
Brain Harmony: A Multimodal Foundation Model Unifying Morphology and Function into 1D Tokens

Zijian Dong^{*1} Ruilin Li^{*1,2} Joanna Su Xian Chong¹ Niousha Dehestani¹ Yinghui Teng¹
Yi Lin¹ Zhizhou Li¹ Yichi Zhang¹ Yapei Xie¹ Leon Qi Rong Ooi¹ B.T. Thomas Yeo¹
Juan Helen Zhou^{†1}

¹National University of Singapore

²Now at MiroMind AI

Abstract

We present **Brain Harmony (BrainHarmonix)**, the first multimodal brain foundation model that unifies structural morphology and functional dynamics into compact 1D token representations. The model was pretrained on two of the largest neuroimaging datasets to date, encompassing 64,594 T1-weighted structural MRI 3D volumes (~14 million images) and 70,933 functional MRI (fMRI) time series. BrainHarmonix is grounded in two foundational neuroscience principles: *structure complements function* - structural and functional modalities offer distinct yet synergistic insights into brain organization; *function follows structure* - brain functional dynamics are shaped by cortical morphology. The modular pre-training process involves single-modality training with geometric pre-alignment followed by modality fusion through shared brain hub tokens. Notably, our dynamics encoder uniquely handles fMRI time series with heterogeneous repetition times (TRs), addressing a major limitation in existing models. BrainHarmonix is also the first to deeply compress high-dimensional neuroimaging signals into unified, continuous 1D tokens, forming a compact latent space of the human brain. BrainHarmonix achieves strong generalization across diverse downstream tasks, including neurodevelopmental and neurodegenerative disorder classification and cognition prediction - consistently outperforming previous approaches. Our models - pretrained on 8 H100 GPUs - aim to catalyze a new era of AI-driven neuroscience powered by large-scale multimodal neuroimaging. Code is available at: <https://github.com/hzlab/Brain-Harmony>

1 Introduction

The human brain is an extraordinarily complex organ, characterized by intricate anatomical architecture and dynamic functional processes. To investigate these aspects in vivo, researchers rely on neuroimaging techniques that probe brain structure and function. However, each neuroimaging modality captures only a single facet of this multifaceted system [1, 2, 3]. This limitation highlights the necessity of multimodal neuroimaging approaches that combine complementary information (*e.g.*, both structural and functional) to offer a holistic understanding of human cognition and improve clinical applications.

Recent advances in brain foundation models have transformed Artificial Intelligence (AI) for neuroimaging analysis from task-specific approaches to self-supervised pretrained models capable of

^{*}Equal contribution

[†]Corresponding author: helen.zhou@nus.edu.sg

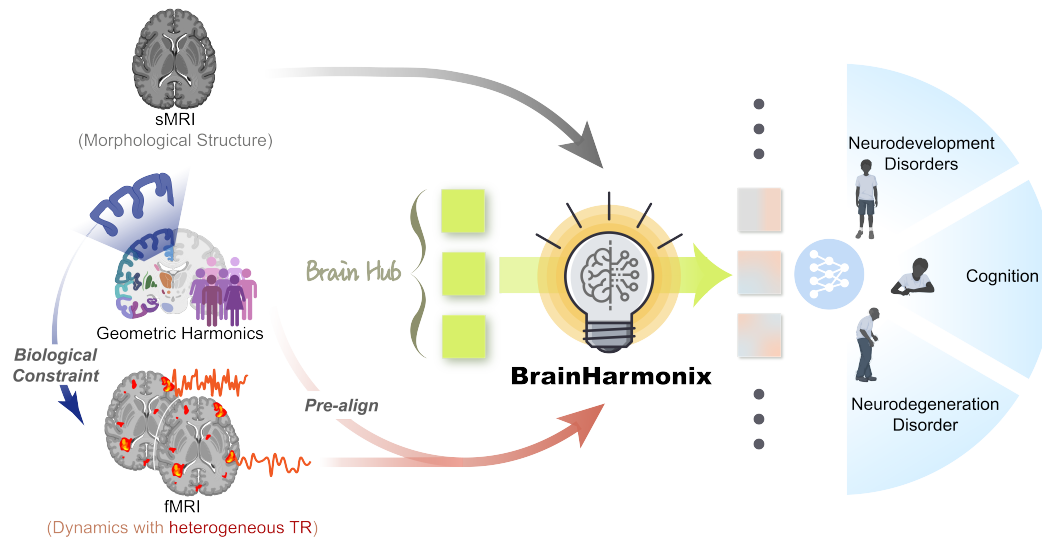


Figure 1: Overview of Brain Harmony (BrainHarmonix). Brain morphology from T1-weighted MRI (sMRI) and functional dynamics from fMRI are unified into compact 1D brain-hub tokens, which can be readily adapted to downstream tasks via an attached projection head. Specifically, functional dynamics are pre-aligned with group-level geometric harmonics, with built-in flexibility to handle heterogeneous repetition times (TRs). This fusion creates a compact yet expressive representation space that effectively captures the interplay between brain structure and function, supporting a broad range of downstream applications, including neurodevelopmental and neurodegenerative disorder classification and cognition prediction.

adaptation across diverse downstream applications [4, 5, 6, 7]. Despite their promising generalizability, these models focus on either brain structure [6] (*e.g.*, T1, T2-weighted MRI) or function [4, 5, 7] (*e.g.*, functional MRI (fMRI)), without capturing the two complementary aspects simultaneously. Furthermore, recent neuroscience findings demonstrate that brain activity can be formulated as excitations of fundamental resonant modes shaped by the brain's geometry, revealing how morphological structure fundamentally constrains functional dynamics [8]. Nevertheless, existing brain dynamics foundation models overlook this crucial constraint imposed by brain morphology. On the other hand, existing brain dynamics foundation models rely exclusively on fMRI datasets with homogeneous temporal resolutions [4, 5], hindering the integration of datasets collected from diverse scanners and protocols with varying repetition times (TRs). Even within individual datasets or real-world clinical scenarios, multiple TRs often coexist [9, 10, 11], rendering previous models infeasible for broader deployment. This limitation substantially reduces available sample sizes and constrains comprehensive modeling of brain dynamics across multiple temporal scales. While connectivity-based approaches are naturally agnostic to variations in TR [7], they aggregate activity across entire scanning sessions, discarding essential non-stationary dynamics (*e.g.*, transient state transitions and evolving co-activation patterns) in the blood-oxygen-level dependent (BOLD) signals [12, 13, 4, 5].

Together, the aforementioned gaps highlight a fundamental challenge: creating comprehensive brain representations that effectively capture both structural and functional neuroimaging data with heterogeneous temporal resolutions. A critical step toward addressing this challenge is developing efficient methods to compress high-dimensional neuroimaging data into compact, information-dense representations. Transforming complex neuroimaging data into sequential 1D tokens offers a promising solution, potentially providing a unified framework for integrating multimodal information across diverse neuroimaging acquisitions.

In this paper, we propose **Brain Harmony (BrainHarmonix)** to address these critical gaps (Figure 1). Our major contributions include: **(1)** Developing the first multimodal brain foundation model to bridge morphological structure *and* functional dynamics in a compact, information-rich representation space with 1D tokens. **(2)** Incorporating geometric harmonics to pre-align cortical morphology and functional organization, embedding structural constraints directly into functional representations. Imposing this population-level, physics-informed inductive bias can further enhance cross-subject and cross-dataset alignment. **(3)** Developing novel Temporal Adaptive Patch Embedding (TAPE) that enables scalable fMRI pretraining across heterogeneous TR values, overcoming a key limitation

of existing models. **(4)** Introducing the first effective data augmentation for fMRI time series - downsampling to hierarchical TR levels - to accommodate heterogeneous TR distributions and enhance performance. **(5)** Finally, BrainHarmonix was benchmarked on a diverse set of downstream tasks, including the diagnosis of neurodevelopmental and neurodegenerative disorders, as well as the prediction of cognition. We demonstrate, for the first time, that *complex brain morphology and dynamics can be deeply compressed into unified continuous-valued 1D tokens* that serve as holistic representations of the human brain.

2 Related Work

Recent brain foundation models have made significant advances in learning human brain representations. BrainLM [4] and Brain-JEPA [5] pioneered self-supervised learning for fMRI time series using masked prediction and joint-embedding approaches, respectively. While these models demonstrated promising generalizability through global representations, they suffer from two critical shortcomings: (1) they ignore brain structural information, and (2) due to their standard choice of patch embedding layer in transformers, cannot accommodate heterogeneous TRs common across - or even within - fMRI datasets. BDO [14] proposed a brain dynamics model based on stochastic optimal control, however, it focuses exclusively on brain dynamics, similar to Brain-JEPA and BrainLM. In addition, BrainMass [7] has been proposed as the first foundation model for brain functional connectivity and pretrained on diverse fMRI datasets. However, it focuses exclusively on static functional connectivity without capturing brain structural information or temporal dynamics. On the other hand, BrainMVP [6] introduced self-supervised pretraining for 3D volumetric brain imaging that excels at learning correspondence among multi-parametric MRI, but fails to capture brain functional dynamics. This makes it suboptimal for gaining a comprehensive understanding of human brain functional organization, capturing individual differences in behavior, and detecting abnormal alterations associated with neuropsychiatric disorders. To the best of our knowledge, Brain Harmony (BrainHarmonix) addresses these limitations as *the first multimodal foundation model that seamlessly integrates structural morphology with functional dynamics while accommodating variable TR values*. By unifying both modalities into 1D tokens, BrainHarmonix creates a compact and effective representational space that captures the holistic nature of the human brain.

