FlexiBrain: Resolution-Agnostic Voxel-Level Encoding for Native fMRI

Mo Wang[1,2,3†], Wenhao Ye[1†], Junfeng Xia[1†], Minghao Xu[2], Hongkai Wen[2*], and Quanying Liu[1,3,4*]

[1] Department of Biomedical Engineering, Southern University of Science and Technology, China
[2] Department of Computer Science, University of Warwick, The UK
[3] Omni-Intelligence, Shenzhen, China
[4] Center for Language, Intelligence and Machines, Shenzhen Loop Area Institute, Shenzhen, China
[†]Equal contribution. [*]Corresponding authors: liuqy@sustech.edu.cn, hongkai.wen@warwick.ac.uk.

Abstract. The success of large-scale deep learning models in neuroscience is fundamentally constrained by severe data heterogeneity. Native fMRI data aggregated from diverse sources exhibit substantial variation in both spatial and temporal resolutions. Consequently, most existing frameworks rely on lengthy, rigid preprocessing pipelines that enforce uniformity across datasets. This practice introduces two critical limitations: (1) potential degradation of subject-specific anatomical information; (2) significant computational overhead, often requiring hours of processing per subject. Here, we propose FlexiBrain, a resolution-agnostic voxel-level encoding framework for native fMRI based on Mamba-JEPA. FlexiBrain defines patch sizes in real-world physical units and employs a dynamic patch resizing, thereby bypassing destructive spatial standardization while enabling direct ingestion of data in native space. We instantiate the framework using an efficient Mamba-JEPA backbone to model high-dimensional 4D fMRI signals. Across five diverse downstream neuroscience tasks, FlexiBrain consistently outperforms recent state-of-the-art methods, achieving gains of up to 12 percentage points without external data augmentation. Importantly, FlexiBrain functions as a seamless plug-in module, substantially reducing preprocessing costs and accelerating the development of robust voxel-level fMRI foundation models. Code is available at https://github.com/OneMore1/FlexiBrain.

Keywords: Functional MRI · Preprocessing · Disease diagnosis · Voxel-level model

1 Introduction

Functional Magnetic Resonance Imaging (fMRI) is an indispensable tool in modern neuroscience, providing crucial, non-invasive insights into human brain activity. By providing voxel-level images of brain activity, it has fundamentally

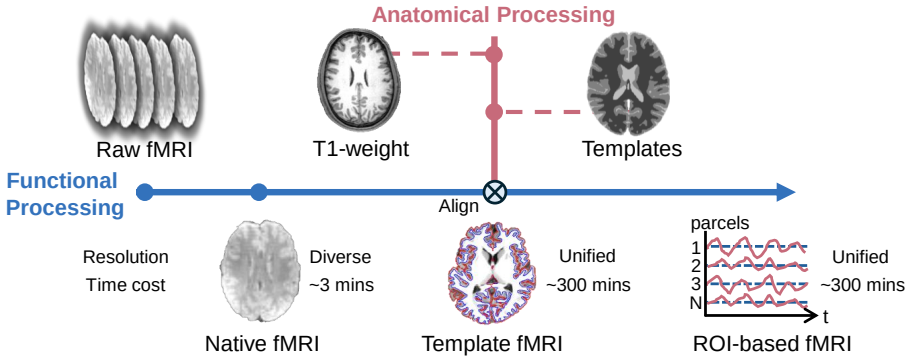

Fig. 1: fMRI preprocessing. Native fMRI is rapid and retains subject-specific anatomical information. However, it presents a challenge for deep learning due to its highly diverse resolution and non-aligned spatial coordinates. **Template fMRI** requires resource-intensive registration to align the image to a standard template brain, often consuming hours per subject. This eliminates individual differences, resulting in a unified and standardized resolution. **ROI-based fMRI** is derived from the registered Template fMRI using brain parcellation. It is reduced to a 2D representation (channels $\times$ time), summarizing activity across defined regions of interest.

advanced our understanding of cognitive processes and has been instrumental in diagnosing and tracking numerous neurological disorders since its inception in the 1990s [2]. Although early scanning protocols were often constrained by low spatial and temporal resolutions, recent technological advancements have yielded data with vastly higher resolution and developed a multitude of advanced acquisition sequences, providing researchers with more resolution options. This progress, while beneficial, has consequently introduced a critical challenge for deep learning methods: data aggregated from various sources are now characterized by substantial variability in spatiotemporal resolution. Although all fMRI data is inherently 4D (3D spatial information across a time dimension), this resolution heterogeneity is pronounced. Mainstream public datasets, for instance, exhibit spatial resolutions varying significantly from $2mm$ to $4mm$, with temporal sampling rates ranging drastically from $720ms$ to over $3000ms$ [6, 8, 22, 26]. Moreover, anatomical differences between subjects mean the explicit 4D data shape (i.e., voxel count) varies across individuals, further compounding this personalized variability.

To achieve the unified input required by deep learning models, the current mainstream fMRI pipeline typically addresses inherent data heterogeneity through comprehensive preprocessing. Despite the development of numerous distinct libraries, the core preprocessing methodology remains fundamentally similar [1, 7, 11, 12]. As shown in Fig. 1, the multi-stage preprocessing pipeline first converts the *Raw fMRI* series into *Native fMRI* through rapid operations such as slice-timing correction and head-motion correction, which are performed in each subject's original scanner space and typically require only a few minutes per

subject. Second, much more computationally demanding stage, the pipeline incorporates the subject's structural MRI (sMRI) to handle inter-subject anatomical variability: the sMRI undergoes intensive anatomical processing and the functional data are then non-linearly registered to a common standard template (e.g., the MNI152 template), a procedure that often takes hours. This yields the *Template fMRI*, in which all subjects' 4D fMRI volumes are warped into the same template-defined coordinate system and thus share an identical spatial structure, i.e., the same voxel grid, spatial layout, and standardized resolution as the template brain. Deep learning models are subsequently trained either directly on these normalized 4D volumes [15, 16, 19, 25, 29, 30, 33] or on dimensionality-reduced *ROI-based fMRI* representations derived using brain parcellation atlases in template space [5, 9, 10, 32, 34].

While full preprocessing pipelines make deep learning on fMRI technically feasible, they introduce several structural limitations. First, model outputs are highly sensitive to preprocessing software and parameter choices, and different pipelines can yield systematically different representations. Second, mapping all subjects into a predefined template space inevitably suppresses subject-specific anatomical variability and introduces interpolation and partial-volume artifacts. Third, such pipelines are computationally intensive and difficult to scale. In the era of large foundation models trained on thousands or tens of thousands of subjects, repeated normalization and resampling become both economically and scientifically inefficient [5, 10, 34]. A natural alternative is to operate directly on native-space fMRI. However, existing approaches typically rely on subject-specific projection networks to map native voxels into a shared latent space [23, 24]. While effective at small scale, this strategy prevents parameter sharing across heterogeneous datasets and becomes impractical for large-cohort foundation modeling.

