BrainWorld: Modeling Whole-Brain Functional Dynamics through Predictive 4D fMRI Generation

Junfeng Xia[1] **Wenhao Ye**[1,3] **Junxiang Zhang**[1,4] **Xuanye Pan**[1]
Mo Wang[2,1,5,*] **Quanying Liu**[1,4,5,*]

[1]Department of Biomedical Engineering, Southern University of Science and Technology, Shenzhen, China
[2]Department of Computer Science, University of Warwick, The UK
[3]Shenzhen University, Shenzhen, China [4]Shenzhen Loop Area Institute, Shenzhen, China
[5]Omni-Intelligence, Shenzhen, China
[*]Co-corresponding authors: 12250099@mail.sustech.edu.cn; liuqy@sustech.edu.cn

Abstract

Whole-brain four-dimensional functional magnetic resonance imaging (4D fMRI) generation is valuable for modeling functional brain dynamics, yet existing fMRI foundation models mainly target representation learning and downstream prediction rather than conditional predictive generation. We introduce BrainWorld, a structural-prior-conditioned generative model for whole-brain 4D fMRI dynamics. BrainWorld uses structural magnetic resonance imaging (sMRI) as subject-level anatomical context and past functional connectivity as recent dynamic context to guide future fMRI generation, integrating both structural and functional conditions into the denoising process. Using 22 datasets for model development and evaluation, BrainWorld supports autoregressive generation of up to 400 frames, improves downstream performance through generated-example augmentation, and learns transferable multimodal representations that outperform baselines. Together, these results establish BrainWorld as a condition-aware generative framework for long-horizon brain dynamics modeling and multimodal representation learning. Code is available at this link.

1 Introduction

Modeling human brain activity is central to understanding how the brain functions. Imagination illustrates a generative aspect of brain function: the construction of coherent internal experiences from partial cues and prior knowledge [3]. Inspired by this perspective, generative modeling offers a promising approach to understanding how brain activity is organized and evolves over time [20]. Functional magnetic resonance imaging (fMRI) provides observations of whole-brain functional activity across space and time. Modeling these 4D signals requires capturing both their spatial patterns and temporal dependencies [26]. We therefore approach whole-brain fMRI modeling through conditional generation, learning the distribution of future functional signals given recent activity and contextual information.

From this perspective, generation serves not only to synthesize additional data, but also to learn and evaluate predictive models of functional dynamics. Generating future fMRI windows from preceding activity and available context provides an explicit learning objective for capturing relationships between past and future signals [49, 46]. The generated trajectories can then be compared with subsequent recordings to assess predictive fidelity over different horizons. Beyond prediction, the learned representations may support downstream analysis [63], while generated examples may improve learning when labeled recordings are limited [42]. These complementary uses motivate studying predictive generation as a framework for brain dynamics modeling, with representation transfer and data augmentation as additional applications.

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

Existing fMRI foundation models learn brain representations at region-of-interest (ROI), network, and voxel scales [8, 67, 58] through diverse pretraining objectives, including masked reconstruction [43, 8], latent prediction [12], contrastive learning [26], and diffusion-based modeling [63]. These advances provide a foundation for learning from large and heterogeneous neuroimaging datasets. Alongside learning informative representations, an important goal of brain dynamics modeling is to characterize how future functional activity may unfold given recent activity and contextual information. Conditional generation directly addresses this goal by learning the distribution of future fMRI signals. Although prior work has explored fMRI synthesis and prediction [23, 49, 46], long-horizon conditional generation of whole-brain voxel-level 4D signals remains relatively underexplored. Modeling such signals requires handling high-dimensional spatial structure and temporal dependencies, placing substantial demands on data and computation. Advances in neuroimaging resources and scalable generative models provide an opportunity to investigate this setting more systematically.

In this context, recent progress in video generation provides a promising modeling route for whole-brain 4D fMRI, as both tasks require long-horizon synthesis of high-dimensional spatiotemporal trajectories [7, 28, 53]. Diffusion Transformers (DiTs) [41] are appealing for whole-brain 4D fMRI generation, as they combine diffusion-based modeling of complex high-dimensional distributions with Transformer-based long-range dependency modeling and flexible condition injection. Importantly, different denoising stages and intermediate DiT layers capture information at multiple levels, from coarse global dynamics to fine-grained local details. This allows the model to function not only as a high-quality generator, but also as a source of informative neural representations [29, 35, 66, 59, 63]. This generative framework also offers a more natural formulation for structural-functional integration. Since brain function evolves within subject-specific anatomy, structural magnetic resonance imaging (sMRI) can play an asymmetric role as anatomical context, rather than being treated as another parallel input modality. Injecting structural information into the denoising process allows anatomical context to modulate the formation of functional dynamics. Based on these observations, we propose BrainWorld, a structural-prior-conditioned generative model for whole-brain 4D fMRI dynamics. BrainWorld uses a variational autoencoder (VAE) to compress voxel-level fMRI into a continuous latent space, where a conditional DiT models brain dynamics under sMRI conditioning, dynamic functional context, and optional stimulus conditions. This design supports condition-aware generation, dynamic representation learning, and generative structure-function fusion.

Our main contributions are threefold:

- We propose BrainWorld, a latent diffusion framework for modeling whole-brain functional dynamics through long-horizon predictive generation of 4D fMRI across diverse cohorts and brain states.
- We introduce a structural-prior conditioned denoising strategy that injects sMRI as subject-level anatomical context throughout the DiT backbone.
- We demonstrate BrainWorld's utility through three complementary evaluations: long-horizon predictive generation on four datasets, generated-example augmentation on two datasets, and representation transfer on downstream tasks.

2 Related Work

2.1 fMRI Foundation Models

Existing fMRI foundation models can be broadly organized by their pretraining objectives. Reconstruction-based methods learn representations through signal-reconstruction objectives [43, 58, 8, 54]. Joint-embedding predictive architecture (JEPA) methods replace direct reconstruction with predictive representation learning [12, 56, 57], emphasizing abstract latent prediction. Contrastive and large language model (LLM) alignment methods encourage transferable or semantically grounded latent representations, but are less directly aligned with explicit generative modeling of brain dynamics [67, 26, 61]. More recently, diffusion-based objectives have been explored for generative pretraining in fMRI modeling, although current formulations remain mainly atlas-based, temporally restricted, and without structural conditioning [63]. Taken together, these lines have substantially advanced fMRI representation learning and generative pretraining.

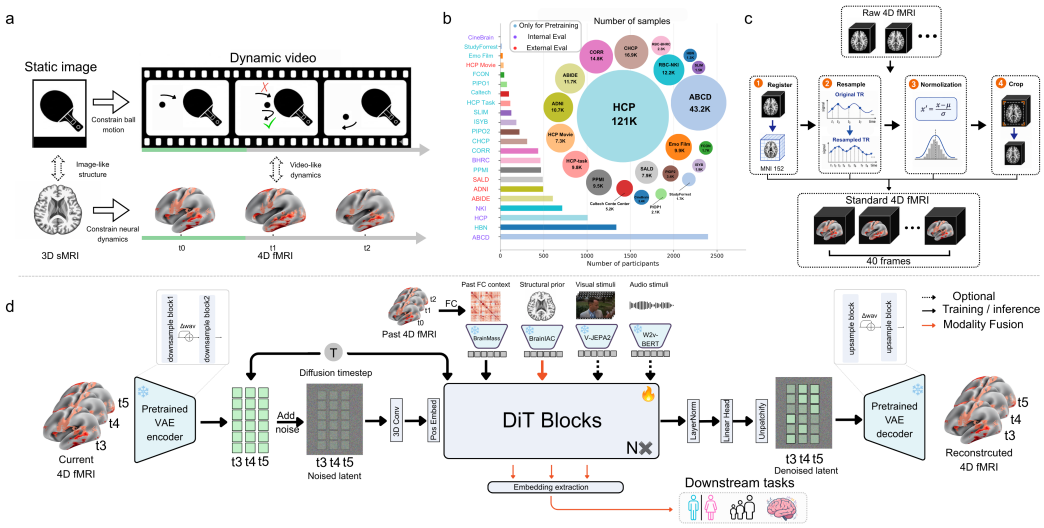

Figure 1: Overview. (a) Motivation. sMRI resembles a static image that provides an anatomical scaffold, whereas 4D fMRI resembles a dynamic video of brain activity. Like a racket influencing a ball's trajectory, sMRI provides anatomical context for fMRI dynamics, motivating structural-condition-aware generation. (b) Dataset. BrainWorld uses 22 fMRI datasets, grouped into pretraining-only, internal evaluation, and external evaluation cohorts. (c) fMRI preprocessing. Raw 4D fMRI is standardized by registration to the Montreal Neurological Institute 152 (MNI152) template, repetition-time (TR) resampling, signal normalization, and spatial cropping, yielding 40-frame inputs. (d) Architecture. BrainWorld encodes voxel-level 4D fMRI into a pretrained VAE latent space and trains a DiT to denoise latent brain dynamics under past functional connectivity, structural MRI, and optional video/audio conditions encoded by frozen encoders. The model supports condition-aware 4D fMRI generation and downstream representation learning via intermediate DiT features.

2.2 Medical Image Generation and fMRI Synthesis

In medical image generation, sMRI and computed tomography (CT) have established mature research paradigms [42]. These models offer not only strong generative capability, but also useful intermediate representations for downstream tasks, highlighting the broader promise of generative pretraining [17]. In contrast, fMRI generation remains less explored, especially for methods that directly model voxel-level whole-brain 4D spatiotemporal dynamics. Prior work addresses different representations and prediction targets, including ROI time series [23], functional connectivity matrices [40], and short-horizon voxel-level prediction [49]. Sequential modeling of task-based fMRI has been studied for synthetic data augmentation [55], while recent work explores event-conditioned flow matching for rest-to-task fMRI time-series synthesis [14]. Seo et al. [46] made an important step toward conditional whole-brain 4D fMRI synthesis. However, their formulation is primarily designed for task-specific conditional generation, leaving broader cross-state modeling of brain dynamics with subject-level structural conditioning less explored.

2.3 Structural-Functional Joint Modeling

Beyond fMRI generation itself, another relevant line concerns how structural information is integrated with functional dynamics. Structural-functional joint modeling has evolved from network-level fusion to fine-grained multimodal learning and, more recently, large-scale multimodal pretraining. Early studies fused structural and functional information through graph modeling or feature fusion to improve disease classification and diagnosis [69]. Later methods directly combined 3D sMRI and 4D fMRI, using co-attention, patch-level alignment, or latent contrastive objectives to mitigate cross-modal heterogeneity [5, 62]. Unlike earlier approaches trained end-to-end on relatively small datasets, BrainSymphony jointly models fMRI time series with diffusion-derived structural connectivity, while Brain Harmony compresses T1 morphology and fMRI dynamics into shared 1D tokens; both move toward multimodal foundation modeling through large-scale pretraining [25, 13]. Nevertheless, these

approaches largely treat structure and function as parallel or complementary inputs. In contrast, BrainWorld uses sMRI as subject-level anatomical context to condition fMRI dynamics generation, motivating generative structure-function fusion for downstream tasks.

3 Method

3.1 Overview

Given a current whole-brain 4D fMRI window and aligned auxiliary conditions, BrainWorld generates future fMRI dynamics while extracting transferable representations from the generative process. To avoid costly voxel-space modeling, we use a two-stage latent framework. First, a 4D fMRI variational autoencoder (VAE) is pretrained on large-scale multi-cohort data to produce a compact and reconstructable latent space. Second, a conditional Diffusion Transformer (DiT) models future latent dynamics under functional, structural, and stimulus-related conditions. At inference, future latent windows are autoregressively generated from the current window and available conditions, then decoded back to voxel-level 4D fMRI by the pretrained VAE decoder. Intermediate DiT states are further used as representations for downstream tasks, supporting both condition-aware generation and foundation-model-style representation learning.

3.2 Stage I: 4D fMRI VAE Pretraining

We first pretrain a frame-wise spatial VAE for 4D fMRI windows to provide a compact and reconstructable latent space for conditional diffusion modeling. To reduce background-dominated computation, each standardized 96^3 fMRI volume is cropped to remove non-brain background regions, yielding an input window $x_i \in \mathbb{R}^{40\times80\times96\times80}$. The VAE preserves the temporal axis and applies a shared 3D encoder frame by frame. With $C = 16$ latent channels, the resulting latent sequence is

$$u_i \in \mathbb{R}^{40\times16\times10\times12\times10}.$$

The encoder uses three spatial downsampling stages to reduce each frame from $80 \times 96 \times 80$ to $10 \times 12 \times 10$, while the decoder symmetrically uses three upsampling stages to reconstruct the original cropped resolution. Following the design of WF-VAE [30], we incorporate lightweight 3D Haar wavelet fusion blocks into intermediate features, where fixed Haar transforms capture low- and high-frequency spatial cues and residual fusion integrates them back into the feature stream. The VAE is trained with

$$\mathcal{L}_{\mathrm{VAE}} = \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{bg}}\mathcal{L}_{\mathrm{bg}} + \lambda_{\mathrm{wav}}\mathcal{L}_{\mathrm{wav}} + \beta\mathcal{L}_{\mathrm{KL}}. \tag{1}$$

The four terms are defined as

$$\begin{aligned} \mathcal{L}_{\mathrm{rec}} &= \frac{\sum_v w_v|x_v - \hat{x}_v|}{\sum_v w_v}, \quad \mathcal{L}_{\mathrm{bg}} = \frac{1}{|\Omega_{\mathrm{bg}}|}\sum_{v\in\Omega_{\mathrm{bg}}} |\hat{x}_v|, \\ \mathcal{L}_{\mathrm{wav}} &= \|H_{\mathrm{haar}}(x) - H_{\mathrm{haar}}(\hat{x})\|_1, \quad \mathcal{L}_{\mathrm{KL}} = \frac{1}{2}\mathbb{E}\left[\mu^2 + \sigma^2 - \log\sigma^2 - 1\right]. \end{aligned} \tag{2}$$

Here, w_v upweights foreground voxels, Ω_{bg} denotes the background region, H_{haar} is applied frame-wise to enforce 3D wavelet-domain consistency, and $\mathcal{L}_{\mathrm{KL}}$ is the Kullback–Leibler (KL) divergence term.