3 Method

The pretraining of Brain Harmony (BrainHarmonix) comprises two sequential stages (Figure 2): **(1) Unimodal Encoding (UE)**: we first separately train modality-specific encoders for T1 (BrainHarmonix-S) and fMRI (BrainHarmonix-F). This separation allows flexible use of unpaired structural and functional data. For BrainHarmonix-S, we employ a 3D Masked Autoencoder (MAE) [15] that effectively captures structural information from the *largest* curation of T1 imaging datasets (given the widely adoption of MAE, readers are referred directly to Section 4.2 for implementation details). For BrainHarmonix-F, we propose two significant innovations to masked brain modeling: first, a geometric harmonics-based alignment method that pre-aligns brain dynamics with structural geometry; second, an innovative Temporal Adaptive Patch Embedding (TAPE) layer, enabling the encoder to flexibly accommodate *any* TR for the first time. Leveraging this unprecedented flexibility, we further introduce data augmentation techniques for fMRI time series, creating hierarchical TR values by downsampling high-resolution data. **(2) Multimodal Fusion (MF)**: modality-specific representations are fused through a set of learnable 1D brain hub tokens. These tokens act as a representational bottleneck, explicitly trained to reconstruct both structural and functional latents, resulting in a highly compact and unified latent space for human brain morphology and function.

3.1 Unimodal Encoding (UE)

In this subsection, we highlight two key innovations introduced in our approach for encoding fMRI dynamics (BrainHarmonix-F). First, we propose a geometric pre-alignment between brain dynamics and geometric harmonics, leveraging the foundational constraint that brain morphology inherently imposes on functional dynamics. Second, to effectively handle datasets with heterogeneous TRs, we introduce the Temporal Adaptive Patch Embedding (TAPE) layer, which enables token generation with consistent temporal length across varying TRs. For the encoding of T1 imaging (BrainHarmonix-S), readers are referred to the implementation details provided in Section 4.2.

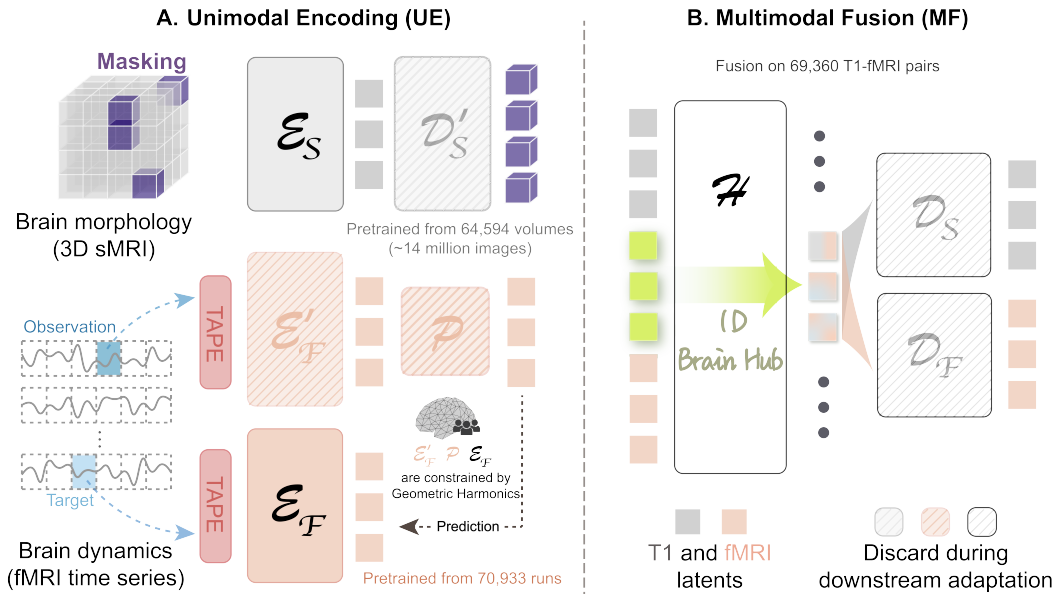


Figure 2: **Pretraining of Brain Harmony (BrainHarmonix).** **A. Unimodal Encoding (UE):** BrainHarmonix-S ($\mathcal{E}_S$) learns T1 structure via a Masked Autoencoder (MAE); gray cubes represent visible patches, while purple cubes are masked and reconstructed by the decoder ($\mathcal{D}'_S$). BrainHarmonix-F ($\mathcal{E}_F$) uses the Joint Embedding Predictive Architecture (JEPA) for fMRI, incorporating our Temporal Adaptive Patch Embedding (TAPE) for heterogeneous TRs and geometric harmonics for cortical alignment, with the observation encoder ($\mathcal{E}'_F$) and predictor ($\mathcal{P}$) following standard JEPA. **B. Multimodal Fusion (MF):** The Harmonizer ($\mathcal{H}$) fuses structural and functional latents into 1D tokens (in green), then decoder ($\mathcal{D}_S$ & $\mathcal{D}_F$) reconstruct modality-specific latents.

3.1.1 Pre-alignment between brain dynamics and geometry

Recent neuroscientific research has revealed the profound relationship between brain morphology and functional dynamics, demonstrating that functional brain activity propagates as waves constrained by cortical geometry [8]. In BrainHarmonix-F, we propose to pre-align brain dynamics with morphology. Specifically, we position brain ROIs in the transformer using geometric harmonics derived from population-level cortical surface mesh, thereby enhancing the structure-function coherence of the learned latent space of brain dynamics.

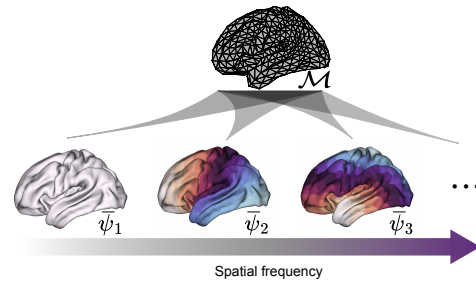


Figure 3: **Geometric harmonics.**

Geometric harmonics are the natural, orthogonal vibration patterns of the brain's folded surface. Given a mesh representation $\mathcal{M}$ of a population-averaged cortical surface derived from T1 imaging, the Laplace-Beltrami operator (LBO) $\Delta_{\mathcal{M}}$ is constructed to capture local vertex-to-vertex spatial relationships and cortical curvature. The corresponding eigenvalue problem can then be solved as follows:

$$\Delta_{\mathcal{M}} \psi = -\lambda \psi, \psi_i \xrightarrow{\downarrow} \bar{\psi}_i \quad (1)$$

where $\psi = \{\psi_1, \psi_2, \dots, \psi_i, \dots\}$ is the sequence of geometric harmonics with the corresponding eigenvalues $\lambda = \{\lambda_1, \lambda_2, \dots, \lambda_i, \dots\}$ ordered regarding spatial frequency (Figure 3). Each $\psi_i \in \mathbb{R}^{V \times 1}$ is further downsampled through averaging within one ROI to formulate $\bar{\psi}_i \in \mathbb{R}^{N \times 1}$, where V represents the number of vertices in the mesh and N denotes the number of ROIs in a brain parcellation. First J downsampled harmonics $\bar{\psi}_J \in \mathbb{R}^{N \times J}$ are selected for learning positional embedding. We incorporate a learnable linear layer to transform geometric harmonics $\bar{\psi}_J$ into positional embeddings $E \in \mathbb{R}^{N \times d}$, where d denotes the embedding dimension of the transformer.

By explicitly encoding the geometric constraints into fMRI representations, we pre-align functional brain organization with cortical structure, enabling more effective integration during subsequent modality fusion. Moreover, embedding this physics-informed inductive bias, derived from population-level observations, can further enhance cross-subject and cross-dataset alignment.

3.1.2 Temporal Adaptive Patch Embedding (TAPE)

fMRI data collection spans scanners, sites and pulse-sequences that sample the BOLD signals anywhere from sub-second to several-second TRs. However, existing brain dynamics foundation models rely on a fixed TR for both pre-training and downstream tasks [4], or downsample datasets with higher temporal resolutions to match lower ones [5]. This limitation restricts their ability to incorporate diverse datasets during pretraining and adapt flexibly to downstream tasks with varying TRs. Furthermore, downsampling inevitably sacrifices finer temporal details, reducing the richness of information encoded at higher temporal resolutions. The inability of existing models to accommodate heterogeneous TRs stems from the use of a uniform patch size and a patch embedding layer with a fixed size in the transformer. When training on fMRI data with diverse TRs, employing a single patch size leads each patch to inadvertently represent different temporal durations across data. The embedding layer, however, fails to accurately interpret the varying temporal extents of the input. As a result, the tokens passed to the transformer lack consistent and well-defined temporal semantics, introducing ambiguity that hampers effective modeling of brain dynamics. This limitation ultimately degrades the quality of learned representations and impairs downstream performance, particularly in capturing brain-behavior associations.