In this paper, we introduce **FlexiBrain**, a resolution-adaptive voxel-level encoding framework for native fMRI. Instead of defining patches by fixed voxel counts, FlexiBrain parameterizes patch sizes in real-world physical units (millimeters and seconds). This design ensures that each token corresponds to a comparable anatomical and temporal receptive field across datasets with different spatial resolutions and fields of view. To enable parameter sharing without destructive resampling, we develop a resolution-aware dynamic patch resizing that continuously resizes embedding kernels according to the input voxel spacing. As a result, resolution heterogeneity no longer requires dataset-specific adapters or retraining, allowing scalable native-space foundation modeling. To further reduce the computational burden of voxel-level modeling, we instantiate FlexiBrain with a purpose-designed Mamba-JEPA backbone. The Mamba architecture replaces quadratic self-attention with linear-complexity state-space modeling, enabling scalable processing of high-resolution 4D fMRI volumes. Meanwhile, the JEPA formulation shifts the objective from full signal reconstruction to latent predictive learning. By masking large background regions and avoiding heterogeneous voxel-wise reconstruction losses, this design improves computational efficiency while focusing representation learning on informative spatiotemporal dynamics.

Across five diverse downstream neuroscience tasks, FlexiBrain consistently outperforms recent state-of-the-art methods, achieving performance gains of up to 12 percentage points, while operating directly in native space without template normalization.

Our main contributions are summarized as follows:

- We identify resolution heterogeneity as a fundamental scalability bottleneck for voxel-level fMRI foundation models, and propose a native-space framework that enables parameter sharing without template normalization or subject-specific adapters.
- We introduce FlexiBrain, a resolution-adaptive tokenization scheme that defines patches in physical units and dynamically resizes embedding kernels, allowing consistent receptive fields across heterogeneous spatial resolutions.
- Extensive experiments on five downstream diagnosis tasks demonstrate that FlexiBrain outperforms template-based methods while operating directly on native 4D fMRI data.

2 Related work

fMRI Preprocessing Recent pipelines for fMRI preprocessing follow a two-track routine [11]: anatomical processing (bias-field correction, skull stripping, tissue segmentation, surface reconstruction) and functional BOLD processing (reference volume creation, motion/susceptibility correction, slice-timing correction, coregistration to anatomy, and resampling). A practical design choice is the output space. Processing in each subject's *native space* is faster (fewer and cheaper resampling/warps) and preserves individual geometry and idiosyncratic anatomy, which benefits subject-specific modeling. In contrast, transforming to a *standard template space* (e.g., MNI) is slower due to non-linear normalization but unifies data into a common coordinate frame, simplifying batching, voxel correspondence, and cross-subject aggregation. Recent voxel-level models adopt the standard space for ease of model design (fixed shapes, consistent spatial priors) and benchmarking [15,16,28]. Yet forcing all subjects into a common template via non-linear warps inevitably introduces interpolation error and partial-volume artifacts, and can suppress subject-specific spatial variation (e.g., regional size/shape), potentially biasing downstream analyses. Beyond voxel-level methods, ROI-based fMRI represents the most widely adopted format in contemporary deep learning for neuroscience [5,9,10,34,35]. This format reduces the fMRI volume to a 2D representation (channels $\times$ time), where each channel contains the average signal within a specific Region-of-Interest (ROI) defined by a brain parcellation atlas. Since these atlases are typically designed and validated in the standard template space (e.g., MNI), ROI-based fMRI must inherently be derived from Template fMRI. Consequently, this highly popular data format inherits all the aforementioned disadvantages associated with non-linear normalization and damages fine-grained spatial information.

Deep Learning in Native fMRI While Native fMRI analysis is favored in traditional neuroscience for its ability to minimize information loss and preserve subject-specific anatomy, its usage in large-scale deep learning remains challenging due to significant cross-site and inter-subject heterogeneity. Data exhibit stark differences in acquisition protocols, such as the spatial/temporal resolution gap between ABIDE ($\approx 3mm, 2000ms$) [8] and UK Biobank ($\approx 2.4mm, 735ms$) [4]. Furthermore, hardware evolution promises an influx of increasingly finer-grained spatiotemporal data, exacerbating this resolution diversity. This heterogeneity has traditionally prevented the application of a single, unified deep learning model, forcing most works into the restrictive standard template space. Research directly leveraging Native fMRI is scarce and largely confined to specialized tasks or small-sample cohorts. For example, recent fMRI-to-image reconstruction models include MindEye2 [23, 24], which focuses on subject-specific mapping.

These methods select crucial voxels based on individual response features and train a dedicated mapping from the subject's native voxels to a shared latent space (e.g., CLIP embeddings) for each participant. While effective for individual decoding, this approach does not scale to variable resolution 4D fMRI from large, diverse cohorts.

3 Method

Fig. 2 illustrates the overall framework of FlexiBrain. Heterogeneous fMRI data with varying spatial and temporal resolutions are tokenized using a dynamic patch resizing defined in physical units. This mechanism ensures that patch tokens correspond to consistent physical receptive fields across heterogeneous fMRI acquisitions, enabling resolution-agnostic parameter sharing and stable representation learning directly in native space. By resizing existing filters to the target physical scale rather than relearning resolution-specific ones, the tokenizer can be seamlessly transferred across datasets without reconfiguration. The resulting fine-grained fMRI tokens are processed by a Mamba backbone, which enables efficient long-range modeling while significantly reducing computational cost. Training is guided by a JEPA-based proxy loss, which avoids heavy voxel-level reconstruction and mitigates loss imbalance arising from heterogeneous resolutions.