3.3 Stage II: Conditional Latent Diffusion Modeling

After VAE pretraining, BrainWorld learns a conditional latent diffusion model $p_\theta(u_{i+1} \mid u_i, c_i)$, where u_i and u_{i+1} are the current and next fMRI-window latents. The condition c_i includes past functional connectivity (FC), sMRI, and optional video/audio embeddings as dynamic, anatomical, and stimulus contexts. Here, "structural prior" denotes sMRI-based conditioning rather than an explicit Bayesian prior. All conditions are extracted by frozen encoders and projected into the DiT conditioning space, with details in Supplementary Section C.3.

We formulate next-window prediction as denoising the future latent. The clean target is $u_0 := u_{i+1}$, and the forward diffusion process is

$$u_t = \sqrt{\bar{\alpha}_t}u_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{3}$$

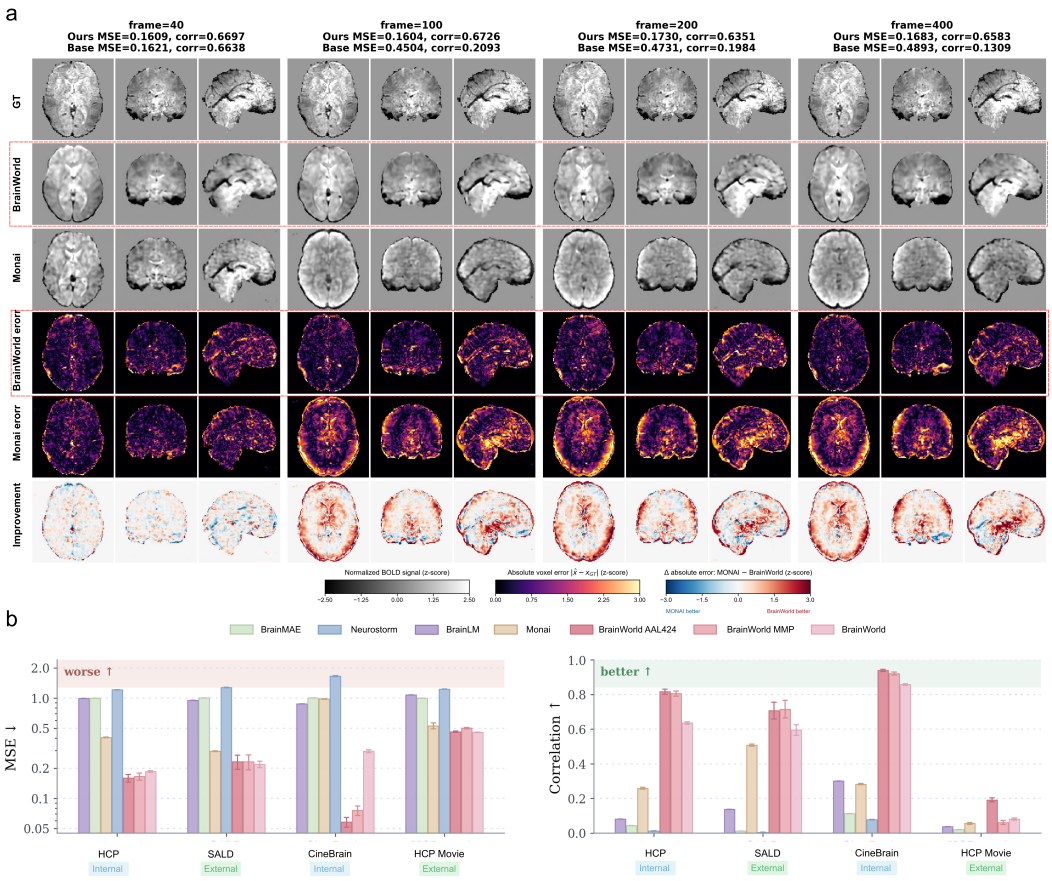

Figure 2: Long-horizon 4D fMRI generation and cross-dataset evaluation. (a) Representative HCP subject visualization at 40, 100, 200, and 400 predicted frames, comparing ground truth (GT), BrainWorld, MONAI, error maps, and voxel-wise improvements. (b) Quantitative results across internal/external datasets and resting-state/visual-stimulus fMRI. BrainWorld-MMP and BrainWorld-AAL424 denote BrainWorld predictions parcellated using the MMP [19] and AAL424 [38] atlases, respectively. Additional visualizations and generation metrics across different prediction horizons are provided in Supplementary Sections B.1 and B.2.

where $t \in \{1, \ldots, T_{\text{diff}}\}$. We use $T_{\text{diff}} = 1000$ diffusion steps with a cosine noise schedule. The noisy latent u_t is spatially patchified into spatiotemporal tokens while preserving the temporal axis, enabling the DiT to model both spatial dependencies and temporal evolution.

Conditions are injected through complementary token-level and global pathways. At the token level, condition embeddings are incorporated by cross-attention to provide content-dependent modulation of latent tokens. At the global level, condition embeddings are fused with the diffusion timestep embedding and used to modulate each DiT block through adaptive normalization and residual gating. This allows sMRI, functional context, and stimulus information to modulate denoising throughout the network, rather than being fused only at the input or output.

We train the DiT with v-prediction:

$$v = \sqrt{\bar{\alpha}_t}\epsilon - \sqrt{1-\bar{\alpha}_t}u_0, \quad \mathcal{L}_v = \|\hat{v}_\theta(u_t, t, u_i, c_i) - v\|_2^2 . \tag{4}$$

To stabilize clean-latent recovery, we reconstruct $\hat{u}_0$ from the predicted v and add an auxiliary loss:

$$\mathcal{L}_{\text{StageII}} = \mathcal{L}_v + \lambda_{x_0}\text{SmoothL1}(\hat{u}_0, u_0), \quad \lambda_{x_0} = 0.05. \tag{5}$$

At inference, the predicted latent is decoded by the frozen VAE decoder and autoregressively rolled out for long-horizon generation; intermediate DiT states are used for downstream representation extraction.

3.4 Autoregressive Latent Rollout and Generation

To generate trajectories longer than one 40-frame block, we autoregressively unroll the one-step latent predictor learned in Stage II. Given the current latent window u_i and condition set c_i, BrainWorld samples the next latent block

$$\hat{u}_{i+1} \sim p_\theta(u_{i+1} \mid u_i, c_i),$$

and reconstructs it into voxel-level 4D fMRI using the frozen VAE decoder:

$$\hat{x}_{i+1} = D_\psi(\hat{u}_{i+1}).$$

During rollout, different conditions follow different update rules. Specifically, each generated voxel-level block is parcellated into Schaefer-100 ROI time series [45], from which an ROI-wise Pearson FC matrix is computed, vectorized, projected into an 800-D embedding, and used as the functional context for the next rollout step. sMRI serves as a subject-level anatomical prior and remains fixed across the rollout. Optional video/audio conditions are externally aligned to the corresponding future time intervals when available. For a target horizon of H frames and block length $B = 40$, we perform

$$K = \left\lceil \frac{H}{B} \right\rceil$$

rollout steps and concatenate the generated blocks in temporal order. This enables efficient long-horizon generation by performing denoising in latent space while updating functional context from generated 4D fMRI.

3.5 Modality-Fused Representation Extraction for Downstream Tasks

Beyond generation, we use the trained DiT as a diffusion-based representation encoder. For each downstream sample, the observed fMRI window x_i is encoded by the pretrained VAE into a latent window u_i, which is then fed into the DiT to extract intermediate features paired with the label of the same window. No future window or generated sample is used for downstream supervision or feature extraction.

For each selected diffusion timestep t, the current latent window u_i is perturbed using the Stage-II noise schedule and fed into the DiT. Hidden states from the selected timesteps and layers are then collected as multi-scale representations. Let $H_{t,l} \in \mathbb{R}^{N \times D}$ denote the hidden tokens at timestep t and layer l. We mean-pool over tokens and concatenate features across selected timesteps and layers:

$$e_{t,l} = \frac{1}{N} \sum_{n=1}^{N} H_{t,l,n}, \quad e = \mathrm{Concat}_{t \in \mathcal{T}_{\mathrm{feat}},\, l \in \mathcal{L}_{\mathrm{feat}}}(e_{t,l}).$$

We use a fixed set of diffusion timesteps and late DiT layers for feature extraction, with detailed settings provided in Supplementary Table S7.

To test how sMRI should be integrated, we compare three variants: BrainWorld$_{\mathrm{F}}$, which extracts features from fMRI latents only; BrainWorld$_{\mathrm{F+S\text{-}concat}}$, which concatenates fMRI features and independently extracted sMRI embeddings before the downstream head; and BrainWorld$_{\mathrm{F+S\text{-}prior}}$, which injects sMRI into the DiT through the structural-prior conditioning pathway. These variants distinguish post-hoc representation fusion from structural-prior fusion within the DiT.

4 Experiments

4.1 Experimental Setup

Datasets and Preprocessing. As shown in Figure 1, we use 22 public datasets spanning resting-state, task-driven, naturalistic movie-based, and clinical cohorts [15, 21, 34, 52, 27, 31, 16, 48, 18, 70, 44, 32, 60, 24, 11, 39, 9, 1, 6]. The datasets are divided into pretraining-only, internal evaluation, and external evaluation splits to assess generalization under distribution shifts. All fMRI scans are spatially aligned to a common template, resampled to 2 mm isotropic resolution, and temporally interpolated to a unified TR of 0.72 s. Each complete fMRI time series undergoes background-aware global Z-scoring before being segmented into 40-frame windows. Each volume is cropped or padded to a fixed spatial size of $96 \times 96 \times 96$. For MRI, images are skull-stripped, bias-field corrected, and resampled to 1 mm MNI152 space, serving as anatomical conditioning inputs. Details of the datasets and preprocessing pipelines are provided in Supplementary Sections C.2 and C.1, respectively.

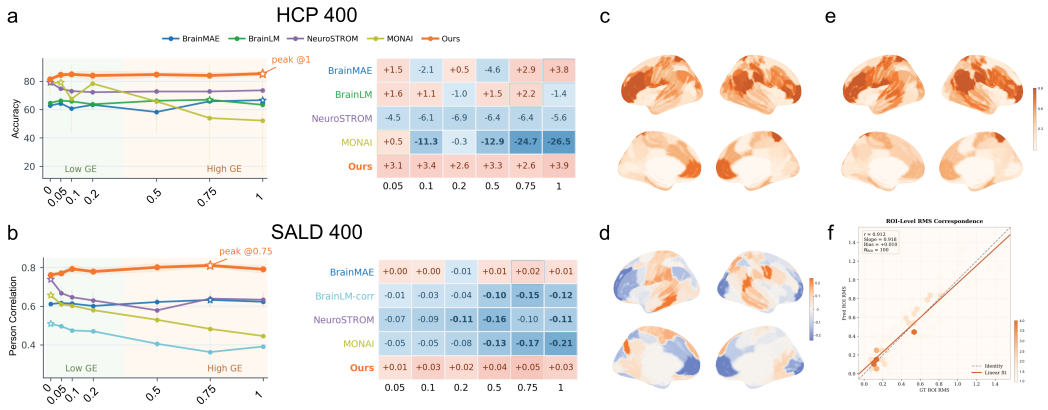

Figure 3: Generated-example (GE) scaling improves downstream performance while preserving ROI-level response structure at a 400-frame horizon. (a) HCP-400 gender classification results across GE ratios, where higher ratios add more generated examples. The line plot shows absolute accuracy, and the heatmap shows gains over each model's GE $= 0$ baseline in percentage points. (b) SALD-400 age regression results in the same format, with gains reported in Pearson correlation units. (c,e) Cohort-level ground-truth and predicted ROI root-mean-square (RMS) surface maps on CineBrain. (d) Prediction-minus-ground-truth ROI RMS difference maps. (f) ROI-wise correspondence between predicted and ground-truth RMS values across Schaefer-100 parcels [45].

Baselines and Evaluation Protocols. We evaluate BrainWorld on two task families: generative forecasting and downstream transfer. For generative forecasting, we compare against representative fMRI foundation models, including BrainMAE (an ROI-level predictive baseline implemented for controlled comparison; see Supplementary Section A.1), BrainLM [8], and NeuroSTORM [54]. We also include a MONAI-based latent diffusion baseline [42] adapted for 4D fMRI prediction (see Supplementary Section C.4). For fair quantitative comparison, voxel-level predictions from BrainWorld and voxel baselines are evaluated in voxel space or parcellated to the corresponding ROI space when compared with ROI-level baselines. For downstream transfer, we compare functional-only (F), structural-only (S), and multimodal (F+S) baselines under two protocols: full fine-tuning for end-to-end adaptability and linear probing for frozen-backbone representation transfer. Detailed baseline descriptions, computational resources, and evaluation protocols are provided in Supplementary Sections C.4, C.6, and C.7, respectively.

4.2 BrainWorld Enables Long-Horizon Predictive Generation of 4D fMRI

We evaluate BrainWorld as a long-horizon 4D fMRI predictive generator across resting-state and stimulus-driven conditions. Figure 2(a) visualizes a representative HCP subject at prediction horizons of 40, 100, 200, and 400 frames, comparing ground truth, BrainWorld, the MONAI baseline, error maps, and voxel-wise improvement maps. BrainWorld preserves clearer whole-brain structure and lower voxel-level error across horizons, with the advantage remaining evident at 200 and 400 frames. In contrast, MONAI progressively loses anatomical detail and accumulates larger errors as the prediction horizon increases.

Figure 2(b) reports 400-frame quantitative results across four settings: internal resting-state HCP, external resting-state SALD, internal visual-stimulus CineBrain, and external visual-stimulus HCP Movie. BrainWorld consistently obtains lower mean squared error (MSE) and higher Pearson correlation than competing baselines across these settings, supporting robust long-horizon predictive generation across both brain states and dataset sources. Its strong performance on SALD and HCP Movie further suggests generalization to unseen cohorts and acquisition paradigms.

