaluations adopt full fine-tuning to fully exploit model capacity, this approach incurs substantial memory costs. To enable more efficient adaptation of Brain-OF, we explore Low-Rank Adaptation (LoRA) [93], a widely used parameter-efficient fine-tuning (PEFT) method. Specifically, we freeze the pretrained backbone and introduce trainable low-rank decomposition matrices (rank $r = 32$, scaling factor $\alpha = 64$), updating only $\sim$2% of the total parameters. As shown in Table 20, LoRA achieves competitive performance compared to full fine-tuning. Notably, on small-scale fMRI datasets, LoRA slightly outperforms full fine-tuning (e.g., ADNI: 68.23% $\rightarrow$ 68.44%, LEMON-fMRI: 58.95% $\rightarrow$ 59.36%). This suggests that the constrained parameter space of LoRA acts as an implicit regularizer, mitigating overfitting in low-data regimes. Although a modest performance drop is observed on more complex tasks (e.g., SEED-V), these results demonstrate that PEFT provides a practical and efficient alternative for adapting Brain-OF, especially in low-resource settings.

Table 20: Comparison between full fine-tuning and LoRA-based parameter-efficient fine-tuning on Brain-OF (Base).

Methods	SEED-V	TUAB	ADNI	ADHD-200	CamCAN	LEMON
	EEG	EEG	fMRI	fMRI	MEG	fMRI
	Kappa(%) ↑	BAC(%) ↑	BAC(%) ↑	BAC(%) ↑	MAE ↓	BAC(%) ↑
Full Finetuning	**23.98±0.29**	**81.88 ±0.17**	68.23±2.07	**62.57±1.19**	**8.99±0.14**	58.95±2.13
LoRA ($\sim$2% Params)	21.81±0.13	81.30±0.04	**68.44±2.13**	62.10±1.04	9.17±0.44	**59.36±3.02**

LoRA updates only $\sim$2% of parameters while maintaining competitive performance, and even improves results on small-scale fMRI datasets, indicating enhanced robustness in low-data regimes.

Q Label Efficiency and Low-Resource Analysis

To evaluate Brain-OF's robustness under data-scarce conditions, we assess the Brain-OF Base using progressively smaller fractions (10%–100%) of the downstream training data. We consider two representative tasks: abnormality detection on TUAB (EEG) and brain-age prediction on CamCAN (MEG). As illustrated in Figure 7, Brain-OF exhibits strong label efficiency across both tasks. For TUAB (left), the model achieves over 81.5% balanced accuracy using only 30% of the labeled data, nearly approaching the full-data performance of 81.88%. For CamCAN (right), the mean absolute

error consistently decreases as more data becomes available, yet the model maintains competitive accuracy even in low-resource settings (e.g., 30% training data). These results indicate that the rich, generalizable representations learned during large-scale multimodal pretraining reduce the dependency on large labeled downstream datasets. Brain-OF thus offers promising potential for low-resource neuroimaging scenarios, where annotated data is often limited or expensive to obtain.