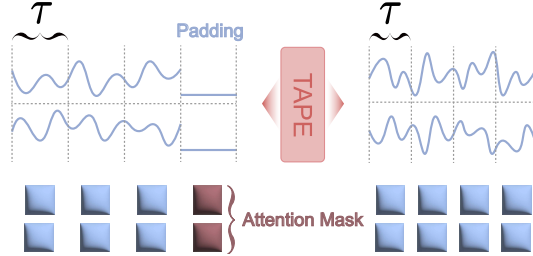


Figure 4: **Temporal Adaptive Patch Embedding (TAPE)**. Tokens represent the same temporal duration τ . Embedding weights are correspondingly resized. Shorter time series with fewer tokens are zero-padded, with attention masks excluding padded tokens afterwards.

To overcome this critical limitation, in BrainHarmonix-F, we propose Temporal Adaptive Patch Embedding (TAPE) that dynamically accommodates varying TRs across fMRI data (Figure 4). We first define a consistent temporal duration τ for any token and a base embedding weight $\omega^* \in \mathbb{R}^{k^*}$ corresponding to the patch size k^* . Given an arbitrary fMRI time series with repetition time $TR = s$, the corresponding patch size k and resized embedding weights $\omega \in \mathbb{R}^k$ are computed as:

$$k = \text{round}\left(\frac{\tau}{s}\right), \omega = ((B_k^{k^*})^T)^\dagger \cdot \omega^* \quad (2)$$

where ω is obtained by pseudoinverse resize (PI-resize) [16], with the linear transformation matrix $B_k^{k^*} \in \mathbb{R}^{k \times k^*}$. Since different fMRI scans may vary in total duration, patchifying with temporally consistent tokens could result in a varying number of tokens across time series. If the maximum number of tokens per time series across the dataset is m , any time series producing fewer tokens ($n < m$) will be zero-padded followed by an attention mask, ensuring tokens derived from padding are excluded from attention computation (Figure 4).

Previously, no established data augmentation techniques existed for fMRI time series; our TAPE uniquely supports arbitrary TRs, enabling the first-ever augmentation by downsampling high-resolution scans into diverse TRs, enhancing model performance (details in Section 4.1).

3.2 Multimodal Fusion (MF)

At BrainHarmonix's MF stage, we introduce learnable 1D brain hub tokens that serve as a representational bottleneck. These tokens are trained through the attention-based model [17] to reconstruct both structural and functional latents, effectively capturing the shared information between modalities (Figure 2.B.). 1D brain hub tokens foster a unified and compact latent space that encapsulates the holistic nature of brain morphology and dynamics.

Let $\mathbf{Z}_S \in \mathbb{R}^{N_S \times d}$, $\mathbf{Z}_F \in \mathbb{R}^{N_F \times d}$ be the modality-specific latents of one paired T1-fMRI produced by BrainHarmonix-S and BrainHarmonix-F, respectively, where d is the common embedding dimen-

sion and N_S, N_F are the numbers of tokens in each modality. We introduce a set of N_H learnable continuous-valued 1D brain hub tokens $\mathbf{H}_0 \in \mathbb{R}^{N_H \times d}$, shared by all pairs and optimized jointly with the network in MF. At every forward pass we concatenate the hubs with the two modality sequences and feed the resulting stream to the Harmonizer transformer ($\mathcal{H}$):

$$\mathbf{Z}_0 = [\mathbf{H}_0; \mathbf{Z}_S; \mathbf{Z}_F] \in \mathbb{R}^{(N_H+N_S+N_F) \times d}, \tilde{\mathbf{H}} = \mathcal{H}(\mathbf{Z}_0)_{1:N_H} \in \mathbb{R}^{N_H \times d} \quad (3)$$

where $\mathbf{Z}_0$ is the concatenated input to $\mathcal{H}$ and $\tilde{\mathbf{H}}$ is the hub tokens updated by $\mathcal{H}$. Self-attention within $\mathcal{H}$ allows the 1D tokens to gather information from *both* structural and functional tokens, while also enabling cross-modal interactions between $\mathbf{Z}_S$ and $\mathbf{Z}_F$.

Two lightweight decoders ($\mathcal{D}_S, \mathcal{D}_F$) project $\tilde{\mathbf{H}}$ back into each modality's latent space (Figure 2.B.). Formally, our training in MF is defined as:

$$\min_{\theta_{\mathcal{H}}, \theta_{\mathcal{D}_S}, \theta_{\mathcal{D}_F}} \mathcal{L}_{\text{fusion}} = \|\mathcal{D}_S(\tilde{\mathbf{H}}) - \mathbf{Z}_S\|_2^2 + \|\mathcal{D}_F(\tilde{\mathbf{H}}) - \mathbf{Z}_F\|_2^2 \quad (4)$$

where $\theta_{\mathcal{H}}, \theta_{\mathcal{D}_S}, \theta_{\mathcal{D}_F}$ represents the parameters in $\mathcal{H}, \mathcal{D}_S$, and $\mathcal{D}_F$, respectively. $\|\cdot\|_2^2$ denotes the Mean Square Error.

4 Experiments

4.1 Datasets

Pre-training. BrainHarmonix was pretrained on two of the largest-scale neuroimaging datasets: UK Biobank (UKB) [18, 19] and Adolescent Brain Cognitive Development (ABCD) [20]. From UKB, we curated neuroimaging data of 43,112 participants aged between 44 and 83 years, comprising 46,455 T1-weighted MRI scans and 40,162 resting-state fMRI time series (TR = 0.735 s). From ABCD, we included 11,221 participants (aged 8 to 11 years at baseline visit), consisting of 18,139 T1-weighted images and 30,771 resting-state fMRI time series (TR = 0.8 s).

During the UE stage, a total of 64,594 T1-weighted images from both datasets were utilized for BrainHarmonix-S pretraining. For fMRI data augmentation, UKB data underwent temporal down-sampling by factors of 1 to 3, resulting in TRs of 0.735 s, 1.47 s, 2.205 s, and 2.94 s. ABCD data were downsampled by factors of 1 to 2, yielding TRs of 0.8 s, 1.6 s, and 2.4 s. Consequently, the total number of pretraining samples for BrainHarmonix-F was 252,961 (UKB: $40,162 \times 4$; ABCD: $30,771 \times 3$). In the MF stage, we extracted 69,360 matched T1-fMRI pairs from both datasets (one T1-weighted image could correspond to multiple fMRI runs within a single session). All fMRI data was parcellated into $N = 400$ ROIs with Schaefer-400 [21]. Further details regarding data preprocessing are provided in Appendix A.

Downstream fine-tuning. We evaluated BrainHarmonix on six neuroimaging benchmark datasets. Three multi-site datasets focused on neurodevelopmental disorder diagnosis (TR distributions are detailed in Figure 7): Autism Brain Imaging Data Exchange datasets (ABIDE-I and ABIDE-II) for distinguishing Autism Spectrum Disorder (ASD) from controls, and the Attention Deficit Hyperactivity Disorder dataset (ADHD-200) for classifying ADHD versus controls. On the other hand, three datasets assessed neurodegenerative disorders and cognitive function: Parkinson's Progression Markers Initiative (PPMI) (TR = 2.5s) for four-class classification involving controls, scans without evidence of dopaminergic deficit (SWEDD), prodromal cases, and Parkinson's disease (PD); Alzheimer's Disease Neuroimaging Initiative (ADNI) (TR = 3.0s) for classification between controls and mild cognitive impairment (MCI); and the Lifespan Human Connectome Project Aging (HCP-A) dataset (TR = 0.8s) for predicting executive function (Flanker task scores). The results were averaged across three independent runs with distinct data splits (train:validation:test = 6:2:2). We adopted the data stratification approach in [22] for splitting the neurodevelopmental datasets. Detailed information regarding class distributions and preprocessing procedures for each benchmark dataset can be found in Appendix A.

4.2 Implementation details

In UE, we adopted Vision Transformer-Base (ViT-B) [23] as backbone for BrainHarmonix-S and BrainHarmonix-F (ε_S and ε_F). We employed MAE [15] as the pretraining framework for

Table 1: Comparison on neurodevelopmental disorder diagnosis. Results are averages over three random splits (standard deviations in Table 4). The best results are highlighted in bold (* indicates statistical significance, $p < 0.05$), and second-best results are underlined. Task details in Section 4.1.

Model	Morphology	Dynamics	Multi-TR	ABIDE-I		ABIDE-II		ADHD-200	
				ACC%	F1%	ACC%	F1%	ACC%	F1%
Structure-based models									
BrainMVP ¹ [6]	✓	✗	N.A.	56.50	62.46	55.71	62.16	<u>67.72</u>	43.97
BrainMVP ² [6]	✓	✗	N.A.	55.06	64.43	55.63	58.76	<u>62.59</u>	49.95
BrainHarmonix-S	✓	✗	N.A.	56.29	62.06	60.00	68.55	64.96	53.53
Function-based models									
BrainNetCNN [26]	✗	✗	✓	60.49	67.13	59.71	67.27	60.54	58.62
BrainGNN [27]	✗	✗	✓	56.72	65.71	58.71	66.48	62.24	60.67
BrainNetTF [22]	✗	✗	✓	56.73	64.75	62.03	67.64	61.91	62.68
BrainMass [7]	✗	✗	✓	65.64*	69.07	59.35	71.86	65.99	61.27
BrainLM [4]	✗	✓	✗	–	–	–	–	–	–
Brain-JEPA [5]	✗	✓	✗	–	–	–	–	–	–
BrainHarmonix-F	✗	✓	✓	57.39	<u>71.24</u>	<u>62.90</u>	<u>72.76</u>	67.69	68.75
Multimodal model									
BrainHarmonix	✓	✓	✓	<u>63.13</u>	72.63*	66.67*	74.88*	70.09*	<u>66.72</u>

¹UniFormer [28] as backbone; ²UNET3D [29] as backbone.