4.3 Generated fMRI Supports Downstream Augmentation

To evaluate whether generated data improves downstream performance, we conduct generated-example (GE) augmentation with varying GE ratios while keeping the real training set fixed. Generated examples inherit labels and sMRI conditions from their conditioning real samples, and are used as label-preserving augmentations rather than independent biological measurements. All GE ratios use the same optimization steps and validation-based early stopping, and are evaluated on held-out real test sets. As shown in Fig. 3(a,b), HCP-400 refers to gender classification on HCP using 400-frame samples, while SALD-400 refers to age regression on SALD. $\mathrm{GE} = 0$ uses only real data, while $\mathrm{GE} = 1.0$ adds an equal number of generated and real samples. BrainWorld achieves the best performance across all GE ratios on both tasks. On HCP-400, accuracy improves from 81.45% to 85.39% at $\mathrm{GE} = 1.0$ ($+3.94$ points), and on SALD-400, Pearson correlation increases from 0.7613 to 0.8104 at $\mathrm{GE} = 0.75$ ($+0.0491$). In contrast, the strongest baseline improves by only $+0.0210$, with most baselines degrading as GE increases.

To assess spatial fidelity, we evaluate Schaefer-100 ROI temporal RMS on 38 unseen CineBrain runs, as shown in Fig. 3(c–f). The predicted RMS closely matches ground truth ($r = 0.912$, cosine similarity $= 0.973$, slope $= 0.918$, bias $= +0.010$), with localized errors. This suggests that BrainWorld preserves ROI-wise functional response patterns in generated samples, which may explain their utility for downstream augmentation.

4.4 Structural-Prior Conditioning Improves Multimodal Representation Transfer

We explore whether sMRI is more effective when injected as a structural conditioning signal than when used as a parallel feature for multimodal fusion. To this end, we compare BrainWorld$_{\mathrm{F+S\text{-}prior}}$ with functional-only (F), structural-only (S), and multimodal fusion baselines, including Brain Harmony as a representative representation-level fusion method. To further isolate the fusion mechanism from backbone differences, we introduce a same-backbone concat baseline, BrainWorld$_{\mathrm{F+S\text{-}concat}}$, which independently extracts fMRI and sMRI features and concatenates them before the downstream head. BrainWorld$_{\mathrm{F}}$ denotes the corresponding fMRI-only variant without sMRI conditioning. Two transfer protocols are used: full fine-tuning, which evaluates end-to-end downstream adaptability, and linear probing with a frozen backbone, which evaluates representation transferability. For age regression, target ages are z-score normalized within the training set. Given the limited number of runs ($n = 3$), Cohen's d [10] is used to quantify the effect size of performance differences.

Table 1 shows that BrainWorld$_{\mathrm{F+S\text{-}prior}}$ achieves the best full fine-tuning performance across five datasets, demonstrating strong adaptability to diverse demographic tasks. This advantage also holds under the frozen-backbone protocol in Table 2, indicating that denoising-time structural conditioning improves not only task-specific adaptation but also the transferability of learned representations.

To further examine the effect of predictive pretraining, we evaluate frozen DiT representations at different training checkpoints. Downstream performance improves consistently as pretraining progresses, suggesting increasingly transferable representations (Supplementary Table S4).

Notably, BrainWorld$_{\mathrm{F+S\text{-}concat}}$ does not consistently improve over BrainWorld$_{\mathrm{F}}$, suggesting that sMRI is not automatically beneficial when used as a post-hoc feature. In contrast, BrainWorld$_{\mathrm{F+S\text{-}prior}}$ achieves the strongest overall performance, supporting the benefit of injecting sMRI into the DiT denoising pathway rather than simply concatenating it as a post-hoc representation.

5 Discussion

Our results support formulating whole-brain 4D fMRI generation as structural-condition-aware brain dynamics modeling. BrainWorld achieves stable long-horizon predictive generation across datasets and brain states, and the generated samples further improve downstream performance in limited-data regimes. Beyond generation quality, the strong downstream performance of intermediate DiT representations indicates that the generative backbone learns transferable neural features, rather than merely fitting realistic signal trajectories. The advantage of BrainWorld$_{\mathrm{F+S\text{-}prior}}$ over BrainWorld$_{\mathrm{F}}$ further suggests that sMRI conditioning benefits both predictive generation and representation learning. A key implication is that structure-function fusion can be more effective when embedded into the generative process rather than applied as post-hoc feature fusion. Unlike multimodal models such as Brain Harmony that compress sMRI and fMRI into shared representations, BrainWorld

Table 1: Full fine-tuning performance on demographic tasks. We report accuracy (ACC) and F1 score for classification, and MSE and Pearson correlation for regression. F, S, and F+S denote functional-only, structural-only, and multimodal models, respectively. Red indicates the best performance, and underlining indicates the second-best performance. * denotes a large favorable effect size with Cohen's $d \geq 0.8$ compared with the best-performing non-BrainWorld baseline for the same task and metric.

Model \ Dataset	Modality	ABIDE *Age Regression*		NKI *Age Regression*		SALD *Age Regression*		ABCD *Gender Classif.*		BHRC *Gender Classif.*	
		MSE ↓	R ↑	MSE ↓	R ↑	MSE ↓	R ↑	ACC ↑	F1 ↑	ACC ↑	F1 ↑
BrainLM [8]	F	0.907±0.007	0.239±0.023	0.503±0.015	0.678±0.014	0.677±0.055	0.634±0.061	59.24±1.03	58.74±0.70	57.09±1.33	54.69±1.03
Brain-JEPA [12]	F	0.982±0.066	0.166±0.017	1.030±0.080	0.330±0.030	1.130±0.108	0.299±0.060	-	-	58.16±6.10	56.16±6.05
BrainMass [67]	F	0.695±0.038	0.500±0.044	0.600±0.062	0.620±0.036	0.704±0.080	0.626±0.062	57.92±1.19	57.86±1.18	62.76±1.50	61.05±1.22
Brain-DiT [63]	F	0.750±0.010	0.610±0.010	0.380±0.010	0.810±0.010	0.330±0.010	0.790±0.010	61.67±1.15	60.01±3.40	56.60±2.35	48.86±5.85
SwiFT [26]	F	0.524±0.010	0.704±0.011	0.127±0.004	0.946±0.002	0.164±0.080	0.926±0.003	66.67±3.51	68.69±0.44	70.21±0.02	69.56±0.02
SLIMBrain [56]	F	0.493±0.012	0.285±0.015	0.667±0.083	0.613±0.078	0.705±0.054	0.590±0.023	67.92±1.10	67.18±1.76	71.74±2.17	69.35±4.25
NeuroSTORM [54]	F	0.537±0.016	0.648±0.013	0.152±0.033	0.835±0.035	0.194±0.019	0.892±0.010	76.51±1.88	76.24±1.70	73.02±2.19	72.49±1.96
Omni-fMRI [58]	F	0.427±0.006	0.734±0.002	0.088±0.004	0.959±0.003	0.182±0.019	0.909±0.005	77.01±0.55	76.91±0.54	75.86±1.35	75.96±1.31
Brain Harmony-S [13]	S	0.318±0.016	0.873±0.008	0.266±0.012	0.868±0.005	0.553±0.054	0.632±0.038	74.33±0.94	73.96±0.75	65.96±3.13	65.76±3.22
Brain Harmony [13]	F + S	0.373±0.028	0.753±0.023	0.171±0.016	0.822±0.018	0.188±0.017	0.667±0.013	75.17±2.08	75.16±2.09	69.50±1.23	68.74±1.17
$\text{BrainWorld}_{\text{F}}$	F	0.127±0.058	0.915±0.026	0.163±0.036	0.949±0.007	0.143±0.015	0.912±0.009	79.43±0.45	79.42±0.46	78.37±2.92	78.09±3.17
$\text{BrainWorld}_{\text{F+S-prior}}$	F + S	**0.062±0.011***	**0.948±0.008***	**0.083±0.001***	**0.962±0.001***	**0.107±0.007***	**0.935±0.005***	**80.75±1.71***	**80.75±1.71***	**79.94±2.07***	**79.84±2.03***

Table 2: Performance on demographic and disease diagnosis tasks using a frozen backbone. We report accuracy (ACC) and F1 score for classification, and MSE and Pearson correlation for regression. MCI denotes mild cognitive impairment. F, S, and F+S denote functional-only, structural-only, and multimodal models, respectively. Red indicates the best performance and underline indicates the second-best performance. * denotes a large favorable effect size with Cohen's $d \geq 0.8$ compared with the best-performing non-BrainWorld baseline for the same task and metric.

Model \ Dataset	Modality	NKI *Education Classif.*		ABCD *Gender Classif.*		SALD *Age Regression*		ADNI (MCI) *Diagnosis*	
		ACC ↑	F1 ↑	ACC ↑	F1 ↑	MSE ↓	R ↑	ACC ↑	F1 ↑
Brain-DiT [63]	F	62.80±0.63	60.26±1.22	55.47±0.03	52.37±0.39	0.666±0.005	0.495±0.007	47.03±0.49	46.94±0.54
SwiFT [26]	F	61.55±2.52	58.23±4.20	59.31±2.09	58.16±2.84	0.933±0.042	0.460±0.131	69.78±4.63	69.31±5.23
SLIMBrain [56]	F	59.94±1.43	59.10±1.79	67.92±1.10	67.18±1.76	0.680±0.100	0.590±0.093	59.14±7.00	57.98±6.85
NeuroSTORM [54]	F	66.78±0.51	57.37±0.86	63.06±6.49	60.76±10.16	0.386±0.035	0.774±0.004	57.25±3.18	56.76±3.21
Omni-fMRI [58]	F	68.13±0.60	63.65±2.04	73.96±1.80	73.74±2.18	0.354±0.048	0.814±0.024	63.40±2.55	56.67±9.60
BrainIAC [51]	S	63.93±2.33	62.97±2.36	65.76±1.61	65.69±1.57	0.416±0.015	0.750±0.003	73.73±0.69	73.56±0.66
Brain Harmony-S [13]	S	82.10±5.72	82.09±5.19	69.87±1.07	67.09±0.80	0.343±0.025	0.803±0.018	52.27±1.86	52.15±1.86
Brain Harmony [13]	F + S	81.48±0.01	80.66±0.85	70.13±3.44	69.42±3.21	0.386±0.034	0.802±0.025	58.33±3.86	58.28±4.01
$\text{BrainWorld}_{\text{F}}$	F	86.76±0.20	86.27±0.17	69.81±0.07	69.80±0.08	0.380±0.020	0.792±0.008	67.82±0.59	67.77±0.53
$\text{BrainWorld}_{\text{F+S-concat}}$	F + S	81.62±1.96	81.53±1.84	67.31±0.62	67.25±0.64	0.442±0.040	0.754±0.007	67.36±2.99	67.31±3.13
$\text{BrainWorld}_{\text{F+S-prior}}$	F + S	**87.67±0.68***	**87.21±0.80***	**74.09±0.24**	**74.10±0.24**	**0.269±0.007***	**0.830±0.005***	**76.27±6.14**	**76.26±6.12**

assigns sMRI a distinct role as subject-level anatomical context that modulates functional dynamics generation. Moreover, $\text{BrainWorld}_{\text{F+S-prior}}$ outperforms the same-backbone $\text{BrainWorld}_{\text{F+S-concat}}$ baseline, suggesting that the gain does not simply come from adding sMRI features, but from injecting anatomical information through the DiT structural-prior pathway. This design reflects a useful division of labor: the VAE preserves voxel-level spatial organization in a compact latent space, the DiT models temporal evolution, FC embeddings provide dynamic functional context during generation, and sMRI supplies stable subject-level anatomical context.

Limitations. BrainWorld currently relies on next-window prediction with diffusion-based autoregressive rollout, which may accumulate errors and increase inference latency over very long horizons. Although the framework supports multiple conditions, our experiments evaluate condition-aware prediction rather than explicit counterfactual control; fine-grained task- and stimulus-level control also depends on the quality and temporal alignment of external embeddings. Moreover, generated examples are used as label-preserving augmentations rather than independent biological measurements, and the benefit of structural conditioning should be interpreted as computational evidence for useful structure-function coupling rather than biological causality. Future work should explore hierarchical temporal rollout, accelerated samplers or distillation, richer stimulus and task conditioning, uncertainty-aware generation, and more mechanistic validation of structure-function relationships.

Acknowledgment

This work was supported by National Natural Science Foundation of China (3254100307, 62472206), Brain Science and Brain-like Intelligence Technology-National Science and Technology Major Project (2021ZD0200500, 2026ZD0220402), National Key R&D Program of China (2025YFC3410000), Shenzhen Science and Technology Innovation Committee (RCYX20231211090405003), 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] Lindsay M Alexander, Jasmine Escalera, Lei Ai, Charissa Andreotti, Karina Febre, Alexander Mangone, Natan Vega-Potler, Nicolas Langer, Alexis Alexander, Meagan Kovacs, et al. An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific data*, 4(1):170181, 2017.

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

[3] Amir H Azizi, Laurenz Wiskott, and Sen Cheng. A computational model for preplay in the hippocampus. *Frontiers in computational neuroscience*, 7:161, 2013.

[4] Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*, 2023.

[5] Yuda Bi, Anees Abrol, Zening Fu, and Vince D Calhoun. A multimodal vision transformer for interpretable fusion of functional and structural neuroimaging data. *Human brain mapping*, 45 (17):e26783, 2024.

[6] Bharat B Biswal, Maarten Mennes, Xi-Nian Zuo, Suril Gohel, Clare Kelly, Steve M Smith, Christian F Beckmann, Jonathan S Adelstein, Randy L Buckner, Stan Colcombe, et al. Toward discovery science of human brain function. *Proceedings of the national academy of sciences*, 107(10):4734–4739, 2010.

[7] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

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

[9] 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 (abcd) study: imaging acquisition across 21 sites. *Developmental cognitive neuroscience*, 32:43–54, 2018.

[10] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013.

[11] 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.

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

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

[14] Peiyu Duan, Jiyao Wang, Nicha C Dvornek, Junlin Yang, Ziqi Gao, Lawrence H Staib, and James S Duncan. FM-fMRI: Event conditioned flow matching for rest-to-task fMRI time-series synthesis. *arXiv preprint arXiv:2605.26423*, 2026.

[15] Jianxiong Gao, Yichang Liu, Baofeng Yang, Jianfeng Feng, and Yanwei Fu. Cinebrain: A large-scale multi-modal brain dataset during naturalistic audiovisual narrative processing. *arXiv preprint arXiv:2503.06940*, 2025.

[16] Peng Gao, Hao-Ming Dong, Si-Man Liu, Xue-Ru Fan, Chao Jiang, Yin-Shan Wang, Daniel Margulies, Hai-Fang Li, and Xi-Nian Zuo. A chinese multi-modal neuroimaging data release for increasing diversity of human brain mapping. *Scientific Data*, 9(1):286, 2022.