3.1 Dynamic Patch Resizing

Motivation. Voxel-level fMRI data exhibit substantial heterogeneity in spatial resolution and temporal sampling, making resolution-specific tokenizers brittle and difficult to transfer across datasets. To address the heterogeneity in 4D fMRI data across varying spatial resolutions and temporal resolution (TR), we introduce a Dynamic Patch Resizing (Fig. 2 a). Rather than relearning or discretely redefining filters for each resolution, we continuously resize a shared set of base kernels to the target physical scale, ensuring smooth parameter sharing

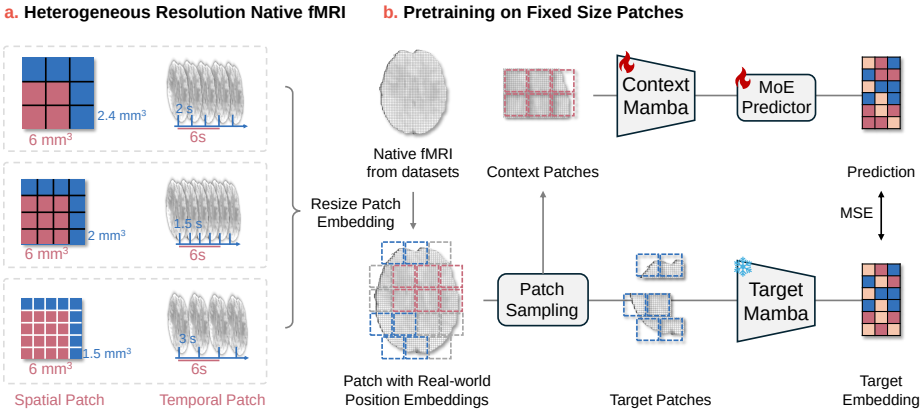

Fig. 2: FlexiBrain Pipeline. (a) FlexiBrain accepts Native fMRI with heterogeneous resolution from diverse datasets. To achieve resolution-agnosticism, the foundational spatial patch size (millimeters) and temporal patch size (seconds) are defined in real-world physical units. FlexiBrain uses a dynamic patch resizing to accommodate the variable number of voxels (due to differing resolutions) within a fixed physical patch size. **(b)** The physical patches with position embeddings in real-world coordinate system are sampled to create target and context patches for a self-supervised approach. The framework employs a novel Mamba-JEPA that is specifically designed to avoid background input and utilizes the Mixture-of-Experts module to mitigate representation collapse.

across heterogeneous inputs. This continuous adaptation preserves the functional semantics of the tokenizer across resolutions, enabling direct transfer without reconfiguration or retraining.

Dynamic Patch Size Determination. Given a target physical temporal duration τ (e.g., 6.0 seconds) and target spatial resolution $\rho = (\rho_x, \rho_y, \rho_z)$ (e.g., 12 mm), the subject-specific patch sizes k_t and k_d as the discrete temporal and spatial patch dimensions are dynamically calculated based on the input's time resolution (tr) and voxel sizes (v_x, v_y, v_z),$\lfloor\rceil$ indicates nearest integer rounding:

$$k_t = \left\lfloor \frac{\tau}{tr} \right\rceil, \quad k_d = \left\lfloor \frac{\rho_d}{v_d} \right\rceil \quad \text{for } d \in \{x, y, z\} \tag{1}$$

Continuous Kernel Resizing. Instead of randomly initializing distinct weights for every possible resolution, we maintain learnable base kernels $W_t^{(base)}$ for the temporal dimension and $W_{sp}^{(base)}$ for the spatial dimensions. To adapt these base kernels to the dynamically computed sizes k_t and (k_x, k_y, k_z), we utilize $\mathcal{R}^+$, a geometric operator that optimally projects the weights via the Moore-Penrose pseudo-inverse of the forward interpolation matrix to preserve their functional

mapping [3]:

$$W_t = \mathcal{R}^{+}_{1D}\left(W_t^{(base)} \to k_t\right) \tag{2}$$

$$W_{sp} = \mathcal{R}^{+}_{3D}\left(W_{sp}^{(base)} \to (k_x, k_y, k_z)\right) \tag{3}$$

Time-to-Space Forward Tokenization. First, a 1D temporal convolution with weight W_t and stride k_t aggregates short-range temporal dynamics within each physical window, transforming the input $X \in \mathbb{R}^{D\times H\times W\times T}$ into an intermediate feature map $X_t \in \mathbb{R}^{D\times H\times W\times D_m\times T'}$, where $T' = T/k_t$ and D_m denotes the intermediate channel dimension.

To enable efficient spatial tokenization, we fold the reduced temporal dimension T' into the channel dimension,

$$\tilde{X}_t = \mathrm{Reshape}(X_t) \in \mathbb{R}^{(D_m\cdot T')\times D\times H\times W},$$

such that temporal information is encoded as channel-wise features. Correspondingly, the spatial kernel W_{sp} is repeated T' times along its input channel dimension to match $\tilde{X}_t$, reflecting the assumption that spatial structure is invariant to temporal indexing. A 3D convolution with stride (k_x, k_y, k_z) is then applied to project the spatial grid into a D_{out}-dimensional latent space, yielding the final joint spatiotemporal patch embeddings. Finally, purely background patches are discarded using a subject-specific foreground mask, and the remaining valid tokens are flattened into a 1D sequence.

Continuous Physical 3D Positional Encoding. Discarding background patches disrupts the regular 3D grid, rendering standard discrete grid-based positional encodings ineffective, especially across subjects with varying spatial resolutions. To preserve precise anatomical awareness, we introduce a continuous physical 3D positional encoding. First, we determine the exact physical center of each valid patch in the anatomical world space. Let $p_i \in \mathbb{R}^3$ denote the discrete voxel-space center index of the i-th patch. Using the subject-specific affine transformation matrix $A \in \mathbb{R}^{3\times 3}$ and translation vector $t \in \mathbb{R}^3$ extracted from the NIfTI headers, the continuous physical coordinate $c_i \in \mathbb{R}^3$ is computed as:

$$c_i = p_i A^{\top} + t$$

To map these continuous physical coordinates into a high-dimensional representation, we apply a spatial scaling factor to mitigate unit variance and generate sinusoidal embeddings using logarithmically spaced frequency bands for each spatial axis. The concatenated features from all three axes are finally mapped to the target token dimension D_{out} via a learnable linear projection. Finally, these positional embeddings are added to the flattened 1D spatial token sequence.

3.2 Model Architecture: Mamba-JEPA

Motivation. Our architectural design is motivated by the need to efficiently model ultra-long, fine-grained token sequences while learning robust semantic

representations from noisy 4D fMRI data. Native-space voxel-level tokenization yields thousands of tokens per subject, rendering standard self-attention prohibitively expensive. Moreover, voxel-level reconstruction objectives are ill-suited for fMRI, as they emphasize high-frequency noise rather than meaningful functional structure. To address these challenges, we combine a linear-time state-space backbone (Mamba) with a latent predictive learning objective (JEPA), further augmented by structural asymmetry via a lightweight Mixture-of-Experts (MoE) to prevent representation collapse (Fig. 2 b).

Mamba for Ultra-Long Sequences. Let $\mathbf{X} \in \mathbb{R}^{D\times H\times W\times T}$ denote the input 4D fMRI signal. A dynamic patch resizing first tokenizes $\mathbf{X}$ into a sequence $\mathbf{Z}$ with real-world physical positional embeddings.