BrainHarmonix-S with Brain-JEPA [5] for BrainHarmonix-F. In BrainHarmonix-F, it patchified fMRI time series into 1D patches, with the length k dynamically determined by TR. Geometric harmonics and brain gradients [5] were each linearly projected, then averaged to produce the final positioning. T1 images were randomly masked while we followed [5] to use spatiotemporal masking for fMRI.

In MF, harmonizer ($\mathcal{H}$) was employed with ViT-B encoder, paired with an MAE-style decoder ($\mathcal{D}_S$ & $\mathcal{D}_F$) whose design matches the encoder’s size [15]. Throughout both MF and downstream fine-tuning, both BrainHarmonix-S and BrainHarmonix-F were *frozen*, providing modality latents only. For downstream fine-tuning, we average-pooled the brain hub tokens to generate a global multimodal representation followed by a linear projection head. The main results in Section 4.3 were all based on $N_H = 128$ 1D tokens. We employed FlashAttention [24, 25] in our self-attention implementation to improve computational efficiency and reduce memory usage. Each pre-training process utilized 8 NVIDIA H100 GPUs (80GB). The pretraining of $\mathcal{H}$ with 128 1D tokens took around 10 hours. The readers are referred to Appendix B for detailed optimization settings.

4.3 Main results

BrainHarmonix demonstrated strong generalization capabilities across neurodevelopmental and neurodegenerative disorder diagnoses as well as cognition prediction (Table 1, 2). As the first multimodal brain foundation model, BrainHarmonix was benchmarked against both structure-based and function-based neuroimaging models. For structure-based comparisons, we included BrainMVP [6], a state-of-the-art structural foundation model originally designed for multi-parametric MRI. Given its incompatibility of pretraining with T1 images only, we adopted BrainMVP’s pretrained weights and fine-tune it on downstream datasets. Many task-specific models for fMRI based on deep learning were proposed before foundation models emerged. These models could only be applied to specific tasks rather than a wide range of downstream applications [26, 27, 22, 30, 31]. For functional comparisons, BrainHarmonix was evaluated against both task-specific (BrainNetCNN [26], BrainGNN [27], and BrainNetTF [22]) and foundational fMRI models (BrainMass [7], BrainLM [4], and Brain-JEPA [5]). Previous brain dynamics foundation models, including BrainLM and Brain-JEPA, are not able to handle heterogeneous TRs. Consequently, these models were only assessed on datasets with homogeneous TRs (PPMI, ADNI, and HCP-A) following their original downsampling strategies (Table 2). BrainMass was pretrained on our pretraining datasets following the original settings.

Table 2: Comparison on neurodegenerative disease diagnosis and cognition prediction (standard deviations in Table 5). Task details in Section 4.1.

Model	Morphology	Dynamics	Multi-TR	PPMI		ADNI		HCP-A	
				ACC%	F1%	ACC%	F1%	MAE	ρ
Structure-based models									
BrainMVP ¹ [6]	✓	✗	N.A.	58.94	50.71	57.41	54.88	5.80	0.25
BrainMVP ² [6]	✓	✗	N.A.	55.04	40.82	60.61	44.67	5.39*	0.36
BrainHarmonix-S	✓	✗	N.A.	59.69	51.04	57.59	56.09	6.05	<u>0.38</u>
Function-based models									
BrainNetCNN [26]	✗	✗	✓	56.59	46.59	56.57	54.59	6.82	0.23
BrainGNN [27]	✗	✗	✓	58.14	47.79	58.59	57.27	6.78	0.22
BrainNetTF [22]	✗	✗	✓	58.92	48.56	60.61	58.00	6.70	0.25
BrainMass [7]	✗	✗	✓	59.77	48.31	59.60	56.73	6.45	0.28
BrainLM [4]	✗	✓	✗	53.49	44.58	57.58	59.57	7.03	0.25
Brain-JEPA [5]	✗	✓	✗	60.36	48.76	59.60	60.78	<u>5.62</u>	0.26
BrainHarmonix-F	✗	✓	✓	<u>62.79</u>	<u>52.90</u>	<u>61.62</u>	<u>64.80</u>	5.77	0.30
Multimodal model									
BrainHarmonix	✓	✓	✓	64.34*	56.40*	64.65*	68.75*	6.56	0.42*

¹UniFormer [28] as backbone; ²UNET3D [29] as backbone.

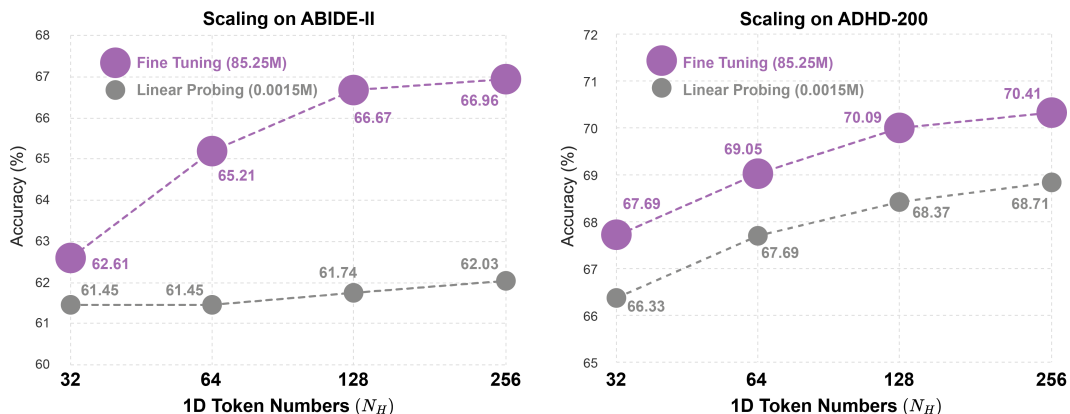


Figure 5: **Scaling across different numbers of 1D tokens in fine-tuning and linear probing.** We plot both fine-tuning (85.25M trainable parameters) and linear probing (0.0015M trainable parameters). Increasing the token count from 32 to 256 steadily improves accuracy before reaching a plateau. Notably, *even our linear-probing approach achieves very strong performance, surpassing prior state-of-the-art results* despite using a minimal set of learnable parameters.

Overall, BrainHarmonix consistently outperformed both structure-based and function-based models. Among its ablations, BrainHarmonix-F, which exclusively captures functional dynamics, achieved superior performance compared to existing fMRI models, highlighting the effectiveness of modeling heterogeneous dynamics from large-scale neuroimaging data. On the other hand, BrainHarmonix-S achieved performance that is superior or comparable to BrainMVP through an MAE framework, pre-trained on large-scale T1 datasets without multi-parametric MRI. The performance of BrainHarmonix-S can be attributed to its pretraining on a significantly larger T1 dataset, resulting in more robust brain morphological representations. The further improvements observed after structural-functional fusion in BrainHarmonix underscore the significance of integrating multimodal heuristics for comprehensive human brain representation.

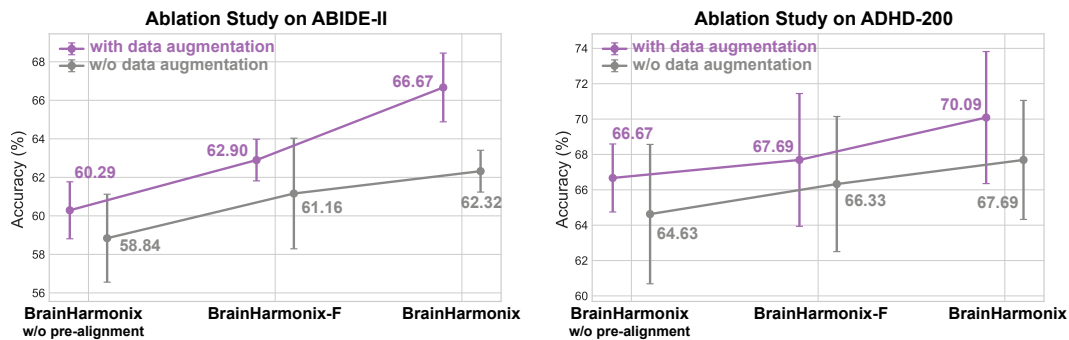


Figure 6: **Ablation study showing the impact of pre-alignment, data augmentation, and multimodal fusion.** In BrainHarmonix-F without pre-alignment, we stuck to the brain gradient positioning in Brain-JEPA, while the others (BrainHarmonix-F and BrainHarmonix) were injected with geometric harmonics on top of brain gradient (Section 4.2). Results were averaged over three random splits with error bars indicating standard deviations.