[17] Zebin Gao, Guoxun Zhang, Hengrui Liang, Jiaxin Liu, Liangdi Ma, Tianyun Wang, Yanchen Guo, YuJia Chen, Zeping Yan, Xiangru Chen, et al. A lung ct vision foundation model facilitating disease diagnosis and medical imaging. *Nature Communications*, 2025.

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

[19] Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

[20] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018.

[21] Michael Hanke, Nico Adelhöfer, Daniel Kottke, Vittorio Iacovella, Ayan Sengupta, Falko R Kaule, Roland Nigbur, Alexander Q Waite, Florian J Baumgartner, and Jörg Stadler. A studyforrest extension, simultaneous fMRI and eye gaze recordings during prolonged natural stimulation. *Scientific Data*, 3:160092, 2016. doi: 10.1038/sdata.2016.92.

[22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[23] Yufan Hu, Wuyang Li, Yixuan Yuan, et al. Synthesizing realistic fmri: a physiological dynamics-driven hierarchical diffusion model for efficient fmri acquisition. In *The Thirteenth International Conference on Learning Representations*, 2025.

[24] 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.

[25] Moein Khajehnejad, Forough Habibollahi, and Adeel Razi. Brainsymphony: A transformer-driven fusion of fmri time series and structural connectivity. *arXiv preprint arXiv:2506.18314*, 2025.

[26] Peter Kim, Junbeom Kwon, Sunghwan Joo, Sangyoon Bae, Donggyu Lee, Yoonho Jung, Shinjae Yoo, Jiook Cha, and Taesup Moon. SwiFT: Swin 4d fMRI transformer. *Advances in Neural Information Processing Systems*, 36:42015–42037, 2023.

[27] Dorit Kliemann, Ralph Adolphs, Tim Armstrong, Paola Galdi, David A Kahn, Tessa Rusch, A Zeynep Enkavi, Deuhua Liang, Steven Lograsso, Wenying Zhu, et al. Caltech conte center, a multimodal data resource for exploring social cognition and decision-making. *Scientific Data*, 9 (1):138, 2022.

[28] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[29] Tian Lan, Yiming Zheng, and Jianxin Yin. Diffusion-based cross-modal feature extraction for multi-label classification. *arXiv preprint arXiv:2509.15553*, 2025.

[30] Zongjian Li, Bin Lin, Yang Ye, Liuhan Chen, Xinhua Cheng, Shenghai Yuan, and Li Yuan. Wf-vae: Enhancing video vae by wavelet-driven energy flow for latent video diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17778–17788, 2025.

[31] Wei Liu, Dongtao Wei, Qunlin Chen, Wenjing Yang, Jie Meng, Guorong Wu, Taiyong Bi, Qinglin Zhang, Xi-Nian Zuo, and Jiang Qiu. Longitudinal test-retest neuroimaging data from healthy young adults in southwest china. *Scientific data*, 4(1):170017, 2017.

[32] 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.

[33] Michael P. Milham, Damien Fair, Maarten Mennes, and Stewart H. Mostofsky. The adhd-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in Systems Neuroscience*, Volume 6 - 2012, 2012. ISSN 1662-5137.

[34] Elenor Morgenroth, Stefano Moia, Laura Vilaclara, Raphael Fournier, Michal Muszynski, Maria Ploumitsakou, Marina Almató-Bellavista, Patrik Vuilleumier, and Dimitri Van De Ville. Emo-film: a multimodal dataset for affective neuroscience using naturalistic stimuli. *Scientific Data*, 12(1):684, 2025.

[35] Soumik Mukhopadhyay, Matthew Gwilliam, Yosuke Yamaguchi, Vatsal Agarwal, Namitha Padmanabhan, Archana Swaminathan, Tianyi Zhou, Jun Ohya, and Abhinav Shrivastava. Do text-free diffusion models learn discriminative visual representations? In *European Conference on Computer Vision*, pages 253–272. Springer, 2024.

[36] Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mahmoud Assran, Koustuv Sinha, Michael Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

[37] Tomoya Nakai and Shinji Nishimoto. Quantitative models reveal the organization of diverse cognitive functions in the brain. *Nature communications*, 11(1):1142, 2020.

[38] Samaneh Nemati, Teddy J Akiki, Jeremy Roscoe, Yumeng Ju, Christopher L Averill, Samar Fouda, Arpan Dutta, Shane McKie, John H Krystal, JF William Deakin, et al. A unique brain connectome fingerprint predates and predicts response to antidepressants. *IScience*, 23(1), 2020.

[39] Kate Brody Nooner, Stanley J. Colcombe, Russell H. Tobe, Maarten Mennes, Melissa M. Benedict, Alexis L. Moreno, Laura J. Panek, Shaquanna Brown, Stephen T. Zavitz, Qingyang Li, Sharad Sikka, David Gutman, Saroja Bangaru, Rochelle Tziona Schlachter, Stephanie M. Kamiel, Ayesha R. Anwar, Caitlin M. Hinz, Michelle S. Kaplan, Anna B. Rachlin, Samantha Adelsberg, Brian Cheung, Ranjit Khanuja, Chaogan Yan, Cameron C. Craddock, Vincent Calhoun, William Courtney, Margaret King, Dylan Wood, Christine L. Cox, A. M. Clare Kelly, Adriana Di Martino, Eva Petkova, Philip T. Reiss, Nancy Duan, Dawn Thomsen, Bharat Biswal, Barbara Coffey, Matthew J. Hoptman, Daniel C. Javitt, Nunzio Pomara, John J. Sidtis, Harold S. Koplewicz, Francisco Xavier Castellanos, Bennett L. Leventhal, and Michael P. Milham. The NKI-Rockland sample: A model for accelerating the pace of discovery science in psychiatry. *Frontiers in Neuroscience*, 6:152, 2012. doi: 10.3389/fnins.2012.00152.

[40] Anton Orlichenko, Gang Qu, Ziyu Zhou, Anqi Liu, Hong-Wen Deng, Zhengming Ding, Julia M Stephen, Tony W Wilson, Vince D Calhoun, and Yu-Ping Wang. A demographic-conditioned variational autoencoder for fmri distribution sampling and removal of confounds. *ArXiv*, pages arXiv–2405, 2024.

[41] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[42] Walter HL Pinaya, Mark S Graham, Eric Kerfoot, Petru-Daniel Tudosiu, Jessica Dafflon, Virginia Fernandez, Pedro Sanchez, Julia Wolleb, Pedro F Da Costa, Ashay Patel, et al. Generative ai for medical imaging: extending the monai framework. *arXiv preprint arXiv:2307.15208*, 2023.

[43] Youzhi Qu, Junfeng Xia, Xinyao Jian, Wendu Li, Kaining Peng, Zhichao Liang, Haiyan Wu, and Quanying Liu. Uncovering cognitive taskonomy through transfer learning in masked autoencoder-based fmri reconstruction. In *International Workshop on Human Brain and Artificial Intelligence*, pages 35–50. Springer, 2024.

[44] Giovanni Abrahão Salum, Ary Gadelha, Pedro Mario Pan, Tais Silveira Moriyama, Ana Soledade Graeff-Martins, Ana Carina Tamanaha, Pedro Alvarenga, Fernanda Valle Krieger, Bacy Fleitlich-Bilyk, Andrea Jackowski, et al. High risk cohort study for psychiatric disorders in childhood: rationale, design, methods and preliminary results. *International Journal of Methods in Psychiatric Research*, 24(1):58–73, 2015. doi: 10.1002/mpr.1459.

[45] 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.

[46] Jungwoo Seo, David Keetae Park, Shinjae Yoo, and Jiook Cha. Scalable diffusion transformer for conditional 4d fmri synthesis. *arXiv preprint arXiv:2511.22870*, 2025.

[47] Golia Shafiei, Nathalia B Esper, Mauricio S Hoffmann, Lei Ai, Andrew A Chen, Jon Cluce, Sydney Covitz, Steven Giavasis, Connor Lane, Kahini Mehta, et al. Reproducible brain charts: An open data resource for mapping brain development and its associations with mental health. *Neuron*, 113(22):3758–3779, 2025.

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

[49] Yifei Sun, Daniel Chahine, Qinghao Wen, Tianming Liu, Xiang Li, Yixuan Yuan, Fernando Calamante, and Jinglei Lv. Voxel-level brain states prediction using swin transformer. *IEEE Journal of Biomedical and Health Informatics*, 29(12):8719–8726, 2025.

[50] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[51] Divyanshu Tak, Biniam A Garomsa, Tafadzwa L Chaunzwa, Anna Zapaishchykova, Juan Carlos Climent Pardo, Zezhong Ye, John Zielke, Yashwanth Ravipati, Sri Vajapeyam, Maryam Mahootiha, et al. A foundation model for generalized brain mri analysis. *MedRxiv*, 2024.

[52] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The WU-Minn Human Connectome Project: An Overview. *Neuroimage*, 80:62–79, 2013.

[53] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 3(4):6, 2025.

[54] Cheng Wang, Yu Jiang, Zhihao Peng, Chenxin Li, Changbae Bang, Lin Zhao, Jinglei Lv, Jorge Sepulcre, Carl Yang, Lifang He, et al. Towards a general-purpose foundation model for fmri analysis. *arXiv preprint arXiv:2506.11167*, 2025.

[55] Jiyao Wang, Nicha C Dvornek, Lawrence H Staib, and James S Duncan. Learning sequential information in task-based fMRI for synthetic data augmentation. In *International Workshop on Machine Learning in Clinical Neuroimaging*, pages 79–88. Springer, 2023.

[56] Mo Wang, Junfeng Xia, Wenhao Ye, Enyu Liu, Kaining Peng, Jianfeng Feng, Quanying Liu, and Hongkai Wen. SLIM-Brain: A Data-and Training-Efficient Foundation Model for fMRI Data Analysis. *arXiv preprint arXiv:2512.21881*, 2025.

[57] Mo Wang, Wenhao Ye, Junfeng Xia, Minghao Xu, Hongkai Wen, and Quanying Liu. Flexibrain: Resolution-agnostic voxel-level encoding for native fmri. *arXiv preprint arXiv:2606.11500*, 2026.

[58] Mo Wang, Wenhao Ye, Junfeng Xia, Junxiang Zhang, Xuanye Pan, Minghao Xu, Haotian Deng, Hongkai Wen, and Quanying Liu. Omni-fmri: A universal atlas-free fmri foundation model. *arXiv preprint arXiv:2601.23090*, 2026.

[59] Shaocong Wang, Tong Liu, Yihan Li, Ming Li, Kairui Wen, Pei Yang, Wenqi Ji, Minjing Yu, and Yong-Jin Liu. Eegdm: Learning eeg representation with latent diffusion model. *arXiv preprint arXiv:2508.20705*, 2025.

[60] Dongtao Wei, Kaixiang Zhuang, Lei Ai, Qunlin Chen, Wenjing Yang, Wei Liu, Kangcheng Wang, Jiangzhou Sun, and Jiang Qiu. Structural and functional brain scans from the cross-sectional Southwest University adult lifespan dataset. *Scientific Data*, 5(1):180134, July 2018. ISSN 2052-4463.

[61] Yuxiang Wei, Yanteng Zhang, Xi Xiao, Chengxuan Qian, Tianyang Wang, and Vince D Calhoun. fmri-lm: Towards a universal foundation model for language-aligned fmri understanding. *arXiv preprint arXiv:2511.21760*, 2025.

[62] Yuxiang Wei, Yanteng Zhang, Xi Xiao, Tianyang Wang, Xiao Wang, and Vince D Calhoun. 4d multimodal co-attention fusion network with latent contrastive alignment for alzheimer's diagnosis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5457–5466, 2026.

[63] Junfeng Xia, Wenhao Ye, Xuanye Pan, Xinke Shen, Mo Wang, and Quanying Liu. Brain-dit: A universal multi-state fmri foundation model with metadata-conditioned pretraining. *arXiv preprint arXiv:2604.12683*, 2026.

[64] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022.

[65] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14138–14148, 2021.

[66] Xingyi Yang and Xinchao Wang. Diffusion model as representation learner. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18938–18949, 2023.

[67] Yanwu Yang, Chenfei Ye, Guinan Su, Ziyao Zhang, Zhikai Chang, Hairui Chen, Piu Chan, Yue Yu, and Ting Ma. Brainmass: Advancing brain network analysis for diagnosis with large-scale self-supervised learning. *IEEE transactions on medical imaging*, 43(11):4004–4016, 2024.

[68] B. T. Thomas Yeo, Fenna M Krienen, Jorge Sepulcre, Mert R Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L Roffman, Jordan W Smoller, Lilla Z"ollei, Jonathan R Polimeni, Bruce Fischl, Hesheng Liu, and Randy L Buckner. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3):1125–1165, 2011. doi: 10.1152/jn.00338.2011.

[69] Lu Zhang, Li Wang, Jean Gao, Shannon L Risacher, Jingwen Yan, Gang Li, Tianming Liu, Dajiang Zhu, Alzheimer's Disease Neuroimaging Initiative, et al. Deep fusion of brain structure-function in mild cognitive impairment. *Medical image analysis*, 72:102082, 2021.

[70] Xi-Nian Zuo, Jeffrey S Anderson, Pierre Bellec, Rasmus M Birn, Bharat B Biswal, Janusch Blautzik, John Breitner, Randy L Buckner, Vince D Calhoun, F Xavier Castellanos, et al. An open science resource for establishing reliability and reproducibility in functional connectomics. *Scientific data*, 1(1):140049, 2014.

Supplementary Material

A. ROI baseline for comparison: BrainMAE

A.1 Architecture and Qualitative Reconstruction

To provide a controlled ROI-level comparison, we implement BrainMAE as a masked-reconstruction baseline using MMP-360 ROI time series. It operates on MMP-360 ROI time series [19], yielding inputs of shape $(B, 360, T)$. As shown in Fig. S1(a,b), a 1D convolutional patch embedding with kernel size 5 and stride 5 tokenizes the temporal axis into $T/5$ tokens. BrainMAE follows a SimMIM-style objective [64], where random temporal patches are masked and the full ROI–time sequence is reconstructed from visible tokens. The backbone is a Nystromformer encoder–decoder [65] with hidden size 1024 and 8 attention heads, trained with MSE loss over all ROI–time positions.