Even after aggressively discarding approximately 80% of the purely background patches, the resulting token sequence $\mathbf{Z}$ still contains around 4,000 tokens per subject. Processing sequences of this magnitude with standard self-attention incurs an intractable $\mathcal{O}(N^2)$ computational overhead. To accelerate training while preserving the ability to model complex spatial-temporal dependencies, we replace the standard self-attention mechanism with the Selective State Space Model (Mamba), which operates in linear time $\mathcal{O}(N)$.

During pre-training, a random masking strategy partitions the token sequence into a visible set $\mathcal{V}$ and a masked set $\mathcal{M}$. The context encoder f_θ, built upon stacked Mamba blocks, processes only the visible tokens $\mathbf{Z}_\mathcal{V}$.

Latent Predictive Objective. For the self-supervised proxy task, Masked Autoencoders (MAE) attempt to reconstruct masked tokens in the raw voxel space. However, fMRI data is notoriously noisy; forcing the network to reconstruct high-frequency voxel-level noise wastes computational capacity and degrades semantic representation quality. This issue is amplified in native-space, resolution-heterogeneous fMRI, where a reconstruction decoder would need to balance losses across different voxel grids and acquisition protocols. Meanwhile, given the variable resolution of fMRI inputs, fine-grained data would disproportionately dominate the MAE loss without appropriate weighting. Therefore, we adopt JEPA, which shifts the predictive reconstruction strictly into the semantic latent space.

The target encoder f_ξ (sharing the same structural design as f_θ) processes the ground-truth masked tokens $\mathbf{Z}_\mathcal{M}$ to generate latent targets $\mathbf{H}_\mathcal{M}$. To stabilize learning, the target encoder's parameters ξ are updated via an exponential moving average (EMA) of the context encoder's parameters θ and μ is the weight:

$$\xi \leftarrow \mu\xi + (1-\mu)\theta$$

Asymmetric Encoding via MoE to Prevent Collapse. A fundamental vulnerability of JEPA is representation collapse, where the encoder trivializes the objective by mapping all inputs to a constant vector. To explicitly combat this, we introduce structural asymmetry between the encoding pathways by appending a lightweight MoE module h_ψ to the tail of the context encoder. The MoE is not

intended to replace JEPA's standard asymmetric context–target pathways or the EMA target encoder; rather, it complements them by adding token-adaptive transformations on the context side. This architectural asymmetry prevents trivial solutions by ensuring that the predictive pathway cannot collapse to a constant mapping shared by both encoders. Instead of relying on complex routing equations, the MoE module simply employs a token-level router to adaptively compute soft mixture weights, routing each visible token to a subset of MLP-based expert networks. This yields enriched, token-adaptive context representations $\widehat{\mathbf{H}}_{\mathcal{V}} = h_{\psi}(f_{\theta}(\mathbf{Z}_{\mathcal{V}}))$.

State-Space Prediction and Loss. To perform the predictive task, we construct a full sequence by scattering the MoE-processed visible tokens $\widehat{\mathbf{H}}_{\mathcal{V}}$ back to their original sequence positions and inserting a learned mask token $\mathbf{m}$ at the masked locations $\mathcal{M}$. A shallow, Mamba-based predictor g_{ϕ} then diffuses information over this sequence to yield latent predictions $\widehat{\mathbf{H}}_{\mathcal{M}}$ at the masked locations.

We train the model by minimizing the ℓ_2 distance between the normalized predictions and the EMA-generated targets:

$$\mathcal{L} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \left\| \frac{\widehat{\mathbf{H}}_i}{\|\widehat{\mathbf{H}}_i\|_2} - \frac{\mathbf{H}_i}{\|\mathbf{H}_i\|_2} \right\|_2^2$$

Downstream Adaptation. For downstream inference, the pre-trained Mamba backbone extracts feature representations from the input tokens. A learnable class token is prepended to this sequence, and the concatenated representation is processed by a lightweight transformer classifier head. Finally, the feature corresponding to the class token is normalized and passed through an MLP to yield the final prediction logits.

4 Experiments

4.1 Data Preprocessing

We leveraged self-supervised training on three public neuroimaging datasets: the Autism Brain Imaging Data Exchange (ABIDE) [8]; the Alzheimer's Disease Neuroimaging Initiative (ADNI) [13]; the ADHD-200 Sample (ADHD) [6] and used the Parkinson progression marker initiative (PPMI) [18] as an external dataset to assess the model's generalization performance. These are all **multi-site datasets** comprising images collected across diverse centers and protocols; consequently, the temporal and spatial resolutions vary significantly across the cohorts. Details of datasets (e.g., voxel spacing and TR distribution) and task can be found in **Appendix C**. This exclusion was applied solely to match the subject set used by baselines that require structural scans, ensuring a fair comparison. For FlexiBrain,

all images retain their temporal and spatial resolution. We further verified that neither temporal resolution nor spatial resolution was associated with the

Table 1: Mean accuracy and standard deviations of predictions on disorder diagnosis over three random seeds. The best results are highlighted in **bold**, and second-best results are underlined. MCI: Mild cognitive impairment; AD: Alzheimer's disease. * denotes a large effect size with Cohen's d ≥ 0.8.