4.4 1D token scaling & linear probing

We investigated how performance changed when we varied the number of 1D tokens from 32 to 256 and evaluated both fine-tuning and linear probing (Figure 5). As the number of tokens grows, model accuracy consistently increases - highlighting the benefit of richer token-based representations - yet the gains begin to saturate from 128 to 256 tokens. Notably, our linear-probing approach, which uses only a simple linear head on top of the frozen BrainHarmonix, already achieves performance that outperforms previous advanced baselines. This underscores the strength of the learned brain representations and their capacity to generalize even with minimal downstream adaptation.

4.5 Ablation study

We compared BrainHarmonix with its ablated versions in Figure 6. The comparison between the purple ("with data augmentation") and gray ("w/o data augmentation") lines demonstrates the consistent performance gains from augmenting fMRI time series through multi-TR downsampling. It illustrates the effectiveness of enriching the model's temporal representation. Comparing the center bars ("BrainHarmonix-F") to the left ("BrainHarmonix-F w/o pre-alignment") further reveals that pre-alignment of fMRI signals to cortical geometry significantly boosts performance. Finally, the rightmost bar ("BrainHarmonix") underscores the value of fusing structural and functional information: integrating these two complementary views of the brain yields the highest accuracy.

4.6 Latent space analyses and interpretation

Table 3: Comparison of significant modes in t-SNE dimensions

Dimension	Model	# Significant Modes	Avg. P-value	Avg. Correlation
<i>Dim1 in t-SNE</i>				
	Brain-JEPA [5]	7	0.00769	0.1562
	BrainHarmonix-F	12	0.00456	0.1717
<i>Dim2 in t-SNE</i>				
	Brain-JEPA [5]	8	0.0115	0.1506
	BrainHarmonix-F	15	0.00477	0.1726

= number of; Avg. = average; Dim = dimension.

We extracted fMRI embeddings from BrainHarmonix-F and Brain-JEPA (400 ROIs, each represented by a 768-dimensional embedding) and applied t-SNE to project these embeddings onto a 2D plane. Specifically, we correlated each dimension of the t-SNE embedding with each of the 200 geometric harmonic modes across 400 ROIs. Compared to Brain-JEPA, our geometry-constrained embeddings

exhibit a greater number of significantly correlated modes ($p < 0.05$), with higher correlation strengths and significance levels on the top 5 most significant modes (Table 3). On the other hand, we applied the Fisher r -to- z transformation to all correlations from the 200 harmonics for each model and conducted a two-sample t -test. Results demonstrate that correlations from our model are significantly higher overall. The correlation strength and statistical significance from the top modes, along with the overall comparison, confirm that our model is constrained by structural information more than Brain-JEPA.

We examined the attention patterns between the 128 learned 1D tokens and the modality-specific tokens (400 fMRI ROI + 1200 T1 tokens) in ASD diagnosis using ABIDE-II data. For the 400 fMRI ROI tokens, each is obtained by averaging all tokens within the corresponding ROI. We found differentiation in modality attention among the 1D tokens: 93/128 tokens attended exclusively to fMRI, 30/128 exclusively to T1, and 5/128 tokens exhibited cross-modal attention. For cross-modal tokens, we found that they exhibited key structure-function coupling such as medial prefrontal cortex in brain morphometry and default model network in brain dynamics, which have previously been demonstrated in the literature to be associated with ASD.

For the 93 fMRI-specific tokens, further analysis revealed network-level functional differentiation relevant to ASD behavioral traits. Specifically, 60/93 were network-specific, predominantly focusing on a single brain network (with $>70\%$ salient ROIs within one network), while the remaining 33 were identified as “bridge” tokens capturing interactions across multiple networks. Among the most salient network-specific tokens, temporoparietal network (implicated in social perception and language processing deficits), somatomotor network (associated with sensorimotor integration impairments), and default mode network (linked to mentalizing deficits) emerged prominently. The identified “bridge” tokens primarily captured interactions involving default, limbic, and control networks, reflecting impaired integration across sensorimotor, socioemotional, and higher-order cognitive processes - a mechanism implicated in pathophysiology of ASD.

5 Conclusion

In this paper, we introduced **Brain Harmony (BrainHarmonix)**, the first multimodal brain foundation model that unifies structural morphology and functional dynamics into compact 1D tokens. By integrating geometric harmonics for structural-functional pre-alignment and introducing the Temporal Adaptive Patch Embedding (TAPE) for handling heterogeneous repetition times (TRs) in fMRI datasets, BrainHarmonix effectively bridges critical gaps existing in previous brain representation learning frameworks. Our approach provides a unified, expressive latent space that significantly enhances the representation of complex brain morphology and dynamics. Extensive experiments demonstrated BrainHarmonix’s superior generalization across diverse neuroimaging benchmarks, consistently outperforming state-of-the-art models in neurodevelopmental and neurodegenerative disorder classification and cognition prediction. BrainHarmonix is positioned to fundamentally advance AI-driven neuroscience research and clinical applications through multimodal neuroimaging.

6 Limitations and Future Work

We acknowledge several limitations in our study and highlight directions for future work. First, although BrainHarmonix was pretrained on the largest curation of structural-functional neuroimaging datasets to date, the age distribution of the data could be further expanded to better represent the entire human lifespan, particularly infancy and young adulthood. On the other hand, jointly optimizing the unimodal encoders and the fusion module could potentially lead to further performance gains. Exploring efficient training strategies for such long-sequence transformer models, particularly in the context of neuroimaging, is a promising direction. Beyond the demonstrated gains in neuropsychiatric disease diagnosis and cognition prediction, our multimodal brain foundation model, BrainHarmonix, holds promise -if further developed- as an AI-driven brain digital twin: a neuroscientific tool capable of validating and potentially uncovering novel neuroscience insights such as specific brain structure-function coupling related to human behavioral phenotypes. Realizing this potential, however, will require rigorous evaluation across diverse tasks and populations as well as systematic investigation into the model’s interpretability and translational pathways.

Acknowledgements

This study was supported by the Singapore National Medical Research Council (NMRC/OFLCG19May-0035, NMRC/CIRG/1485/2018, NMRC/CSA-SI/0007/2016, NMRC/MOH-00707-01, NMRC/CG/435 M009/2017-NUH/NUHS, CIRG21nov-0007, HLCA23Feb-0004, and OFIRG24Jul-0049), RIE2020 AME Programmatic Fund from A*STAR, Singapore (No. A20G8b0102), Ministry of Education (MOE-T2EP40120-0007 & T2EP2-0223-0025, MOE-T2EP20220-0001), and Yong Loo Lin School of Medicine Research Core Funding, National University of Singapore, Singapore. We would like to acknowledge that GPU resource used in this work was supported by NUS IT's Research Computing group using grant numbers NUSREC-HPC-00001.

References

- [1] Nikos K Logothetis. What we can do and what we cannot do with fmri. *Nature*, 453(7197): 869–878, 2008.
- [2] Russell A Poldrack and Martha J Farah. Progress and challenges in probing the human brain. *Nature*, 526(7573):371–379, 2015.
- [3] Vince D Calhoun and Jing Sui. Multimodal fusion of brain imaging data: a key to finding the missing link (s) in complex mental illness. *Biological psychiatry: cognitive neuroscience and neuroimaging*, 1(3):230–244, 2016.
- [4] Josue Ortega Caro, Antonio Henrique de Oliveira Fonseca, Syed A Rizvi, Matteo Rosati, Christopher Averill, James L Cross, Prateek Mittal, Emanuele Zappala, Rahul Madhav Dhodapkar, Chadi Abdallah, et al. Brainlm: A foundation model for brain activity recordings. In *The Twelfth International Conference on Learning Representations*.
- [5] 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.
- [6] Shaohao Rui, Lingzhi Chen, Zhenyu Tang, Lilong Wang, Mianxin Liu, Shaoting Zhang, and Xiaosong Wang. Brainmvp: Multi-modal vision pre-training for brain image analysis using multi-parametric mri. *arXiv preprint arXiv:2410.10604*, 2024.
- [7] Yanwu Yang, Chenfei Ye, Guinan Su, Ziyao Zhang, Zhikai Chang, Hairui Chen, Piu Chan, Yue Yu, and Ting Ma. Brainmass: Advancing brain network analysis for diagnosis with large-scale self-supervised learning. *IEEE Transactions on Medical Imaging*, 2024.
- [8] James C Pang, Kevin M Aquino, Marianne Oldehinkel, Peter A Robinson, Ben D Fulcher, Michael Breakspear, and Alex Fornito. Geometric constraints on human brain function. *Nature*, 618(7965):566–574, 2023.
- [9] 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.
- [10] 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.
- [11] Adriana Di Martino, David O'connor, Bosi Chen, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Joshua H Balsters, Leslie Baxter, Anita Beggato, 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.
- [12] Catie Chang and Gary H Glover. Time–frequency dynamics of resting-state brain connectivity measured with fmri. *Neuroimage*, 50(1):81–98, 2010.