We train two BrainMAE variants. An HCP-only model is trained on HCP resting-state data to evaluate reconstruction generalization across new subjects, tasks, stimuli, and disease cohorts. For the main baseline comparisons, we train a larger multi-dataset model on HCP [52], CHCP [18], HCP Task [52], the Nakai–Nishimoto cognitive-task dataset [37], HCP Movie [52], StudyForrest [21], NSD [2], ABIDE [11], ADHD [33], and ADNI [24]. This model is used only as the BrainMAE baseline and is not counted in the 22-dataset BrainWorld corpus.

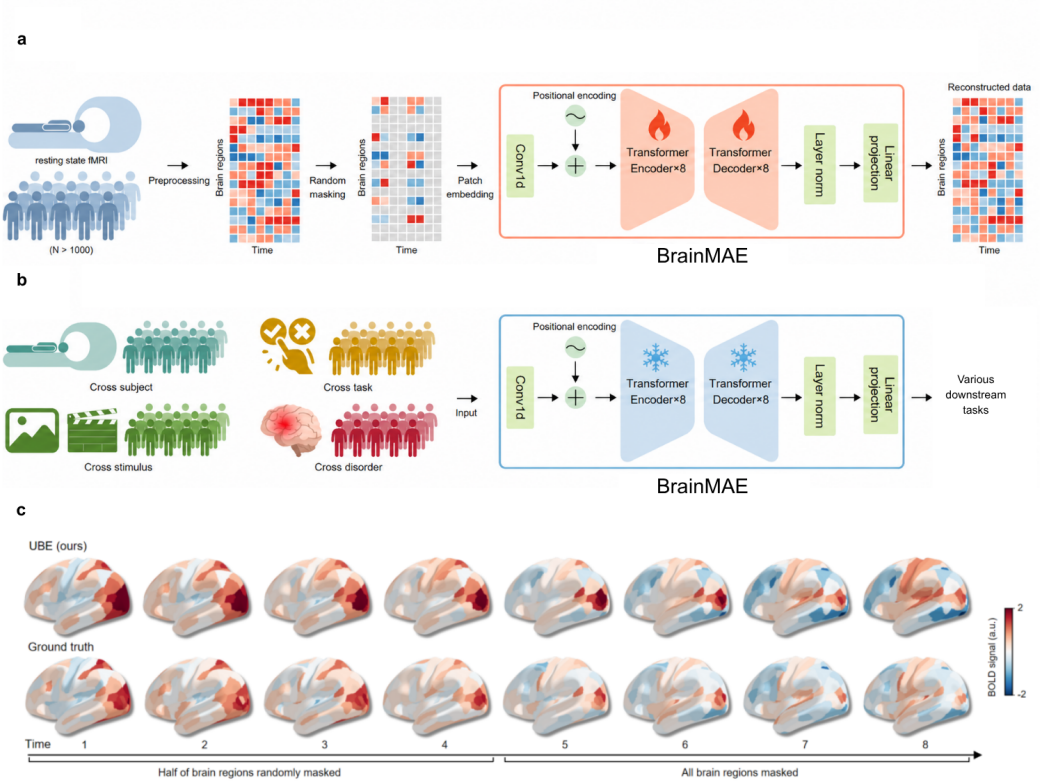

Figure S1: Overview of BrainMAE. (a) BrainMAE is pretrained on MMP-360 ROI time series using random temporal masking and full-sequence reconstruction. (b) After pretraining, the encoder is reused as a frozen feature extractor for probing-style transfer across subjects, tasks, naturalistic stimuli, and disease cohorts. (c) Qualitative reconstruction examples under half-region masking and full-region masking, showing that BrainMAE recovers temporally coherent ROI dynamics even under strong masking.

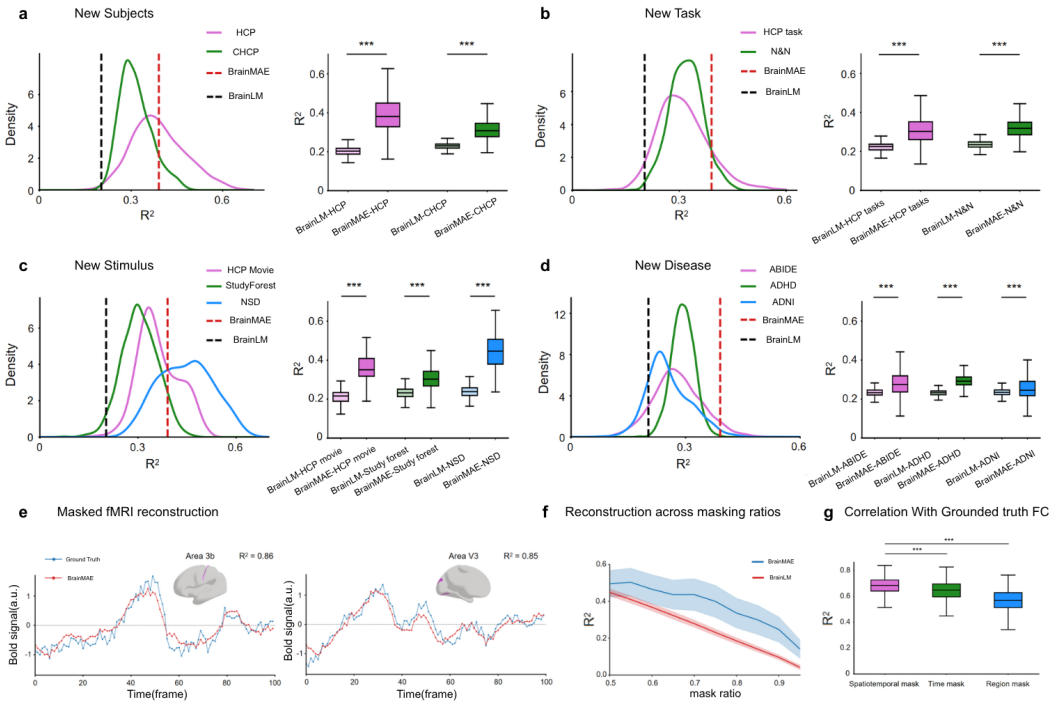

Figure S2: Reconstruction statistics of BrainMAE across transfer conditions and masking settings. (a–d) Reconstruction generalization of the HCP-only BrainMAE on new subjects, unseen tasks, naturalistic-stimulus data, and disease cohorts, compared with BrainLM. (e) Representative ROI-level time-course reconstruction. (f) Reconstruction performance across masking ratios. (g) Correlation between reconstructed and ground-truth functional connectivity under different masking strategies.

A.2 Reconstruction Statistics Across Conditions and Masking Settings

Figure S2 evaluates the reconstruction generalization of the HCP-only BrainMAE across multiple transfer settings. Panels (a–d) show that, after being trained only on HCP resting-state data, BrainMAE remains competitive on new subjects, unseen tasks, naturalistic-stimulus data, and disease cohorts, consistently outperforming BrainLM under the same ROI-level reconstruction setting. Panels (e–g) further examine reconstruction behavior under different masking settings. BrainMAE recovers representative ROI time courses, remains robust across masking ratios, and preserves functional-connectivity structure. These results indicate that BrainMAE learns stable spatiotemporal dynamics, supporting its use as a strong predictive-reconstruction baseline.

A.3 Integrated-Gradient and Connectivity Analysis

We further analyze whether BrainMAE captures inter-regional dependencies during reconstruction using an attribution-based approach. Let $X \in \mathbb{R}^{N\times T}$ denote an ROI–time sequence with $N = 360$ ROIs and T time points. For each target ROI j, we mask its signal to obtain $X^{(j)}_{\text{mask}}$, reconstruct it with the pretrained BrainMAE f_θ, and compute integrated gradients from a constant baseline X_0 to the masked input [50]. The interpolation path is defined as

$$X^{(j)}_\alpha = X_0 + \alpha\left(X^{(j)}_{\text{mask}} - X_0\right), \quad \alpha \in [0,1].$$

The target reconstruction loss is

$$\mathcal{L}_j(\alpha) = \left\| f_\theta\left(X^{(j)}_\alpha\right)_{j,:} - X_{j,:} \right\|_2^2.$$

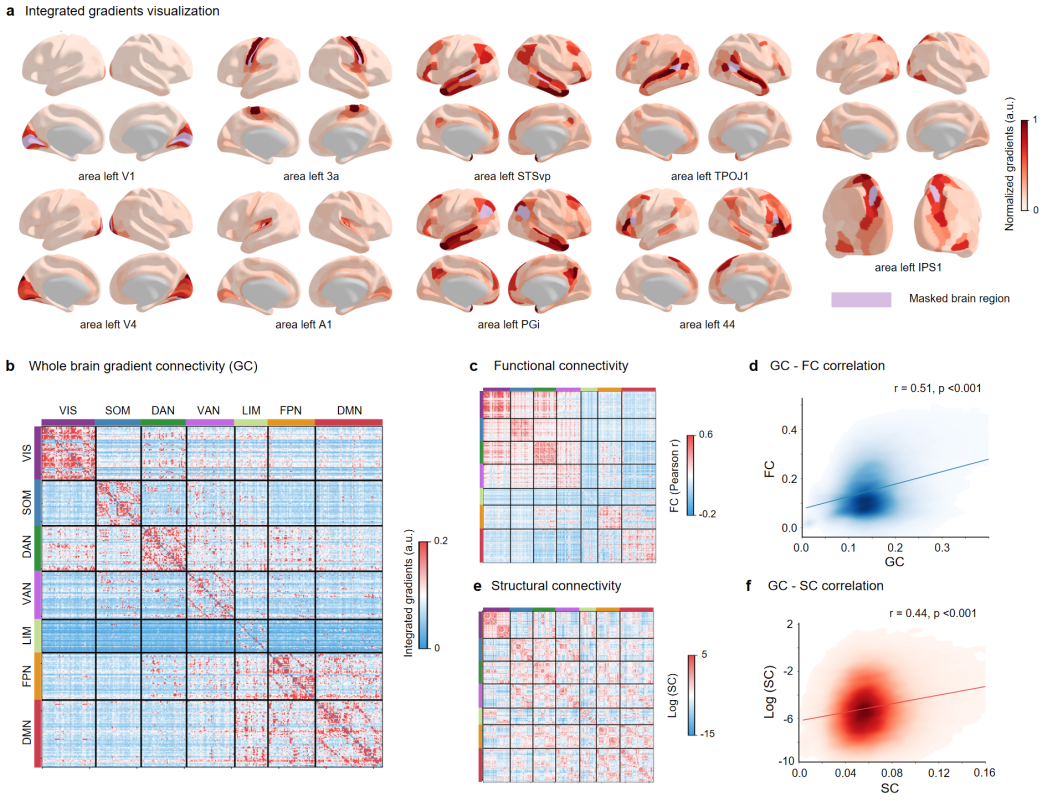

Figure S3: Integrated-gradient and connectivity analysis of BrainMAE. (a) Integrated-gradient attribution maps for representative masked target ROIs. (b) Whole-brain gradient-connectivity (GC) matrix aggregated across all 360 ROIs. (c) Functional connectivity (FC) matrix. (d) Correlation between GC and FC. (e) Structural connectivity (SC) matrix. (f) Correlation between GC and SC. Together, these results suggest that BrainMAE learns distributed inter-regional dependencies consistent with functional and structural brain organization during reconstruction.

For source ROI i, we define its attribution to target ROI j as

$$A_{i\to j} = \frac{1}{T}\sum_{t=1}^{T}\left|X^{(j)}_{\text{mask},i,t} - X_{0,i,t}\right| \cdot \frac{1}{K}\sum_{k=1}^{K}\left|\frac{\partial \mathcal{L}_j(\alpha_k)}{\partial X^{(j)}_{\alpha_k,i,t}}\right|, \quad \alpha_k = \frac{k-1}{K-1},$$

where $K = 50$ interpolation steps are used. The resulting attributions form a directed gradient-connectivity matrix $GC \in \mathbb{R}^{N\times N}$, with $GC_{i,j} = A_{i\to j}$, and diagonal entries set to zero. Larger $GC_{i,j}$ values indicate that source ROI i has a stronger influence on reconstructing the masked target ROI j.

Figure S3 visualizes the resulting attribution and connectivity patterns. Panel (a) shows representative integrated-gradient maps, indicating that reconstruction depends on distributed and anatomically structured patterns rather than purely local information. Panel (b) aggregates these attributions across all 360 ROIs to form a whole-brain gradient-connectivity matrix. Panels (c–f) compare GC against functional connectivity (FC) and structural connectivity (SC). GC is positively correlated with both FC ($r = 0.51, p < 0.001$) and SC ($r = 0.44, p < 0.001$), suggesting that BrainMAE learns inter-regional dependencies consistent with functional and structural brain organization during reconstruction.

A.4 Downstream Transfer Evaluation

We further evaluate whether BrainMAE learns representations that transfer beyond masked reconstruction. The pretrained BrainMAE encoder is frozen and used as a feature extractor, followed

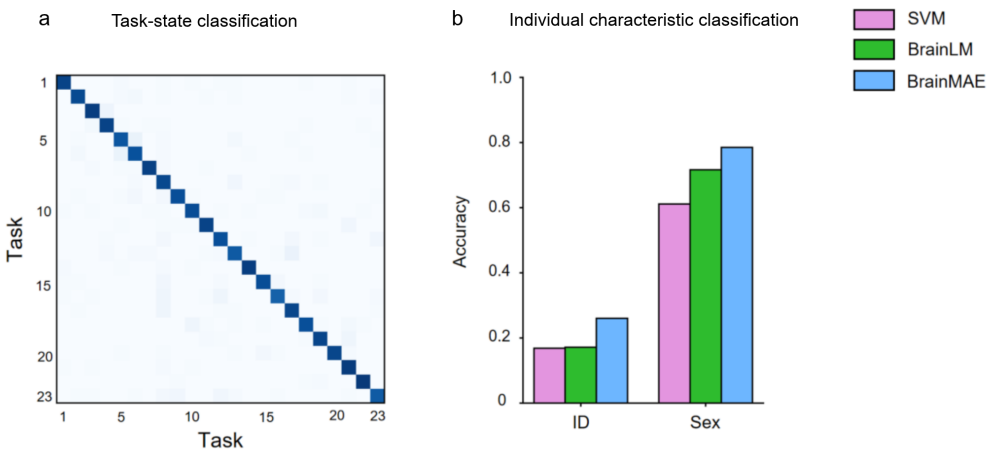

Figure S4: Downstream transfer evaluation of BrainMAE. (a) HCP task-state classification across 23 task categories, showing strong separation of task-specific representations. (b) Individual-level characteristic prediction, including subject identification and sex classification, compared with simpler representation baselines.

by lightweight downstream classifiers. Figure S4(a) shows HCP task-state classification across 23 task categories. The near-diagonal confusion matrix indicates that BrainMAE representations preserve strong task-discriminative information. Figure S4(b) further evaluates individual-level characteristics, including subject identification and sex classification, where BrainMAE achieves higher accuracy than simpler representation baselines. These results show that BrainMAE is not only a reconstruction baseline, but also a competitive transfer baseline with meaningful task- and subject-level representations.