Model	ADHD-200		ABIDE		ADNI (MCI)		ADNI (AD)		PPMI	
	ACC%	F1%	ACC%	F1%	ACC%	F1%	ACC%	F1%	ACC%	F1%
BrainNetCNN [14]	$54.46_{\pm 2.47}$	$53.07_{\pm 3.89}$	$57.90_{\pm 3.25}$	$57.84_{\pm 3.21}$	$59.74_{\pm 4.50}$	$56.48_{\pm 6.04}$	$74.01_{\pm 3.91}$	$66.47_{\pm 4.18}$	$64.24_{\pm 2.41}$	$58.15_{\pm 1.44}$
BrainGNN [17]	$55.87_{\pm 3.88}$	$54.93_{\pm 4.46}$	$54.67_{\pm 3.72}$	$53.45_{\pm 3.28}$	$63.20_{\pm 7.38}$	$58.36_{\pm 12.50}$	$78.53_{\pm 2.59}$	$72.03_{\pm 5.38}$	$55.56_{\pm 4.81}$	$51.31_{\pm 2.99}$
BrainMass [34]	$60.78_{\pm 0.49}$	$\underline{60.05_{\pm 0.79}}$	$\underline{63.33_{\pm 1.18}}$	$\mathbf{65.35}_{\pm 2.00}$	$62.39_{\pm 0.60}$	$61.35_{\pm 0.60}$	$73.45_{\pm 3.91}$	$62.10_{\pm 0.70}$	$58.68_{\pm 4.21}$	$51.95_{\pm 2.23}$
Brain-JEPA [10]	$59.74_{\pm 0.23}$	$55.20_{\pm 6.40}$	$59.74_{\pm 2.64}$	$54.94_{\pm 11.98}$	$64.53_{\pm 0.60}$	$55.17_{\pm 0.66}$	$80.23_{\pm 0.80}$	$\underline{79.66_{\pm 1.03}}$	$47.91_{\pm 3.61}$	$48.60_{\pm 2.42}$
BrainHarmonix-F [9]	$59.74_{\pm 1.43}$	$58.76_{\pm 1.52}$	$54.03_{\pm 0.55}$	$43.04_{\pm 7.12}$	$60.61_{\pm 0.75}$	$51.86_{\pm 5.89}$	$76.84_{\pm 0.98}$	$71.17_{\pm 1.55}$	$65.63_{\pm 3.61}$	$55.19_{\pm 2.42}$
SwiFT [15]	$\underline{62.50_{\pm 1.50}}$	$59.99_{\pm 4.71}$	$61.02_{\pm 0.28}$	$61.36_{\pm 3.05}$	$65.38_{\pm 1.29}$	$\underline{72.54_{\pm 0.16}}$	$74.90_{\pm 1.88}$	$80.76_{\pm 0.33}$	$61.55_{\pm 2.52}$	$58.23_{\pm 4.20}$
NeuroSTORM [27]	$59.51_{\pm 1.01}$	$59.05_{\pm 1.01}$	$54.52_{\pm 2.40}$	$53.04_{\pm 0.70}$	$\underline{66.67_{\pm 1.06}}$	$62.17_{\pm 8.93}$	$\mathbf{84.26}_{\pm 0.80}$	$\mathbf{83.91}^{*}_{\pm 0.75}$	$\underline{68.25_{\pm 1.18}}$	$\underline{65.18_{\pm 1.19}}$
FlexiBrain	$\mathbf{69.76}^{*}_{\pm 1.75}$	$\mathbf{69.85}^{*}_{\pm 1.69}$	$\mathbf{65.45}^{*}_{\pm 0.69}$	$\underline{64.82_{\pm 0.52}}$	$\mathbf{79.16}^{*}_{\pm 1.48}$	$\mathbf{78.82}^{*}_{\pm 1.51}$	$\underline{82.34_{\pm 2.39}}$	$79.41_{\pm 4.42}$	$\mathbf{74.31}^{*}_{\pm 1.77}$	$\mathbf{70.54}^{*}_{\pm 1.19}$

class labels, indicating that resolution heterogeneity does not induce shortcut learning or spurious label–resolution confounds. All images for baselines were preprocessed following the specific requirements in the corresponding paper.

4.2 Disease Diagnosis

To validate the robustness of our model to heterogeneous data resolutions, we curated a mixed-resolution dataset for self-supervised pre-training by combining three public fMRI sources: ABIDE, ADHD-200, and ADNI. For downstream evaluation, the pretrained model was fine-tuned and benchmarked against multiple baselines on four tasks, using accuracy (ACC) and F1-score as evaluation metrics. Specifically, ABIDE and ADHD-200 were formulated as binary classification tasks, while ADNI was evaluated using two clinically meaningful binary settings: (i) mild cognitive impairment (MCI) vs. cognitively normal (CN), and (ii) Alzheimer's disease (AD) vs. CN. To explicitly assess out-of-distribution generalization, we further evaluated the pretrained model on the external, unseen PPMI dataset, where a three-class classification task was performed. Even compared to large-scale pretraining baselines [9, 10, 27, 34], our model consistently achieved strong performance as reported in Table 1. Detailed hyperparameter settings, and implementation details for both our model and the baselines are provided in **Appendices A and B**. We further audited this result under acquisition heterogeneity. PPMI is treated as a leave-one-dataset-out evaluation because it is not used during pretraining; when stratified by voxel spacing, FlexiBrain obtains 74.40% weighted F1 on the dominant higher-resolution bin and 66.96% weighted F1 on the lower-resolution bin (Supplementary Table 2). A controlled PPMI downsampling study shows that performance declines under spatial and temporal degradation, especially when both are degraded, indicating that the physical-unit tokenizer shares parameters across resolutions without collapsing all resolutions into an equivalent representation (Supplementary Table 3). We also include gender prediction on the heterogeneous ABIDE and ADHD-200 cohorts as an additional downstream check (Supplementary Table 4). A stricter

ABIDE leave-one-scanner-center-out analysis degrades when the held-out scanner center is unseen during both pretraining and downstream training; thus, FlexiBrain addresses resolution heterogeneity, but scanner-center domain shift remains an open challenge.

4.3 Choice of Preprocessing

The effect of the transformation on the predictions. Here we used different preprocessing states of fMRI to compare their performance by training the models from scratch. (i) Native fMRI / minimal **(ours)**: brain extraction only; head-motion parameters and slice-timing information are estimated, and the series remains in native space. (ii) T1w (weighted) fMRI: the native fMRI is first affinely registered to each subject's T1w anatomical scan, and not nonlinearly warped into a population template. (iii) Template fMRI (standard space): a nonlinear warp maps the data to a standard template, which is the most common preprocessing choice to align subjects into a standard space under heterogeneous resolutions. As shown in Fig. 3 Left, the native-space approach achieves the best performance while dramatically reducing preprocessing time from around 5 hours to 3 minutes. We hypothesize that the inferior performance of template fMRI is because spatial normalization reduces inter-subject variability and information entropy, while the requisite interpolation introduces noise artifacts.

Ablation of Patch Size. We systematically ablate the temporal duration τ and the spatial resolution ρ (Fig. 3 Right). We evaluated performance with ADHD by training it from scratch and find the best settings with $\tau = 6$ seconds and $\rho = 12$ millimeters. Increasing τ beyond 6 seconds and ρ beyond 12 millimeters leads to coarser patches that blur details, while very small τ or voxel spacings cause discretization, reducing sensitivity to subtle differences and harming generalization across protocols.

4.4 Choice of Architecture

Effectiveness of Mamba over ViT. To validate our backbone choice, we compare the Mamba backbone against a standard attention-based ViT under a comparable parameter budget ($\sim$50M). Profiling results reveal that the ViT backbone incurs substantial computational and memory overheads (**28.24** GFLOPs, **8.29** GB memory). In contrast, Mamba significantly reduces the resource footprint to just **18.59** GFLOPs and **1.97** GB memory. Beyond these critical efficiency gains, we further evaluated both pre-trained architectures on downstream tasks. As shown in Table 2, the Mamba backbone achieves superior downstream performance. We attribute this empirical advantage to Mamba's inherent capacity for modeling ultra-long, irregular voxel sequences (nearly 4,000 tokens), effectively bypassing the quadratic context bottleneck that degrades representation learning in standard self-attention mechanisms.