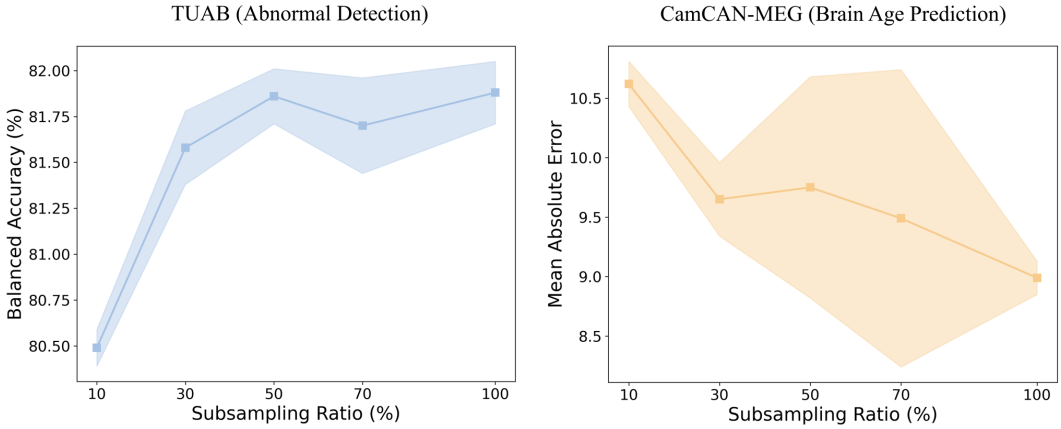


Figure 7: **Label efficiency of Brain-OF under limited supervision.** Brain-OF Base is evaluated across varying training data fractions (10%, 30%, 50%, 70%, 100%). Shaded regions denote standard deviation across five random seeds. Brain-OF achieves strong performance even with limited labeled data, highlighting the effectiveness of heterogeneous multimodal pretraining for low-resource settings.

R Limitations and Future Work

Although we curate one of the largest neuroimaging corpora to date, which spans over 40 datasets across fMRI, EEG and MEG, the overall data scale remains orders of magnitude smaller than that of foundation models in natural language processing or computer vision. Moreover, the pretraining corpus exhibits notable modality imbalance, with fMRI samples being considerably scarcer than EEG and MEG recordings. Future iterations of Brain-OF may address these challenges by incorporating additional modalities such as fNIRS and PET, and by extending the framework to support more general multivariate time series.

Currently, Brain-OF employs a relatively straightforward multimodal fusion strategy. While effective, this approach may not fully capture complex and nonlinear cross-modal interactions. Future work could explore more expressive fusion mechanisms, such as adaptive gating strategies or dedicated cross-modal attention modules, to better model fine-grained inter-modality dependencies.

S Broader Impacts

This work presents Brain-OF, an omnifunctional foundation model jointly pretrained on fMRI, EEG and MEG, representing a significant advance toward unifying heterogeneous neuroimaging modalities within a single framework. Compared to single-modality foundation models, Brain-OF exhibits potentially stronger generalization capabilities, even to unseen or specialized functional modalities. Consequently, Brain-OF has a broader impact on multimodal fusion in neuroscience research and clinical applications, empowering researchers to integrate the high spatial resolution of fMRI with the temporal precision of EEG/MEG for more accurate diagnosis in epilepsy, neurodegenerative diseases and cognitive disorders. Moreover, by transferring knowledge from capital-intensive modalities (fMRI/MEG) to accessible, wearable ones (EEG), this work accelerates the deployment of high-performance neurotechnologies in real-world and portable settings. We release Brain-OF Huge, one of the largest open-source brain foundation models to date. This provides a reliable, pre-validated backbone for the neuroscience and brain-computer interface community, substantially lowering the barrier to entry for researchers who lack resources to pretrain large-scale models from scratch. By enabling the reuse of rich, cross-modal representations, we aim to broaden access to powerful neural representations and foster a collaborative, data-centric ecosystem for future neuroimaging research.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The paper's main contributions and scope are explicitly stated in the abstract and introduction.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The limitations of the work are discussed in the Limitations and Future Work section (Appendix R).

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [N/A]

Justification: The work primarily focuses on applications in neuroscience and therefore does not include theoretical results.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: The main text describes the model architecture and performance, while detailed hyperparameter configurations are provided in Appendix J. In addition, all datasets are publicly available, and the code will be released upon acceptance.

Guidelines:

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

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

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

   Answer: [Yes]

   Justification: All datasets are publicly available, and the code and pretrained model weights have been released.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: All experimental details are provided in Appendix A.3 and Appendix J.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: The work reports the average performance and standard deviation for each downstream task, as shown in Table 2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The computational efficiency of model pretraining is reported in Appendix K.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in this paper adhered to the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The broader impacts of this work are discussed in Appendix S.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The model is released for research use only and does not currently pose known security risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All datasets used in this work are described in detail and properly cited in Appendix A.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: The code, pretrained checkpoints for all model variants, and related documentation will be made publicly available.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing. All human-subject data used in this study come from publicly available neuroimaging datasets.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing. All human-subject data used in this study come from publicly available neuroimaging datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

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