B. Supplementary Results

B.1 Additional 4D fMRI Generation Visualizations

To complement the quantitative generation results, we provide qualitative visualizations of BrainWorld predictions on three representative datasets: CineBrain (Fig. S5), SALD (Fig. S6), and HCP Movie (Fig. S7). For each dataset, we visualize predictions at multiple rollout horizons using axial, coronal, and sagittal slices, together with the corresponding ground truth and absolute error maps. To avoid cherry-picking visually favorable examples, the displayed subject or run is selected as the median-performing case according to MSE among all evaluated samples of the corresponding dataset. Across CineBrain and SALD, BrainWorld preserves coherent whole-brain spatial structure over increasing horizons, while the HCP Movie results illustrate the greater difficulty of external naturalistic-stimulus prediction.

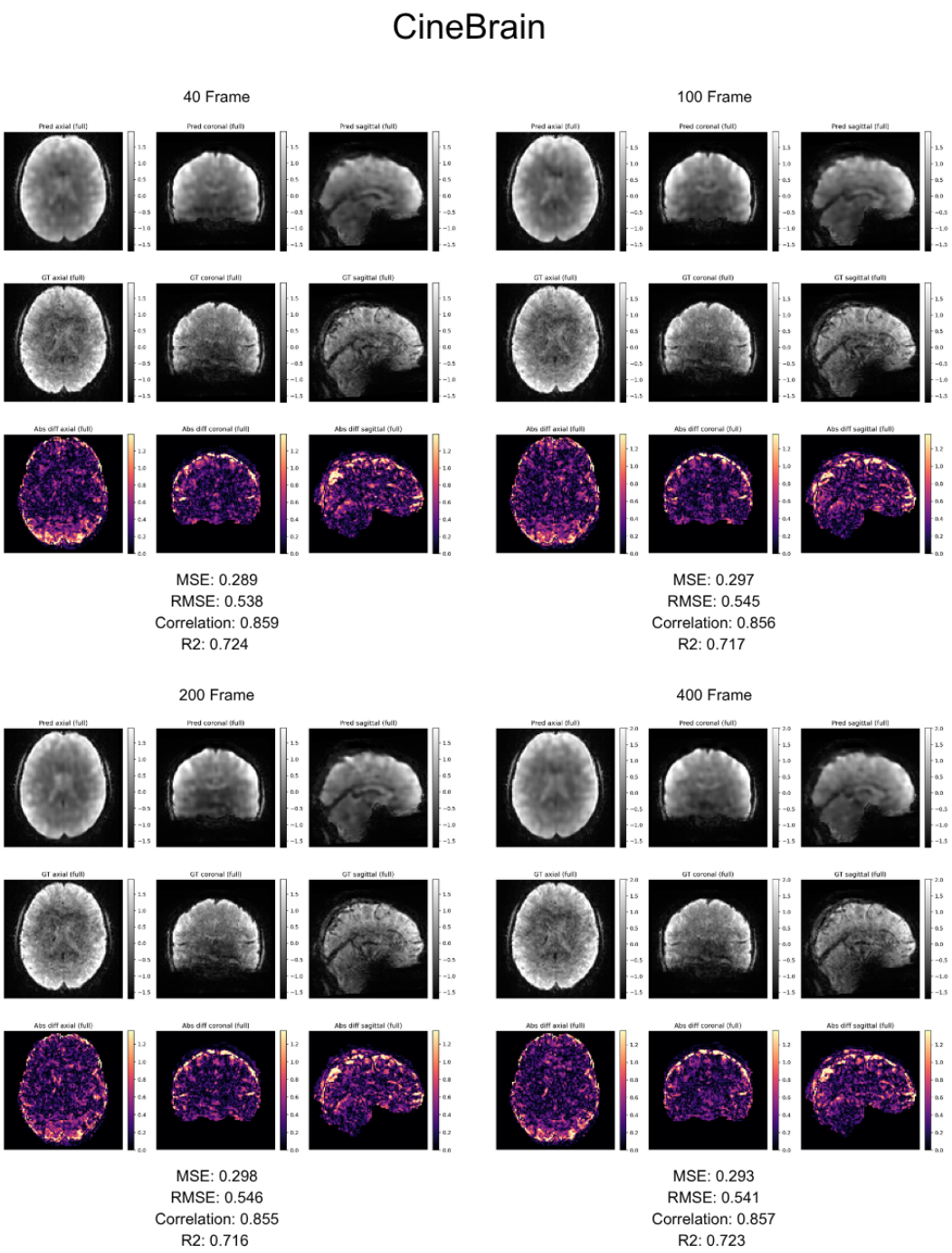

Figure S5: Qualitative visualization of BrainWorld generation on CineBrain across prediction horizons. For each horizon, axial, coronal, and sagittal slices are shown for generated fMRI, ground truth, and absolute error maps. The displayed subject/run is selected as the median-performing case according to MSE among all evaluated CineBrain samples, rather than as the best case. Metrics below each panel report mean squared error (MSE), root mean squared error (RMSE), Pearson correlation, and R^2.

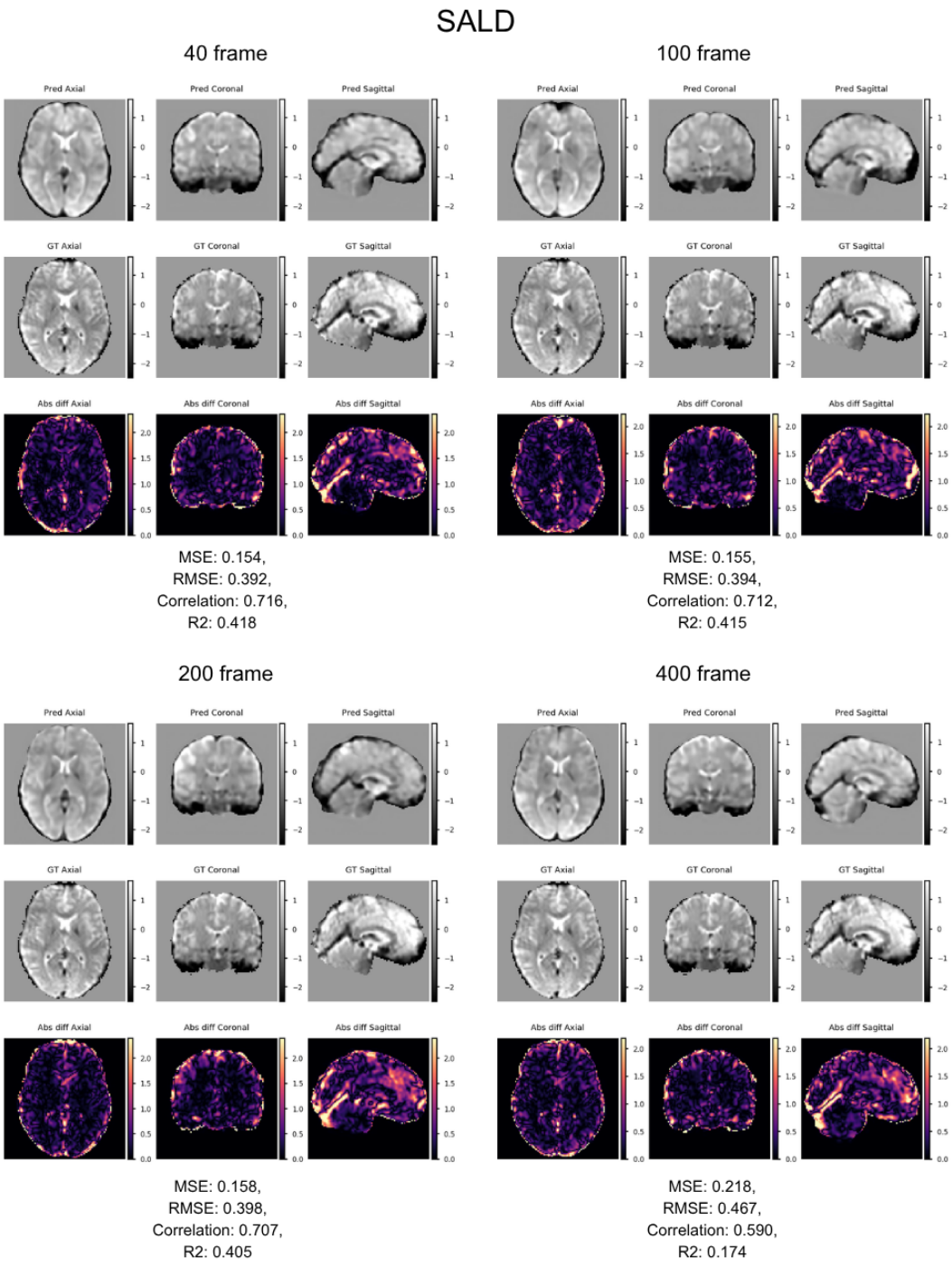

Figure S6: Qualitative visualization of BrainWorld generation on SALD across prediction horizons. For each horizon, axial, coronal, and sagittal slices are shown for generated fMRI, ground truth, and absolute error maps. The displayed subject is selected as the median-performing case according to MSE among all evaluated SALD samples, rather than as the best case. Metrics below each panel report MSE, RMSE, Pearson correlation, and R^2.

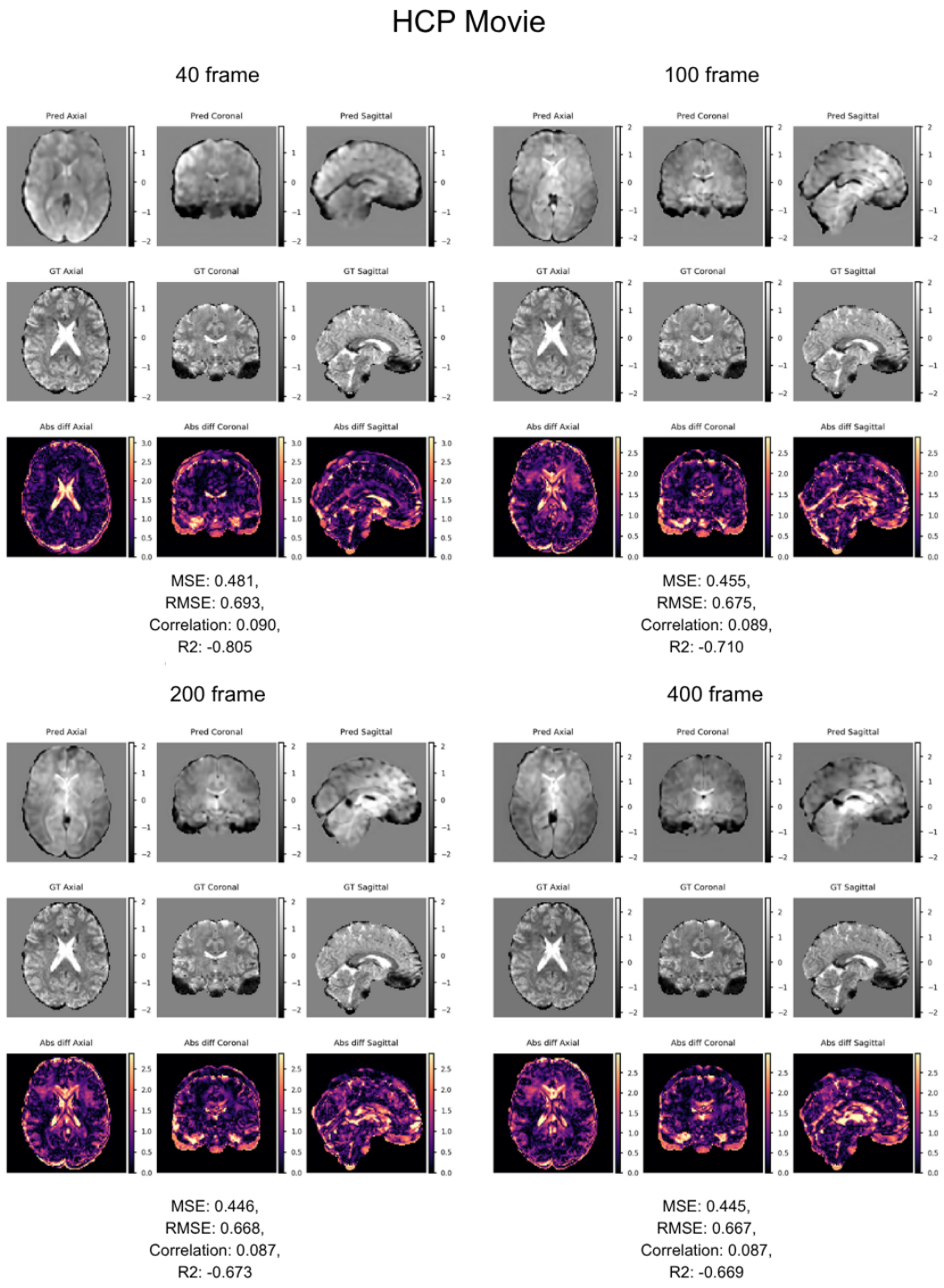

Figure S7: Qualitative visualization of BrainWorld generation on HCP Movie across prediction horizons. For each horizon, axial, coronal, and sagittal slices are shown for generated fMRI, ground truth, and absolute error maps. The displayed run is selected as the median-performing case according to MSE among all evaluated HCP Movie runs, rather than as the best case. Metrics below each panel report MSE, RMSE, Pearson correlation, and R^2.