- [13] Vince D Calhoun, Robyn Miller, Godfrey Pearlson, and Tulay Adalı. The chronectome: time-varying connectivity networks as the next frontier in fmri data discovery. *Neuron*, 84(2): 262–274, 2014.
- [14] Joonhyeong Park, Byoungwoo Park, Chang-Bae Bang, Jungwon Choi, Hyungjin Chung, Byung-Hoon Kim, and Juho Lee. A foundational brain dynamics model via stochastic optimal control. *arXiv preprint arXiv:2502.04892*, 2025.
- [15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
- [16] Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. Flexivit: One model for all patch sizes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14496–14506, 2023.
- [17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [18] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O’Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.
- [19] Karla L Miller, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, Saad Jbabdi, Stamatios N Sotiropoulos, Jesper LR Andersson, et al. Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523–1536, 2016.
- [20] Betty Jo Casey, Tariq Cannonier, May I Conley, Alexandra O Cohen, Deanna M Barch, Mary M Heitzeg, Mary E Soules, Theresa Teslovich, Danielle V Dellarco, Hugh Garavan, et al. The adolescent brain cognitive development (ab cd) study: imaging acquisition across 21 sites. *Developmental cognitive neuroscience*, 32:43–54, 2018.
- [21] Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and BT Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral cortex*, 28(9):3095–3114, 2018.
- [22] 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.
- [23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [24] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [25] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.
- [26] 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.
- [27] Xiaoxiao Li, Yuan Zhou, Nicha Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H Staib, Pamela Ventola, and James S Duncan. Braingnn: Interpretable brain graph neural network for fmri analysis. *Medical Image Analysis*, 74:102233, 2021.

- [28] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12581–12600, 2023.
- [29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.
- [30] Zijian Dong, Yilei Wu, Yu Xiao, Joanna Su Xian Chong, Yueming Jin, and Juan Helen Zhou. Beyond the snapshot: Brain tokenized graph transformer for longitudinal brain functional connectome embedding. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 348–357. Springer, 2023.
- [31] Zijian Dong, Joanna Su Xian Chong, Bing Cai Kok, and Juan Helen Zhou. Coop-dhgnn: a framework for joint classification and prediction of brain functional connectivity using sparse trajectory dataset with application to early dementia. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 4972–4978. IEEE, 2022.
- [32] Esten H Leonardsen, Han Peng, Tobias Kaufmann, Ingrid Agartz, Ole A Andreassen, Elisabeth Gulowsen Celius, Thomas Espeseth, Hanne F Harbo, Einar A Høgestøl, Ann-Marie De Lange, et al. Deep neural networks learn general and clinically relevant representations of the ageing brain. *NeuroImage*, 256:119210, 2022.
- [33] Florent Ségonne, Anders M Dale, Evelina Busa, Maureen Glessner, David Salat, Horst Karl Hahn, and Bruce Fischl. A hybrid approach to the skull stripping problem in mri. *Neuroimage*, 22(3):1060–1075, 2004.
- [34] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012.
- [35] Mark Jenkinson and Stephen Smith. A global optimisation method for robust affine registration of brain images. *Medical image analysis*, 5(2):143–156, 2001.
- [36] Maria A Di Biase, Robert E Smith, Andrew Zalesky, Caio Seguin, et al. Connectomes for 40,000 uk biobank participants: A multi-modal, multi-scale brain network resource. *NeuroImage*, 283:120407, 2023.
- [37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=Bkg6RiCqY7>.
- [38] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.
- [39] Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, et al. The parkinson progression marker initiative (ppmi). *Progress in neurobiology*, 95(4):629–635, 2011.
- [40] Jiaxing Xu, Yunhan Yang, David Huang, Sophi Shilpa Gururajapathy, Yiping Ke, Miao Qiao, Alan Wang, Haribalan Kumar, Josh McGeown, and Eryn Kwon. Data-driven network neuroscience: On data collection and benchmark. *Advances in Neural Information Processing Systems*, 36:21841–21856, 2023.
- [41] 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. fmripiprep: a robust preprocessing pipeline for functional mri. *Nature methods*, 16(1):111–116, 2019.

- [42] 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.
- [43] Michael P Harms, Leah H Somerville, Beau M Ances, Jesper Andersson, Deanna M Barch, Matteo Bastiani, Susan Y Bookheimer, Timothy B Brown, Randy L Buckner, Gregory C Burgess, et al. Extending the human connectome project across ages: Imaging protocols for the lifespan development and aging projects. *Neuroimage*, 183:972–984, 2018.
- [44] Jianxiao Wu, Jingwei Li, Simon B Eickhoff, Felix Hoffstaedter, Michael Hanke, BT Thomas Yeo, and Sarah Genon. Cross-cohort replicability and generalizability of connectivity-based psychometric prediction patterns. *Neuroimage*, 262:119569, 2022.
- [45] Joshua Faskowitz, Daniel Moyer, Daniel A Handwerker, Javier Gonzalez-Castillo, Peter A Bandettini, Saad Jbabdi, and Richard Betzel. Commentary on pang et al.(2023) nature. *bioRxiv*, pages 2023–07, 2023.
- [46] Zijian Dong, Yilei Wu, Zijiao Chen, Yichi Zhang, Yueming Jin, and Juan Helen Zhou. Prompt your brain: Scaffold prompt tuning for efficient adaptation of fmri pre-trained model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 512–521. Springer, 2024.

NeurIPS Paper Checklist

1. Claims

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

Answer: [Yes]

Justification: The last paragraph in introduction (Section 1) explicitly lists the paper's five core contributions. All claims have been verified through experimental results in Section 4.3, 4.4, and 4.5.

Guidelines:

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

2. Limitations

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

Answer: [Yes]

Justification: Limitations and future work are discussed in Section 6.

Guidelines:

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

3. Theory assumptions and proofs

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

Answer: [NA]

Justification: The paper does not include any theoretical result.

Guidelines:

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

4. Experimental result reproducibility

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

Answer: [Yes]

Justification: All datasets used in this work are publicly available, with detailed preprocessing information in Section 4.1 and Appendix A. The model checkpoints and source code are included in the supplementary materials, accompanied by detailed instructions in the README file.

Guidelines:

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

5. Open access to data and code

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

Answer: [\[Yes\]](#)

Justification: All datasets used in this work are publicly available. The model checkpoints and source code are included in the supplementary materials, accompanied by detailed instructions in the README file.

Guidelines:

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

6. Experimental setting/details

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

Answer: [\[Yes\]](#)

Justification: Full codebase is provided in the supplementary materials. Detailed training settings with hyperparameters are provided in Section 4.1, 4.2 and Appendix B.

Guidelines:

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

7. Experiment statistical significance

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

Answer: [\[Yes\]](#)

Justification: Statistical significance is clearly labeled in the main results (Table 1 and 2), with error bars in Table 4 and 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

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

8. Experiments compute resources

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

Answer: [Yes]

Justification: We have provided the computer resources in Section 4.2.

Guidelines:

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

9. Code of ethics

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

Answer: [Yes]

Justification: This work conforms the NeurIPS Code of Ethics in every respect.

Guidelines:

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

10. Broader impacts

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

Answer: [Yes]

Justification: Societal impacts are discussed in Appendix C.

Guidelines:

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

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

11. Safeguards

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

Answer: [NA]

Justification: The proposed research poses no such risks.

Guidelines:

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

12. Licenses for existing assets

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

Answer: [Yes]

Justification: The original papers that produced the code package or dataset are all properly cited.

Guidelines:

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

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

13. **New assets**

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

Answer: [\[Yes\]](#)

Justification: Code and model checkpoints are provided in the supplementary materials, with detailed documentation.

Guidelines:

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

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

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

Answer: [\[NA\]](#)

Justification: The datasets of human subjects used in this paper are all publicly available.

Guidelines:

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

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

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

Answer: [\[NA\]](#)

Justification: The datasets of human subjects used in this paper are all publicly available.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

Answer: [NA]

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

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Details of datasets

In this section, we detail the datasets used for pretraining and downstream evaluation of BrainHarmonix, describing their characteristics and associated preprocessing procedures.

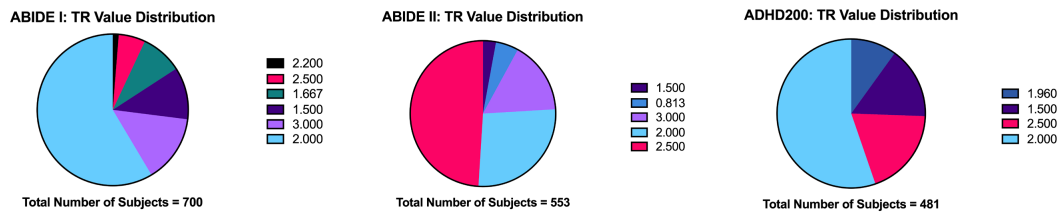


Figure 7: TR distributions of three multi-site datasets.

A.1 T1 preprocessing (shared by all datasets)

All T1-weighted images underwent a standardized preprocessing pipeline following [32]: first, images were skull-stripped using FreeSurfer [33]; then reoriented to match the standard orientation defined by FMRIB Software Library (FSL) [34]; and subsequently registered to the Montreal Neurological Institute (MNI) 152 template using FSL's linear registration tool (FLIRT) [35]. Finally, images were cropped to dimensions of $167 \times 212 \times 160$ voxels, with voxel intensities normalized to a $[0, 1]$ range.