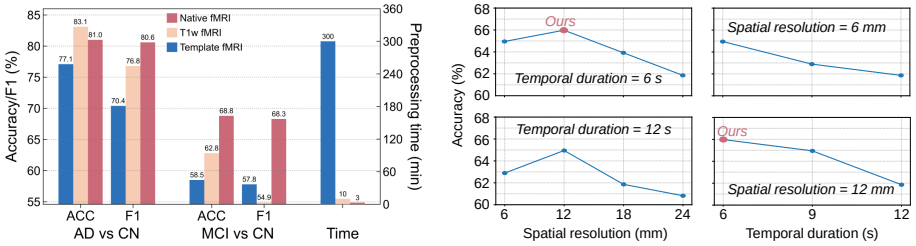

Fig. 3: Left: Ablation of preprocessing steps on ADNI dataset. Comparison of classification accuracy and preprocessing time between three fMRI states (Native fMRI, T1w fMRI, Template fMRI), with models trained from scratch. **Right: Ablation study on patch size.** The figure shows the accuracy of classification on ADHD dataset under different temporal duration and spatial resolution. Our settings with 6 s and 12 mm show best performance. AD: Alzheimer's disease; MCI: Mild cognitive impairment; CN: Cognitively normal.

Table 2: Ablation study on architecture. We compared four downstream tasks under different settings. The best results are highlighted in bold. MCI: Mild cognitive impairment; AD: Alzheimer's disease.

Backbone	Pre-training	MoE	ABIDE		ADHD-200		ADNI (MCI)		ADNI (AD)	
			ACC%	F1%	ACC%	F1%	ACC%	F1%	ACC%	F1%
Mamba	✓	✗	52.55	52.32	61.86	58.76	71.88	71.63	71.43	70.40
Mamba	✗	✓	59.85	59.66	65.98	61.37	68.75	68.25	80.95	**80.59**
ViT	✓	✓	61.31	60.41	63.92	59.03	66.66	66.60	76.85	72.92
Mamba	✓	✓	**65.45**	**64.82**	**69.76**	**69.85**	**79.16**	**78.82**	**82.34**	79.41

MoE and Pre-training. We compared the performance of our model with and without the MoE and pre-training. As shown in Table 2, there is a noticeable decline in model performance when the MoE module or the pre-training stage is removed. Pre-training on diverse data allows the model to learn a robust and generalizable feature representation. This representation serves as an initial distribution for fine-tuning, which in turn significantly boosts performance and data efficiency on downstream tasks. MoE effectively prevents model collapse. We visualized the tokens output by the target encoder (ordered by the position of tokens in the flattened 1D se-

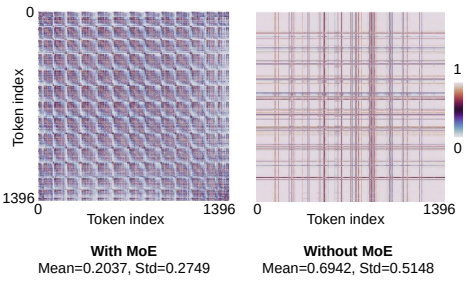

Fig. 4: Target embedding similarity matrix exhibits strong vertical/horizontal streaks, indicating low-rank, directional convergence to a few dominant directions.

quence of valid patch tokens) in cosine-similarity map (Fig. 4), which reveals a marked difference in representational geometry. With MoE, the mean similarity drops substantially and the pattern becomes fine-grained and predominantly near-diagonal, indicating content-dependent structure rather than convergence to a few dominant directions. Additional experiments are provided in **Appendix B**.

4.5 Interpretability

Transfer learning improves performance on a target task by leveraging knowledge acquired from a related source task. In general, stronger transfer gains indicate a closer relationship between the source and target tasks, thereby revealing their underlying structural and semantic connections [20, 21, 31, 36]. Following this principle, we employ a taskonomy-based analysis to investigate whether native-space fMRI data capture richer individual-specific and structural information than template-space data, leading to improved transferability.

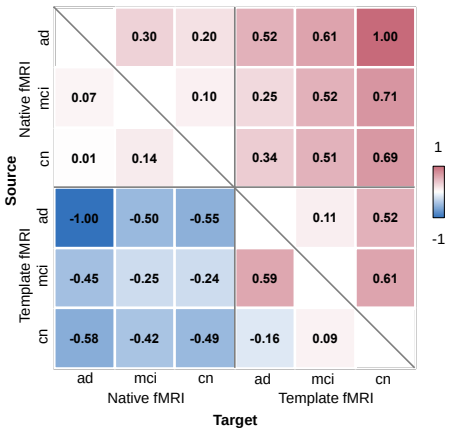

Fig. 5: Transfer task illustrates the normalized domain relations and disease-dependent transferability. Warm color denotes higher transferability, while cool colors indicate weaker transfer.

Specifically, we train six source models using ADNI data across two spatial domains—Native space (FSL) and Template space (MNI)—and three diagnostic groups (AD, MCI, CN). This results in 30 directed source-to-target transfer experiments. All models share the same architecture and training objective as used during pretraining. For each ordered task pair (s, t), we freeze the encoder pretrained on the source task and fine-tune a lightweight decoder on a small subset of the target task for 30 epochs, yielding a transfer loss $L_{s\to t}$, where lower values indicate stronger transferability. For visualization, we apply global min–max normalization to map transfer scores to $[-1, 1]$ and render the resulting taskonomy matrix as a heatmap (Fig. 5).

The taskonomy reveals clear relationships between spatial domain choice and diagnostic transferability. Cross-domain transfers from Native-space fMRI (FSL) to Template-space fMRI (MNI) outperform the reverse direction, suggesting that native-space representations preserve richer individual-specific and structural information, whereas template normalization improves spatial alignment at the cost of smoothing subject-specific signals. Within-domain transfers, both in Native and Template spaces, exhibit consistently lower transfer losses, indicating strong internal consistency under a shared spatial processing framework.

Across both within- and cross-domain transfers, AD demonstrates the strongest transferability, likely due to its pronounced and diverse pathological patterns. MCI shows intermediate transfer performance, consistent with its transitional clinical nature, while CN exhibits weaker cross-diagnostic transferability, reflecting more stable brain activity and reduced signal variability. Together, these results indicate that native-space representations contain richer transferable information and that disease-related heterogeneity plays a key role in determining transferability across diagnostic tasks.

5 Discussion

The transition from ROI-based analysis to fine-grained voxel-level modeling represents a natural evolution in fMRI research, analogous to the shift from static images to video modeling in computer vision. However, current voxel-level pipelines remain tightly coupled to heavy preprocessing, particularly spatial normalization to a common template. While such alignment enables cross-subject comparison, it also introduces structural and scalability constraints.