Table S1: Generation performance for 40-frame prediction across HCP, SALD, CineBrain, and HCP Movie. MSE, RMSE, and Pearson correlation are reported. Lower MSE/RMSE and higher correlation indicate better performance. Red indicates the best performance and Underline indicates the second-best performance. * denotes a large favorable effect size with Cohen's $d \geq 0.8$ compared with the best-performing non-BrainWorld baseline for the same dataset, horizon, and metric.

Model \ Dataset	HCP LR1 *Internal*			SALD *External*		
	MSE ↓	**R ↑**	**RMSE ↓**	**MSE ↓**	**R ↑**	**RMSE ↓**
BrainMAE	0.986±0.126	0.147±0.140	0.991±0.062	1.038±0.097	0.055±0.181	1.018±0.047
BrainLM	0.925±0.084	0.265±0.087	0.961±0.041	0.573±0.292	0.429±0.156	0.738±0.171
NeuroSTORM	0.514±0.020	0.182±0.033	0.716±0.014	0.590±0.023	0.170±0.019	0.768±0.015
MONAI	**0.157±0.014**	**0.666±0.032**	**0.396±0.017**	0.269±0.030	0.496±0.069	0.517±0.029
BrainWorld	0.175±0.027	0.659±0.052	0.417±0.029	**0.186±0.074***	**0.657±0.140***	**0.425±0.074***

Model \ Dataset	CineBrain *Internal*			HCP Movie *External*		
	MSE ↓	**R ↑**	**RMSE ↓**	**MSE ↓**	**R ↑**	**RMSE ↓**
BrainMAE	0.781±0.083	0.456±0.107	0.882±0.048	1.001±0.031	0.075±0.083	1.000±0.016
BrainLM	0.766±0.407	0.522±0.128	0.850±0.214	1.059±0.162	0.133±0.069	1.027±0.071
NeuroSTORM	0.795±0.034	0.628±0.004	0.891±0.019	0.517±0.008	0.082±0.009	0.719±0.006
MONAI	0.422±0.022	0.752±0.016	0.649±0.017	0.871±0.756	0.056±0.041	0.857±0.370
BrainWorld	**0.292±0.021***	**0.859±0.009***	**0.540±0.019***	**0.405±0.095***	**0.180±0.152**	**0.631±0.077***

B.2 Generation Statistics at Additional Rollout Lengths

We report quantitative generation performance at three rollout lengths: 40, 100, and 400 frames (Tables S1–S3). For each horizon, we evaluate generation quality on HCP LR1, SALD, CineBrain, and HCP Movie using MSE, RMSE, and Pearson correlation. Lower MSE/RMSE and higher correlation indicate better reconstruction of future fMRI dynamics. Across rollout lengths, BrainWorld$_{\text{F+S-prior}}$ maintains stable performance and achieves the best or near-best results on most external evaluation settings, especially on SALD and HCP Movie. The results also show that performance generally degrades as the prediction horizon increases, reflecting the increasing difficulty of long-horizon autoregressive generation.

B.3 Generated-Example Augmentation at 200-Frame Horizon

To further examine whether the augmentation effect observed in the main 400-frame setting also holds at a shorter generation horizon, we conduct an additional GE augmentation ablation on HCP-200 and SALD-200 in Fig. S8. The real training set is kept fixed, and generated samples from the 200-frame setting are added at different GE ratios. For HCP-200, we evaluate gender classification accuracy, whereas for SALD-200, we evaluate age-regression performance using Pearson correlation. The line plots show the absolute downstream performance under each GE ratio, and the heatmaps summarize the corresponding performance changes relative to $\mathrm{GE} = 0$. Although the gains from 200-frame generated samples are generally weaker than those obtained with 400-frame generation in the main experiments, BrainWorld still provides consistent positive improvements and remains stronger than the competing baselines, while several baselines show limited gains or degradation as more generated samples are added.

Table S2: Generation performance for 100-frame prediction across HCP, SALD, CineBrain, and HCP Movie. MSE, RMSE, and Pearson correlation are reported. Lower MSE/RMSE and higher correlation indicate better performance. Red indicates the best performance and Underline indicates the second-best performance. * denotes a large favorable effect size with Cohen's $d \geq 0.8$ compared with the best-performing non-BrainWorld baseline for the same dataset, horizon, and metric.

Model \ Dataset	HCP LR1 *Internal*			SALD *External*		
	MSE ↓	**R ↑**	**RMSE ↓**	**MSE ↓**	**R ↑**	**RMSE ↓**
BrainMAE	1.001±0.071	0.096±0.102	1.000±0.035	1.018±0.055	0.033±0.122	1.009±0.027
BrainLM	0.971±0.032	0.171±0.059	0.985±0.016	0.724±0.284	0.331±0.148	0.836±0.161
NeuroSTORM	0.944±0.038	0.079±0.022	0.971±0.019	0.991±0.032	0.072±0.017	0.996±0.016
MONAI	0.318±0.017	0.394±0.030	0.564±0.015	0.283±0.017	0.510±0.036	0.532±0.016
BrainWorld	**0.181±0.026***	**0.645±0.050***	**0.425±0.028***	**0.204±0.078***	**0.623±0.146***	**0.445±0.077***

Model \ Dataset	CineBrain *Internal*			HCP Movie *External*		
	MSE ↓	**R ↑**	**RMSE ↓**	**MSE ↓**	**R ↑**	**RMSE ↓**
BrainMAE	0.908±0.068	0.299±0.077	0.952±0.037	1.003±0.018	0.042±0.055	1.001±0.009
BrainLM	0.840±0.574	0.473±0.123	0.878±0.271	1.060±0.138	0.077±0.042	1.028±0.062
NeuroSTORM	1.182±0.040	0.368±0.006	1.087±0.019	0.932±0.017	0.023±0.005	0.965±0.009
MONAI	0.684±0.027	0.549±0.020	0.827±0.017	0.595±0.322	0.046±0.034	0.747±0.191
BrainWorld	**0.291±0.021***	**0.860±0.010***	**0.539±0.019***	**0.437±0.045**	**0.114±0.072**	**0.660±0.035**

Table S3: Generation performance for 400-frame prediction across HCP, SALD, CineBrain, and HCP Movie. MSE, RMSE, and Pearson correlation are reported. Lower MSE/RMSE and higher correlation indicate better performance. Red indicates the best performance and Underline indicates the second-best performance. * denotes a large favorable effect size with Cohen's $d \geq 0.8$ compared with the best-performing non-BrainWorld baseline for the same dataset, horizon, and metric.

Model \ Dataset	HCP LR1 *Internal*			SALD *External*		
	MSE ↓	**R ↑**	**RMSE ↓**	**MSE ↓**	**R ↑**	**RMSE ↓**
BrainMAE	0.998±0.023	0.043±0.055	0.999±0.012	1.006±0.012	0.013±0.060	1.003±0.006
BrainLM	0.995±0.010	0.081±0.032	0.997±0.005	0.951±0.180	0.142±0.175	0.971±0.093
NeuroSTORM	1.208±0.049	0.013±0.015	1.099±0.022	1.270±0.041	0.005±0.011	1.127±0.018
MONAI	0.405±0.024	0.259±0.040	0.636±0.019	0.288±0.015	0.508±0.032	0.537±0.014
BrainWorld	**0.185±0.027***	**0.637±0.052***	**0.429±0.028***	**0.218±0.082***	**0.596±0.155**	**0.460±0.081***

Model \ Dataset	CineBrain *Internal*			HCP Movie *External*		
	MSE ↓	**R ↑**	**RMSE ↓**	**MSE ↓**	**R ↑**	**RMSE ↓**
BrainMAE	1.006±0.014	0.113±0.031	1.003±0.007	0.996±0.004	0.020±0.028	0.998±0.002
BrainLM	0.790±0.256	0.297±0.106	0.877±0.148	1.059±0.164	0.037±0.021	1.027±0.072
NeuroSTORM	1.653±0.048	0.078±0.004	1.286±0.019	1.230±0.038	-0.008±0.006	1.109±0.017
MONAI	0.982±0.013	0.283±0.010	0.991±0.006	0.528±0.230	0.056±0.033	0.713±0.139
BrainWorld	**0.295±0.023***	**0.857±0.011***	**0.543±0.021***	**0.454±0.020**	**0.081±0.037**	**0.674±0.015**

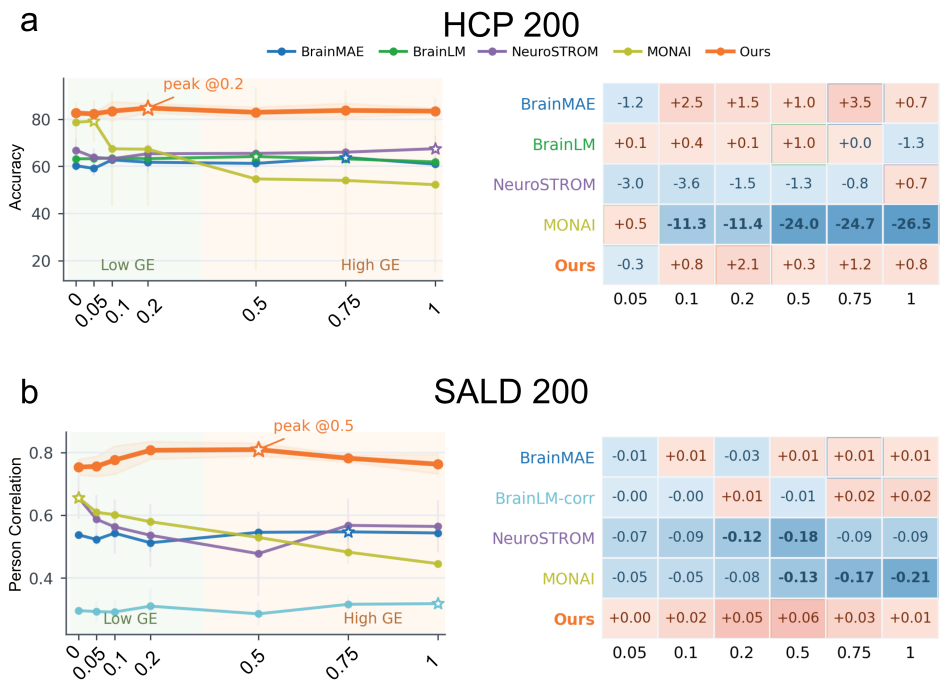

Figure S8: Additional generated-example (GE) augmentation results on HCP-200 and SALD-200. (a) HCP-200 gender classification accuracy under different GE ratios, with the corresponding performance changes relative to GE $= 0$ shown on the right. (b) SALD-200 age regression performance measured by Pearson correlation under different GE ratios, with relative changes shown on the right. BrainWorld achieves consistently strong performance across GE ratios, whereas most baselines show limited gains or performance degradation when more generated samples are added.

B.4 Predictive-Pretraining Checkpoint Transfer

We evaluate four checkpoints from the same predictive-pretraining trajectory using frozen DiT representations and linear downstream probes on identical 40-frame fMRI inputs.

Table S4: Linear-probe performance across predictive-pretraining checkpoints.

Checkpoint (updates)	HCP Gender ACC ↑	SALD Age MSE ↓
0.5k	0.7964 ± 0.0124	0.4305 ± 0.0207
1k	0.8112 ± 0.0057	0.3986 ± 0.0378
2k	0.8161 ± 0.0057	0.3848 ± 0.0274
3k	0.8259 ± 0.0075	0.3672 ± 0.0373

As pretraining progresses, HCP gender accuracy increases by 2.95 percentage points, while SALD age-regression MSE decreases by 14.7%. This indicates improved representation transfer along the training trajectory, without establishing a causal advantage over alternative pretraining objectives.

C. Details of Materials

C.1 Preprocessing

fMRI preprocessing. All fMRI scans are converted to a common 4D voxel space before model training and evaluation. For datasets without released minimally preprocessed volumes, we perform standard preprocessing including motion correction, skull stripping, nuisance removal, and spatial normalization. All functional volumes are registered to the MNI152 template and resampled to 2 mm isotropic resolution. To reduce variability in acquisition protocols, BOLD time series are temporally interpolated to a unified TR of 0.72 s using cubic interpolation. Each complete fMRI time series undergoes background-aware global Z-scoring before being segmented into 40-frame windows. Normalization statistics are computed only over non-background brain voxels, and background regions are kept fixed. Each 3D volume is cropped or zero-padded to a fixed spatial size of $96 \times 96 \times 96$. The resulting non-overlapping windows are used as the basic input units for VAE encoding and diffusion-based future-window prediction. Due to the computational cost of large-scale fMRI preprocessing, for some datasets we use the subset of subjects that have completed preprocessing and quality control, rather than the entire public release. Within each dataset, the retained subjects are split at the subject level into training, validation, and test sets with a ratio of 7:1:2, unless otherwise specified by a task-specific evaluation protocol.

sMRI preprocessing. Structural MRI is processed separately to provide subject-level anatomical conditioning. For each T1-weighted scan, we use a unified T1 preprocessing pipeline consisting of N4 bias-field correction, skull stripping with HD-BET, and affine registration to the MNI152NLin2009cAsym 1-mm T1w template. The registered image is then reoriented to canonical space, non-finite values are set to zero, and the volume is resampled to the fixed 1 mm MNI reference grid with spatial size $193 \times 229 \times 193$. The resulting single-channel T1 volume is encoded by the frozen structural MRI encoder to obtain a subject-level anatomical embedding, which serves as a stable structural prior for conditioning future fMRI generation.

C.2 Datasets

Table S5 summarizes the role of each dataset in BrainWorld training and evaluation. The training row reports the union of datasets used across VAE training, mixed-state DiT training, or stimulus-conditioned DiT fine-tuning; individual datasets need not participate in every stage. In-domain evaluation uses subject-level held-out splits from datasets represented during model development, whereas out-of-domain datasets are excluded from BrainWorld training.

Amsterdam Open MRI Collection (AOMIC). AOMIC contributes two young-adult cohorts to our pretraining pool. PIOP1 [48] contains 168 participants (22.15 ± 1.83 years), with sex labels available for 161 participants (94 males and 67 females). PIOP2 [48] contains 180 participants (21.99 ± 1.83

Table S5: Dataset roles across BrainWorld training and evaluation.