A.2 Adolescent Brain Cognitive Development (ABCD)

The ABCD Study is the largest long-term study of brain development and child health in the United States [20]. We curated data from 11,221 participants, each with one/two visits (baseline (aged 107-133 months) and two-year follow-up), totaling 18,139 T1-weighted images and 30,771 resting-state fMRI time series (TR=0.8s). The fMRI preprocessing pipeline was as follows: fMRI were first aligned to T1-weighted anatomical scans using boundary-based registration. Respiratory pseudo-motion artifacts were reduced by applying a band-stop filter within the 0.31-0.43 Hz range. Frames exhibiting excessive motion - defined as framewise displacement (FD) greater than 0.3 mm or voxel-wise differentiated signal variance (DVARs) exceeding 50 - were flagged. Each flagged frame, along with the preceding frame and two subsequent frames, was censored; additionally, any uncensored data segments containing fewer than five consecutive frames were also removed. Subsequently, nuisance signals - including global, white matter, ventricular signals, six head motion parameters, and their temporal derivatives - were regressed out, with coefficients computed from uncensored data. Missing data from censored frames were interpolated using the Lomb-Scargle periodogram method, after which a band-pass filter of 0.009-0.08 Hz was applied. Finally, processed data were projected onto the FreeSurfer [33] fsaverage6 surface template and spatially smoothed using a Gaussian kernel with a 6 mm full-width at half-maximum (FWHM).

A.3 UK Biobank (UKB)

The UK Biobank is a large-scale biomedical database containing in-depth health information from UK participants, with the neuroimaging component represents the largest brain imaging study [18, 19]. We curated neuroimaging data from 43,112 participants aged 44 to 83, comprising 46,455 T1-weighted MRI and 40,162 fMRI time series (TR=0.735). Following Brain-JEPA [5], we used the preprocessed fMRI data from [36].

A.4 Autism Brain Imaging Data Exchange (ABIDE-I and ABIDE-II)

The Autism Brain Imaging Data Exchange (ABIDE) is a multi-site, open-access initiative that aggregates structural and resting-state fMRI scans - alongside rich phenotypic data - from individuals with autism spectrum disorder (ASD) and matched typically developing controls to accelerate reproducible neuroimaging research. We curated neuroimaging data from two releases: ABIDE-I [10], which contains 700 participants (320 control vs. 380 ASD) with paired T1 and fMRI data collected across 20 different sites, and ABIDE-II [11], which includes 553 participants (230 control

Table 4: Comparison on neurodevelopmental disorder diagnosis (standard deviations).

Model	Morphology	Dynamics	Multi-TR	ABIDE-I		ABIDE-II		ADHD-200	
				ACC%	F1%	ACC%	F1%	ACC%	F1%
Structure-based models									
BrainMVP ¹ [6]	✓	✗	N.A.	0.85	1.68	4.44	13.11	10.57	23.24
BrainMVP ² [6]	✓	✗	N.A.	8.20	5.85	2.95	18.85	1.18	2.71
BrainHarmonix-S	✓	✗	N.A.	6.32	5.74	3.79	4.03	1.56	4.36
Function-based models									
BrainNetCNN [26]	✗	✗	✓	1.38	0.75	1.33	1.90	0.59	3.70
BrainGNN [27]	✗	✗	✓	2.67	3.46	2.19	2.24	2.05	3.78
BrainNetTF [22]	✗	✗	✓	2.13	1.61	3.06	2.55	2.36	1.69
BrainMass [7]	✗	✗	✓	4.21	3.76	3.52	3.00	2.57	1.11
BrainLM [4]	✗	✓	✗	—	—	—	—	—	—
Brain-JEPA [5]	✗	✓	✗	—	—	—	—	—	—
BrainHarmonix-F	✗	✓	✓	1.38	2.30	1.32	1.06	4.60	2.77
Multimodal model									
BrainHarmonix	✓	✓	✓	4.31	1.30	2.18	1.02	4.57	3.31

¹UniFormer [28] as backbone; ²UNET3D [29] as backbone.

Table 5: Comparison on neurodegenerative disease classification and cognition prediction (standard deviations).

Model	Morphology	Dynamics	Multi-TR	PPMI		ADNI		HCP-A	
				ACC%	F1%	ACC%	F1%	MAE	ρ
Structure-based models									
BrainMVP ¹ [6]	✓	✗	N.A.	7.14	10.02	2.55	1.57	0.07	0.11
BrainMVP ² [6]	✓	✗	N.A.	1.34	1.43	5.25	15.02	0.11	0.06
BrainHarmonix-S	✓	✗	N.A.	3.55	6.31	3.06	9.50	0.21	0.11
Function-based models									
BrainNetCNN [26]	✗	✗	✓	1.10	0.82	1.75	3.77	0.59	0.07
BrainGNN [27]	✗	✗	✓	0.00	0.45	4.63	8.32	0.63	0.06
BrainNetTF [22]	✗	✗	✓	1.35	1.06	3.03	2.30	0.23	0.01
BrainMass [7]	✗	✗	✓	1.63	3.28	1.75	5.81	0.69	0.06
BrainLM [4]	✗	✓	✗	2.33	1.56	5.25	5.08	0.26	0.03
Brain-JEPA [5]	✗	✓	✗	2.17	3.67	1.43	0.24	0.61	0.14
BrainHarmonix-F	✗	✓	✓	2.33	3.65	1.75	1.91	0.73	0.11
Multimodal model									
BrainHarmonix	✓	✓	✓	3.55	6.31	4.63	3.87	0.56	0.12

¹UniFormer [28] as backbone; ²UNET3D [29] as backbone.

vs. 323 ASD) from 12 different sites. The distribution of heterogeneous TR values is shown in Figure 7. The fMRI preprocessing begins by de-obliquing and reorienting each fMRI run, then discarding the initial volumes before applying slice-timing correction. Head motion was corrected with FSL’s mcflirt [34], and the time series were coregistered to each subject’s T1 image using FreeSurfer’s bbregister [33]. Next, nuisance signals - including global, white matter, ventricular signals, six head motion parameters, and their temporal derivatives - were regressed out and the data were despiked and band-pass filtered (0.009-0.08 Hz). Finally, the processed data were normalized to MNI space.

Table 6: Pre-training settings.

config	value
<i>Common configs</i>	
d	768
optimizer	AdamW [37]
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$
learning rate schedule	warmup cosine schedule [38]
<i>BrainHarmonix-S & Harmonizer configs</i>	
start learning rate	0
learning rate	5×10^{-4}
final learning rate	0
weight decay schedule	constant
weight decay	0.05
warmup epochs	40
BrainHarmonix-S patch size	16
N_S	1200
BrainHarmonix-S total batch size	150×8 GPU cards = 1200
BrainHarmonix-S training epochs	800
BrainHarmonix total batch size	15×8 GPU cards = 120
BrainHarmonix training epochs	50
<i>BrainHarmonix-F configs</i>	
J	200 [8]
patch size, k^*	48
τ	$48 \times 0.735 = 35.28$ seconds
max number of tokens	18
N_F	$400 \times 18 = 7200$
start learning rate	2.5×10^{-6}
learning rate	5.7×10^{-5}
final learning rate	1×10^{-6}
weight decay schedule	cosine weight decay schedule [38]
weight decay	0.05
final weight decay	0.4
EMA momentum schedule	linear [38]
EMA start momentum	0.996
EMA final momentum	1
total batch size	64×8 GPU Cards = 512
warmup epochs	10
training epochs	100

A.5 Attention Deficit Hyperactivity Disorder (ADHD-200)

The ADHD-200 dataset is a multi-site, open-access repository of structural and resting-state fMRI scans with accompanying phenotypic measures from children and adolescents with attention-deficit/hyperactivity disorder (ADHD) and matched typically developing controls. We curated neuroimaging data from 481 participants (292 control vs. 189 ADHD) with paired T1 and fMRI data collected across 6 different sites. The distribution of heterogeneous TR values is shown in Figure 7. The ADHD-200 dataset underwent the same preprocessing procedure as the ABIDE datasets.

A.6 Parkinson’s Progression Markers Initiative (PPMI)

The PPMI is a longitudinal, multi-center, open-access dataset combining imaging, biospecimens, and detailed clinical assessments from Parkinson’s disease (PD) patients, prodromal cohorts, and healthy controls to accelerate biomarker discovery and disease-progression research [39]. We utilized the open benchmark repository [40] preprocessed by fMRIPrep [41], which contains data from 195 participants (15 control, 14 SWEDD, 53 Prodromal, and 113 PD patients).

A.7 Alzheimer’s Disease Neuroimaging Initiative (ADNI)

The ADNI is a longitudinal, multi-site study providing open-access neuroimaging, biomarker, genetic, and clinical data from cognitively normal (CN), mild cognitive impairment (MCI), and Alzheimer’s disease (AD) participants to advance early diagnosis and therapeutic research [42]. We curated neuroimaging data from 164 participants (83 CN vs. 81 MCI) with paired T1 and fMRI data. The preprocessing procedure is the same as the ABIDE datasets.

A.8 Lifespan Human Connectome Project Aging (HCP-A)

The HCP-A dataset is a Lifespan Human Connectome Project release that provides multimodal MRI and rich behavioral assessments from adults to elucidate brain connectivity changes across healthy aging [43]. The resting-state fMRI data in MNI152 space underwent ICA-FIX denoising. We then performed nuisance regression to control for 24 motion parameters, white matter signal, CSF signal, and their temporal derivatives following [44].