First, template normalization inevitably alters subject-specific anatomical variability through interpolation and smoothing, potentially degrading analyses that depend on fine-grained structural fidelity. Second, the computational cost and diversity of preprocessing standards limit reproducibility and scalability. These issues become critical in the era of foundation models, where training on tens of thousands of subjects makes repeated large-scale normalization economically and practically prohibitive. As open neuroimaging datasets continue to grow, this scalability bottleneck becomes increasingly prominent.

In this work, we propose FlexiBrain, a native-space voxel-level architecture that shifts the burden of alignment from data preprocessing to model design. The central challenge of native fMRI lies in resolution heterogeneity across datasets. While temporal inconsistencies have been partially addressed in prior work [9], voxel-level spatial heterogeneity has remained largely unexplored.

FlexiBrain addresses this challenge at the tokenization level. By defining patches in physical units (millimeters and seconds) and dynamically resizing embedding kernels, the model preserves consistent anatomical and temporal receptive fields across heterogeneous resolutions. Real-world coordinate positional encoding further enforces semantic consistency in native space, enabling parameter sharing without destructive resampling.

To make voxel-level modeling computationally feasible, we instantiate FlexiBrain with a Mamba-JEPA backbone. Linear-complexity state-space modeling replaces quadratic self-attention, enabling scalable processing of high-resolution 4D volumes. Meanwhile, the JEPA formulation shifts learning from voxel-wise reconstruction toward latent predictive objectives, avoiding heavy decoders and focusing representation learning on informative spatiotemporal dynamics. Across five downstream neuroscience tasks, FlexiBrain consistently outperforms recent state-of-the-art methods while operating directly in native space.

Limitations and Future Work. Although FlexiBrain demonstrates the feasibility of template-free native-space modeling, several limitations remain. First, we operate on minimally preprocessed native fMRI (e.g., motion-corrected) to ensure temporal consistency within sessions. Extending the framework to raw acquisition data would move closer to a fully end-to-end learning paradigm, potentially reducing reliance on traditional preprocessing pipelines and enabling the utilization of data currently excluded during quality control. Second, while FlexiBrain enables parameter sharing across resolutions, acquisition-specific artifacts and scanner variability remain open challenges for native-space modeling. Our leave-one-scanner-center-out analysis indicates that resolution-adaptive tokenization alone does not eliminate domain shift when a scanner center is absent from both pretraining and downstream training. Addressing these factors may further enhance robustness and cross-dataset generalization. Future work should combine physical-unit tokenization with explicit site harmonization, nuisance auditing, or domain-adaptive pretraining for deployment across unseen acquisition centers.

6 Conclusion

We present FlexiBrain, a resolution-adaptive voxel-level framework for native fMRI modeling. By redefining tokenization in physical units and introducing dynamic patch resizing, FlexiBrain removes resolution heterogeneity as a barrier to parameter sharing across datasets. Combined with an efficient Mamba-JEPA backbone, the framework enables scalable representation learning directly in native space without template normalization.

Our results demonstrate that native-space modeling can achieve superior performance while reducing computational overhead, suggesting a promising direction for future fMRI foundation models. We hope this work encourages a shift from data-centric alignment toward model-centric adaptation in large-scale neuroimaging research.

Acknowledgements This work was supported by National Science and Technology Major Project (2021ZD0200500), National Natural Science Foundation of China (3254100307, 62472206), National Key R&D Program of China (2025YFC3410000), Shenzhen Science and Technology Innovation Committee (RCYX20231211090405003, JCYJ20220818100213029), Guangdong Basic and Applied Basic Research Foundation (2026B1515020099), Guangdong S&T Program (2026B0101110003), Shanghai Municipal Special Program for Basic Research on General AI Foundation Models (2025SHZDZX026D05), Shenzhen Loop Area Institute (FPF10120250012), and the open research fund of the Guangdong Provincial Key Laboratory of Mathematical and Neural Dynamical Systems, the Center for Computational Science and Engineering at Southern University of Science and Technology, Shenzhen Key Laboratory of Smart Healthcare Engineering.

References

1. Bellec, P., Chu, C., Chouinard-Decorte, F., Benhajali, Y., Margulies, D.S., Craddock, R.C.: The neuro bureau adhd-200 preprocessed repository. Neuroimage **144**, 275–286 (2017)
2. Belliveau, J.W., Kennedy, D.N., McKinstry, R.C., Buchbinder, B.R., Weisskoff, R.M., Cohen, M.S., Vevea, J., Brady, T.J., Rosen, B.R.: Functional mapping of the human visual cortex by magnetic resonance imaging. Science **254**(5032), 716–719 (1991)
3. Beyer, L., Izmailov, P., Kolesnikov, A., Caron, M., Kornblith, S., Zhai, X., Minderer, M., Tschannen, M., Alabdulmohsin, I.M., Pavetic, F.: Flexivit: One model for all patch sizes. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 14496–14506 (2022), `https://api.semanticscholar.org/CorpusID:254685937`, accessed: 2026-06-28
4. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al.: The uk biobank resource with deep phenotyping and genomic data. Nature **562**(7726), 203–209 (2018)
5. Caro, J.O., Fonseca, A.H.d.O., Averill, C., Rizvi, S.A., Rosati, M., Cross, J.L., Mittal, P., Zappala, E., Levine, D., Dhodapkar, R.M., et al.: Brainlm: A foundation model for brain activity recordings. bioRxiv pp. 2023–09 (2023)
6. consortium, A..: The adhd-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. Frontiers in systems neuroscience **6**, 62 (2012)
7. Craddock, C., Benhajali, Y., Chu, C., Chouinard, F., Evans, A., Jakab, A., Khundrakpam, B.S., Lewis, J.D., Li, Q., Milham, M., et al.: The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. Frontiers in Neuroinformatics **7**(27), 5 (2013)
8. Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., 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)
9. Dong, Z., Li, R., Chong, J.S.X., Dehestani, N., Teng, Y., Lin, Y., Li, Z., Zhang, Y., Xie, Y., Ooi, L.Q.R., et al.: Brain harmony: A multimodal foundation model unifying morphology and function into 1d tokens. arXiv preprint arXiv:2509.24693 (2025)
10. Dong, Z., Li, R., Wu, Y., Nguyen, T.T., Chong, J., Ji, F., Tong, N., Chen, C., Zhou, J.H.: Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. Advances in Neural Information Processing Systems **37**, 86048–86073 (2024)
11. Esteban, O., Markiewicz, C.J., Blair, R.W., Moodie, C.A., Isik, A.I., Erramuzpe, A., Kent, J.D., Goncalves, M., DuPre, E., Snyder, M., et al.: fmriprep: a robust preprocessing pipeline for functional mri. Nature methods **16**(1), 111–116 (2019)
12. Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., et al.: The minimal preprocessing pipelines for the human connectome project. Neuroimage **80**, 105–124 (2013)
13. Jack Jr, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L. Whitwell, J., Ward, C., 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)

14. Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., Zwicker, J.G., Hamarneh, G.: Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. NeuroImage **146**, 1038–1049 (2017)
15. Kim, P., Kwon, J., Joo, S., Bae, S., Lee, D., Jung, Y., Yoo, S., Cha, J., Moon, T.: Swift: Swin 4d fmri transformer. Advances in Neural Information Processing Systems **36**, 42015–42037 (2023)
16. Kwon, J., Seo, J., Wang, H., Moon, T., Yoo, S., Cha, J.: Predicting task-related brain activity from resting-state brain dynamics with fmri transformer. Imaging Neuroscience **3**, imag_a_00440 (2025)
17. Li, X., Zhou, Y., Dvornek, N., Zhang, M., Gao, S., Zhuang, J., Scheinost, D., Staib, L.H., Ventola, P., Duncan, J.S.: Braingnn: Interpretable brain graph neural network for fmri analysis. Medical Image Analysis **74**, 102233 (2021)
18. Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., Coffey, C., Kieburtz, K., Flagg, E., Chowdhury, S., et al.: The parkinson progression marker initiative (ppmi). Progress in neurobiology **95**(4), 629–635 (2011)
19. Peng, Y.P., Cheung, V.K.M., Su, L.: Whole-brain transferable representations from large-scale fmri data improve task-evoked brain activity decoding. arXiv preprint arXiv:2507.22378 (2025), `https://arxiv.org/abs/2507.22378`, accessed: 2026-06-28
20. Qu, Y., Jian, X., Che, W., Du, P., Fu, K., Liu, Q.: Transfer learning to decode brain states reflecting the relationship between cognitive tasks. In: International Workshop on Human Brain and Artificial Intelligence. pp. 110–122. Springer (2022)
21. Qu, Y., Xia, J., Jian, X., Li, W., Peng, K., Liang, Z., Wu, H., Liu, Q.: Uncovering cognitive taskonomy through transfer learning in masked autoencoder-based fmri reconstruction. In: International Workshop on Human Brain and Artificial Intelligence. pp. 35–50. Springer (2024)
22. Satterthwaite, T.D., Connolly, J.J., Ruparel, K., Calkins, M.E., Jackson, C., Elliott, M.A., Roalf, D.R., Prabhakaran, K., Hopson, R., Behr, M., Qiu, H., Mentch, F.D., Chiavacci, R., Sleiman, P.M.A., Gur, R.C., Hakonarson, H., Gur, R.E.: The philadelphia neurodevelopmental cohort: A publicly available resource for the study of normal and abnormal brain development in youth. NeuroImage (2016)
23. Scotti, P., Banerjee, A., Goode, J., Shabalin, S., Nguyen, A., Dempster, A., Verlinde, N., Yundler, E., Weisberg, D., Norman, K., et al.: Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. Advances in Neural Information Processing Systems **36**, 24705–24728 (2023)
24. Scotti, P.S., Tripathy, M., Villanueva, C.K.T., Kneeland, R., Chen, T., Narang, A., Santhirasegaran, C., Xu, J., Naselaris, T., Norman, K.A., et al.: Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. arXiv preprint arXiv:2403.11207 (2024)
25. Sun, Y., Chahine, D., Wen, Q., Liu, T., Li, X., Yuan, Y., Calamante, F., Lv, J.: Voxel-level brain states prediction using swin transformer. ArXiv **abs/2506.11455** (2025), `https://api.semanticscholar.org/CorpusID:279392087`, accessed: 2026-06-28
26. Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., Consortium, W.M.H., et al.: The wu-minn human connectome project: an overview. Neuroimage **80**, 62–79 (2013)
27. Wang, C., Jiang, Y., Peng, Z., Li, C., Bang, C., Zhao, L., Lv, J., Sepulcre, J., Yang, C., He, L., Liu, T., Barron, D., Li, Q., Hirschtick, R., Kim, B.H., Li, X., Yuan, Y.: Towards a general-purpose foundation model for fmri analysis. arXiv preprint

arXiv:2506.11167 (2025), https://arxiv.org/abs/2506.11167, accessed: 2026-06-28
28. Wang, M., Peng, K., Tang, J., Wen, H., Liu, Q.: DCA: Graph-guided deep embedding clustering for brain atlases. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2026), https://openreview.net/forum?id=ypPxYsmZPx, accessed: 2026-06-28
29. Wang, M., Xia, J., Ye, W., Liu, E., Peng, K., Feng, J., Liu, Q., Wen, H.: Slimbrain: A data-and training-efficient foundation model for fmri data analysis. arXiv preprint arXiv:2512.21881 (2025)
30. Wang, M., Ye, W., Xia, J., Zhang, J., Pan, X., Xu, M., Deng, H., Wen, H., Liu, Q.: Omni-fMRI: A universal atlas-free fMRI foundation model. In: Forty-third International Conference on Machine Learning (2026), https://openreview.net/forum?id=Rc8th2rXXe, accessed: 2026-06-28
31. Xia, J., Li, W., Zhang, M., Guo, J.: Beyond single-source cognitive taskonomy:multi-source task relations through fmri transfer learning (2026), https://arxiv.org/abs/2606.26279, accessed: 2026-06-28
32. Xia, J., Ye, W., Pan, X., Shen, X., Wang, M., Liu, Q.: Brain-dit: A universal multi-state fmri foundation model with metadata-conditioned pretraining. arXiv preprint arXiv:2604.12683 (2026)
33. Xia, J., Ye, W., Zhang, J., Pan, X., Wang, M., Liu, Q.: Brainworld: A structural-prior-conditioned generative model for whole-brain 4d fmri dynamics. arXiv preprint arXiv:2606.17742 (2026)
34. Yang, Y., Ye, C., Su, G., Zhang, Z., Chang, Z., Chen, H., Chan, P., Yu, Y., Ma, T.: Brainmass: Advancing brain network analysis for diagnosis with large-scale self-supervised learning. IEEE Transactions on Medical Imaging (2024)
35. Ye, Z., Qu, Y., Liang, Z., Wang, M., Liu, Q.: Explainable fmri-based brain decoding via spatial temporal-pyramid graph convolutional network. Human Brain Mapping **44**(7), 2921–2935 (2023)
36. Zamir, A.R., Sax, A., Shen, W., Guibas, L.J., Malik, J., Savarese, S.: Taskonomy: Disentangling task transfer learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3712–3722 (2018)