B Additional implementation details

The default optimization settings for pretraining are detailed in Table 6. We initialized all transformer blocks using the Xavier uniform method, as described in [15]. For downstream adaptation, the default setting follows MAE [15], except for using AdmaW for linear probing.

C Additional analysis & discussion

C.1 Synthetic testing of TAPE

Table 7: Performance comparison on HCP-A test sets

	MAE	Correlation
Original test set	6.56	0.42
Synthetic test set	6.69	0.39

To further evaluate TAPE’s effectiveness, we tested on both the original HCP-A test set and version with samples randomly downsampled by factors of 1 and 2 (equal probability, leading to TR values 1.6, 2.4). The comparable performance across conditions demonstrates TAPE’s robustness (Table 7).

C.2 Performance improvement: parameter count v.s. multimodal integration

We performed additional comparisons to better illustrate the performance gains from incorporating multiple modalities (2nd & 3rd columns in the Table 8 containing the results regarding accuracy

Table 8: Performance comparison across model sizes (accuracy%)

	Single Modality (fMRI)	Concat	22M	86M	307M
ABIDE-II	62.90	63.19	64.06	66.67	66.95
ADHD-200	67.69	68.36	69.39	70.09	70.40

(%)) and from introducing the harmonizer module for fusion (4th-6th columns). Specifically, we concatenated the embeddings from the frozen T1 and fMRI encoders and passed them through a trainable linear layer for the classification task. Furthermore, we conducted experiments using harmonizers of different sizes, ranging from 22M parameters to 307M parameters (4th-6th columns). The results above clearly demonstrate the performance improvements achieved both by adding modalities and by scaling the harmonizer module.

C.3 Ablation of ABCD dataset

Table 9: Model performance comparison on different training data (MCI classification on ADNI)

Model	ACC (%)	F1 Score (%)
Brain-JEPA using UKB only	59.60	60.78
BrainHarmonix-F using UKB only	60.67	63.34
BrainHarmonix-F using both UKB & ABCD	61.62	64.80

We conducted an ablation on ADNI for MCI classification by pretraining BrainHarmonix-F without ABCD data (Table 9). We found it still outperformed the original Brain-JEPA based on the same UKB dataset.

C.4 Ablation studies on ADNI

Table 10: Ablation on ADNI (accuracy%)

	BrainHarmonix w/o pre-alignment	BrainHarmonix-F	BrainHarmonix
with DA	61.35	61.62	64.65
w/o DA	60.07	60.11	62.94

DA: data augmentation.

We additionally performed ablation studies on ADNI (Table 10 regarding accuracy (%)), where we observed a similar trend and performance pattern, reinforcing the effectiveness of our proposed model design.

C.5 Generalization to Asian clinical cohorts

We extended our evaluation to an Asian clinical cohort collected by Memory, Ageing and Cognition Center (MACC), thereby assessing generalizability to non-Western populations and in real-world clinical scenarios. Specifically, we performed an additional task - classification of amyloid-positive/negative participants, which holds significant clinical value for AD prognosis and intervention. As shown in the Table 11, BrainHarmonix achieved state-of-the-art performance in this clinically relevant, in-house setting, underscoring its robustness and cross-population generalizability.

C.6 Evaluation across data portions

We conducted additional analyses by scaling the fine-tuning dataset using increasing proportions (20%, 40%, 60%, 80%, and 100%). The results regarding accuracy (%) are shown in the Table 12. Our results demonstrate a clear and consistent scaling of performance with increasing data portions. Notably, compared with prior leading baseline BrainMass 59.35% on ABIDE-II and 65.99% on

Table 11: Performance comparison on MACC

	ACC (%)	F1 (%)
BrainMVP	65.83	53.64
BrainHarmonix-S	67.68	56.67
BrainNetCNN	57.57	52.00
BrainGNN	62.61	40.57
BrainNetTF	63.03	57.57
BrainMass	64.65	57.93
BrainLM	63.64	54.03
Brain-JEPA	66.67	59.18
BrainHarmonix-F	68.69	62.50
BrainHarmonix	74.75*	65.57*

Table 12: Accuracy vs data portion (fine-tuning)

	Portion (%)	20%	40%	60%	80%	100%
ABIDE-II	Accuracy (%)	55.94	56.52	60.87	63.77	66.67
ADHD-200	Accuracy (%)	57.14	62.72	67.44	69.39	70.09

ADHD, BrainHarmonix achieves state-of-the-art performance even when fine-tuned on only 80% of the dataset, highlighting the efficiency and effectiveness of our pretrained representations.

On the other hand, we investigated the effect of using different portions of the pretraining dataset. Specifically, we applied identical sampling proportions to both the UKB and ABCD datasets for pretraining. The corresponding results regarding accuracy (%) are reported in the Table 13. We observe that the model’s performance improves as the portion of the pretraining dataset increases.

C.7 Scaling with increasing token numbers

For completeness, we have included results with 512 and 1024 tokens as references in addition to the results in the main content. As shown in the Table 14, the accuracy (%) remains relatively stable beyond 256 tokens, confirming our initial observation.

C.8 Efficiency evaluation

We included the pretraining time (on 8 NVIDIA H100 GPUs (80GB)), as well as finetuning time (on 1 H100 GPU) and inference time (on 1 H100 GPU) on ABIDE-II, corresponding to different model sizes (token counts) to provide a more comprehensive view of the computational cost in the Table 15. Larger model or more token counts lead to longer computing time.

C.9 Discussion on dynamic time warping (DTW)

Dynamic time warping (DTW) is an algorithm that measures the similarity between two temporal sequences, or time series, that may vary in speed or timing. It assumes a meaningful temporal correspondence between sequences. However, in the context of resting-state fMRI, there is no ground truth temporal alignment across individuals, as each subject’s brain dynamics evolve independently and asynchronously. Therefore, applying DTW across different scans would impose artificial temporal correspondences not supported by the data.

C.10 Discussion on geometric harmonics in neuroimaging community

There are critiques to [8], which focus on the paper’s claim that geometric harmonics, by themselves, can serve as a “winner-take-all” solution for brain dynamics reconstruction, thereby diminishing the role of the structural connectome [45]. However, the critiques do not affect the validity of our geometric pre-alignment. The harmonics in our work are only used to provide geometry-aware positioning, we make no claim that they can fully explain/reconstruct brain dynamics. On the other

Table 13: Accuracy vs data portion (pretraining)

Pretrain Portion	20%	40%	60%	80%	100%
ABIDE-II	59.12	62.90	64.35	65.21	66.67
ADHD-200	64.96	65.64	67.01	68.37	70.09

Table 14: Scaling with increasing token numbers (accuracy%)

Dataset	Method	32	64	128	256	512	1024
ABIDE-II	Finetune	62.61	65.21	66.67	66.96	67.53	66.96
	Linear Probe	61.45	61.45	61.74	62.03	62.32	62.89
ADHD-200	Finetune	67.69	69.05	70.09	70.41	70.41	70.75
	Linear Probe	66.33	67.69	68.37	68.71	68.37	69.05

Table 15: Training and inference time comparison across model sizes (token numbers) on ABIDE-II

	22M (128)	307M (128)	86M (32)	86M (64)	86M (128)	86M (256)	86M (512)	86M (1024)
Pretraining Time	5h 10m	17h 9m	9h 20m	9h 26m	9h 37m	9h 45m	10h 23m	11h 11m
FT Training Time	0h 21m 52s	1h 07m 28s	0h 25m 33s	0h 26m 54s	0h 27m 41s	0h 29m 54s	0h 30m 17s	0h 31m 54s
Inference Time	5.02s	7.19s	5.90s	5.89s	6.36s	5.77s	6.11s	7.47s

FT: fine tuning.

hand, the harmonics are averaged with large-scale functional gradients, so it does not have a winner-take-all basis. Future work can explore how structural connectome and other biological principles can be encoded into the model and whether they can further improve brain representation learning and generalizability.

C.11 Potential failure mode

In our current evaluations, one notable case where BrainHarmonix underperforms is on the ADHD-200. Its F1 is slightly lower than BrainHarmonix-F. This is likely due to motion artifacts in T1, as ADHD patients exhibit increased head motion during MRI acquisition. Such motion introduces noise and negatively affects structural data quality, potentially reducing multimodal fusion performance. Future work will explore methods to improve robustness against data-quality issues. On the other hand, although we have demonstrated data scaling effects, model performance under low-sample and few-shot learning scenarios remains an area for improvement. Future studies may address few-shot adaptation through approaches such as parameter-efficient fine-tuning or prompt-based tuning [46].

D Broader Impact

The integration and compression of multimodal neuroimaging signals could not only reduce storage and computational demands but also lays the groundwork for deployment on resource-constrained platforms. Coupled with its capabilities for various downstream tasks, it may accelerate exploratory analyses, support richer biomarker discovery, and drive improvements in diagnosis, prognosis, and personalized treatment planning. Moreover, its capacity to harmonize data across different scanners and acquisition protocols could enhance reproducibility and deepen our understanding of large-scale neuroimaging in both health and disease. However, these powerful capabilities also bring ethical responsibilities. Protecting patient confidentiality and ensuring data integrity are essential—deployments must include rigorous de-identification procedures and secure data pipelines.