Role	Datasets
VAE/DiT training	HCP-Rest, ABCD, CHCP, CoRR, Caltech, CineBrain, FCON, HBN, HCP-Task, ISYB, PIOP1, PIOP2, PPMI, RBC-BHRC, RBC-NKI, SLIM, StudyForrest, Emo-FilM
In-domain downstream	HCP-Rest, RBC-NKI, RBC-BHRC, ABCD
Out-of-domain downstream	ABIDE, SALD, ADNI
In-domain generation	HCP-Rest, CineBrain
Out-of-domain generation	SALD, HCP-Movie

years), with sex labels available for 179 participants (98 males and 81 females). The discrepancy between the total number of participants and the reported sex counts is due to missing sex labels for a small number of participants. We treat the two releases as independent datasets because they were acquired as separate AOMIC studies.

Chinese Human Connectome Project (CHCP). CHCP [18] is used to broaden the population coverage of the pretraining corpus beyond Western cohorts. We include 244 participants, with 117 males and 127 females, spanning 18–79 years of age. Its wider age range complements the young-adult HCP and AOMIC cohorts.

Imaging Chinese Young Brains (ISYB). ISYB [16] provides an additional young-adult Chinese cohort. The subset used here includes 130 healthy Han Chinese participants aged 18–30 years (22.52 ± 2.61 years), with 98 females and 32 males. It is included as part of the multi-dataset pretraining corpus.

Adolescent Brain Cognitive Development (ABCD). ABCD [9] introduces early-developmental samples into our dataset collection. We use 2400 participants from the 9–10-year-old baseline cohort. This dataset provides a pediatric contrast to the adult and lifespan datasets used elsewhere in the study.

Autism Brain Imaging Data Exchange (ABIDE). ABIDE [11] is used as an autism-related disease cohort for external evaluation. After our quality-control and filtering steps, the retained subset includes 609 participants from both autism spectrum disorder and typically developing control groups, with ages ranging from 6 to 58 years.

Human Connectome Project (HCP). HCP [52] provides the main high-quality young-adult reference cohort. We use 707 participants (28.79 ± 3.53 years) for pretraining, generation analysis, and downstream evaluation. HCP REST LR1, HCP-400, and HCP-200 in the main text denote different protocols or frame-length settings derived from this same cohort.

Parkinson's Progression Markers Initiative (PPMI). PPMI [32] is included in the training pool to broaden disease-state coverage. We retain 331 participants after quality control and label filtering, covering healthy controls, prodromal cases, and clinically diagnosed Parkinson's disease, with ages between 35 and 84 years.

Alzheimer's Disease Neuroimaging Initiative (ADNI). ADNI [24] supports evaluation on Alzheimer's disease progression. The curated subset contains 497 participants with complete resting-state fMRI and diagnostic labels. The retained subjects span 55–90 years and include cognitively normal controls, mild cognitive impairment, and Alzheimer's disease groups.

Southwest University Adult Lifespan Dataset (SALD). SALD [60] is used for external age-regression and generation evaluation because of its broad adult age coverage. We include resting-state fMRI from 493 participants. SALD-400 and SALD-200 refer to different frame-length protocols used in the main experiments, not to separate datasets.

Brazilian High-Risk Cohort (BHRC). We use the preprocessed BHRC [44] subset released through the Reproducible Brain Charts (RBC) benchmark resource [47], retaining 465 participants after preprocessing and label filtering. This dataset complements the HCP and SALD cohorts by providing an additional developmental evaluation setting.

Nathan Kline Institute (NKI). NKI [39] provides a community-ascertained lifespan neuroimaging cohort with resting-state fMRI and rich phenotypic annotations, making it suitable for evaluating individual-difference prediction across the lifespan. We use the preprocessed NKI subset released through the Reproducible Brain Charts (RBC) benchmark resource [47]. After preprocessing and label filtering, 717 participants are retained for age regression, and 331 participants with valid education labels are used for education-level classification. Education labels are grouped into three categories: primary education (Grades 1–6), secondary education (Grades 7–12), and higher education (those with graduate-level degrees).

HCP Task. HCP Task [52] is treated separately from resting-state HCP because it provides task-evoked brain states. The benchmark covers seven task families and 23 experimental conditions, including working memory, motor, emotion, gambling, language, relational reasoning, and social cognition.

StudyForrest. StudyForrest [21] is included as a naturalistic audiovisual-stimulus dataset. We use a seven-participant subset of the Phase-2 movie-watching acquisition. fMRI was acquired at 3 T while participants watched and listened to the German audiovisual version of *Forrest Gump*. Each participant completed eight runs and contributed 3,599 native volumes, yielding 56 runs in total.

CineBrain. CineBrain [15] is used as a naturalistic audiovisual dataset for evaluating stimulus-related brain dynamics. It contains simultaneous fMRI, electroencephalography (EEG), and electrocardiography (ECG) recordings collected while participants watched narrative video clips from *The Big Bang Theory*. The public description reports approximately six hours of audiovisual stimulation for each of six participants, making it suitable for long-form movie-driven fMRI generation and stimulus-conditioned analysis.

Emo-FilM. Emo-FilM [34] provides a naturalistic emotion-elicitation setting based on short film viewing. The dataset contains fMRI and physiological recordings from 30 healthy participants watching 14 emotion-inducing films, together with dense emotion annotations from an independent group of raters.

HCP Movie. HCP Movie refers to the HCP Young Adult 7T audiovisual movie-watching acquisition [52]. Participants watched and listened to short independent-film and Hollywood movie excerpts concatenated into four movie runs. We use a subset comprising 41 participants and 141 runs after preprocessing and quality control. HCP Movie is excluded from stimulus-adaptation training and reserved for external naturalistic-stimulus evaluation.

Healthy Brain Network (HBN). HBN [1] is included as a pediatric and adolescent neuroimaging cohort with rich phenotypic characterization. The project aims to build a large-scale biobank of participants aged 5–21 years, including multimodal brain imaging and behavioral, cognitive, and psychiatric assessments. In BrainWorld, HBN broadens developmental coverage during training and complements the adult cohorts.

1000 Functional Connectomes Project (FCON). The 1000 Functional Connectomes Project (FCON) [6] is a large multi-site resting-state fMRI resource. The original release aggregates over 1,200 resting-state fMRI datasets collected independently across 33 imaging sites. In BrainWorld, FCON is used to increase acquisition-site diversity and improve robustness to heterogeneous resting-state protocols.

Caltech Conte Center Dataset. The Caltech Conte Center [27] dataset provides multimodal neuroimaging and behavioral data from healthy adults. The release includes structural MRI, resting-state fMRI, movie-watching fMRI, and psychological assessments, with repeated acquisitions available for many participants. We use this dataset to add a behaviorally rich, multi-session adult cohort with both resting-state and naturalistic movie-viewing conditions.

Southwest University Longitudinal Imaging Multimodal Dataset (SLIM). The Southwest University Longitudinal Imaging Multimodal dataset (SLIM) [31] is a longitudinal multimodal neuroimaging cohort from Southwest University. It includes structural MRI, resting-state fMRI, diffusion MRI, and behavioral measures collected across up to three time points over approximately three and a half years. We include SLIM to introduce longitudinal resting-state data and improve robustness to repeated-measure and test–retest variability.

Consortium for Reliability and Reproducibility (CoRR). The Consortium for Reliability and Reproducibility (CoRR) [70] is used as a test–retest-oriented resting-state resource. CoRR aggregates resting-state fMRI and diffusion imaging data from multiple laboratories to support reliability and reproducibility analyses in functional and structural connectomics.

C.3 Condition Preparation for BrainWorld

Overview of conditions. BrainWorld supports two types of conditions: intrinsic subject/context conditions and optional external stimulus conditions. Intrinsic conditions include a subject-level structural MRI embedding and a past-window functional connectivity embedding, whereas stimulus conditions include video and audio embeddings for naturalistic-stimulus datasets. These conditions are prepared by frozen modality-specific encoders and then projected by the condition encoder into token-level and global-level representations for the DiT backbone.

Intrinsic structural and dynamic-context conditions. BrainWorld uses two non-stimulus conditions to describe subject-specific anatomy and recent functional context. The first is an intrinsic structural condition derived from sMRI, which provides a stable anatomical prior for each subject. Specifically, each preprocessed T1-weighted MRI is encoded by a frozen BrainIAC [51] encoder to obtain an MRI embedding. The second is a dynamic-context condition derived from the past fMRI window. We compute past functional connectivity (past FC) using the Schaefer-100 [45] parcellation and extract an FC embedding with a frozen BrainMass [67] encoder. The resulting MRI and past-FC embeddings are projected by the condition encoder into token-level and global-level representations for the DiT.

During training, we adopt classifier-free guidance (CFG)-style condition dropout [22]. When a condition is intentionally dropped, or when an MRI condition is unavailable for a subject, the corresponding condition is replaced by a learned null embedding. This design avoids requiring every dataset to provide all modalities, which is important for large-scale multi-dataset training where structural MRI, stimulus annotations, or other auxiliary conditions may be incomplete. It also enables BrainWorld to support conditional, partially conditional, and unconditional generation during inference.

Stimulus conditions. Video and audio are treated as optional external stimulus conditions for naturalistic-stimulus modeling. For the i-th fMRI window, we denote its blood-oxygen-level-dependent (BOLD) interval as $[b_i, b_i + L)$. To compensate for the hemodynamic delay between external stimuli and BOLD responses, we use a fixed lag of $\Delta = 5$ s and define the corresponding stimulus interval as

$$I_i = [\max(0, b_i - \Delta), \max(0, b_i + L - \Delta)] .$$

Each lagged stimulus interval is divided into $K = 4$ consecutive clips, each with a nominal duration of 7.2 s. Video clips are encoded with Video Joint-Embedding Predictive Architecture 2.1 (V-JEPA 2.1) ViT-L/16 [36] by uniformly sampling 64 frames per clip, resizing/cropping frames to 384×384, and mean-pooling the output tokens to obtain a clip-level video feature. Audio clips are resampled to 16 kHz, converted into 80-dimensional fbank features, encoded with W2v-BERT 2.0 [4], and aggregated with masked mean pooling over valid frames to obtain a clip-level audio feature. When available, each stimulus modality is represented by four embeddings for the corresponding fMRI window.

Naturalistic-stimulus datasets differ in how their stimulus timelines are defined. We therefore first construct a dataset-specific stimulus timeline and then apply the same 5-s lag and 7.2-s clip slicing rule. For CineBrain, multiple subjects watched common movie runs, so we build a shared run-level stimulus timeline and align each subject's fMRI segments to this common timeline. For StudyForrest [21], movie timing is reconstructed separately for each subject from the provided event logs; when a

clip crosses a run boundary, the corresponding subsegments are extracted from adjacent runs and concatenated. For HCP Movie [52], fMRI windows are generated with an overlapping-window protocol: each window contains 40 TRs, corresponding to 28.8 s, and adjacent windows are shifted by 17 TRs, corresponding to 12.24 s. Thus, neighboring fMRI windows may partially share stimulus content. Despite these dataset-specific timeline constructions, all available stimulus streams are converted into a common condition format of four lag-aligned clips per modality, each lasting 7.2 s.

Because stimulus datasets are much smaller than resting-state and disease datasets, we do not inject video/audio conditions during the initial mixed-state training stage. In this stage, naturalistic-stimulus datasets are trained together with resting-state and disease datasets, but their stimulus conditions are dropped and treated as absent. Emo-FilM is also trained in this way: we do not crop its corresponding stimulus clips, and instead treat it as a regular fMRI dataset without stimulus conditioning. Stimulus modalities are dataset-dependent. In the realized stimulus-adaptation training, CineBrain provides available V-JEPA video conditions, whereas StudyForrest provides W2v-BERT audio conditions. HCP Movie is excluded from stimulus-adaptation training and is evaluated using video conditions. Unavailable modalities are represented using explicit condition masks or learned null embeddings.

Condition injection. In the DiT backbone, MRI and past-FC embeddings provide intrinsic structural and dynamic-context conditions, while video and audio embeddings provide optional external stimulus conditions. MRI and FC embeddings are projected into token-level and global-level condition representations. Video and audio embeddings are encoded as four-token sequences, with an additional mean-pooled global vector for each modality. All available condition tokens are concatenated into a unified condition-token sequence and injected into each DiT block through cross-attention. In parallel, modality-level global vectors are fused by attention into a single global condition. This global condition is combined with the diffusion timestep embedding to generate adaptive shift, scale, and gate parameters for each DiT block, modulating the self-attention, cross-attention, and MLP branches. Therefore, BrainWorld uses a three-level conditioning mechanism: token-level cross-attention, global-level modality fusion, and block-level adaptive modulation. Missing or dropped modalities are represented by learned null tokens and null global embeddings, enabling robust training and inference under incomplete conditions.

C.4 Baseline Models

We compare BrainWorld with representative fMRI foundation models. These baselines are used for two evaluation purposes: 4D fMRI generation and downstream task prediction. BrainLM and NeuroSTORM are evaluated in both settings, serving as baselines for generation quality and downstream transfer performance, whereas MONAI is used only as a generation baseline. For each baseline, we follow the released implementation or reported architecture when available, and adapt the prediction head to the corresponding task setting used in our experiments. Unless otherwise specified, all baselines are evaluated under the same train/validation/test split as BrainWorld.

Table S6 positions BrainWorld relative to representative fMRI foundation and multimodal models in terms of input scale, learning objective, conditioning information, and intended output.

Table S6: Comparison of representative fMRI foundation and multimodal models.

Model	Input representation	Pretraining objective	Context / conditioning	Main output / use
BrainLM [8]	Atlas-parcellated ROI time series	Masked ROI–time reconstruction	Visible ROI patches	Signal reconstruction; representations
Brain-JEPA [12]	Atlas-parcellated ROI time series	Predict masked latent representations	Visible ROI blocks; position information	Transferable representations
BrainMass [67]	ROI-level FC networks	Masked connectivity modeling and latent alignment	Visible FC features; augmented views	Brain-network representations
Omni-fMRI [58]	Atlas-free voxel-level 4D fMRI	Masked voxel-patch reconstruction	Visible voxel patches	Voxel-level representations
Brain-DiT [63]	ROI time-series windows	Diffusion denoising of observed ROI–time windows	Demographic and clinical metadata	Diffusion-based representation learning
Brain Harmony [13]	Structural MRI and ROI-level fMRI	Unimodal pretraining and multimodal fusion	Structural and functional representations	Unified multimodal representations
BrainWorld	Voxel-level 4D fMRI latents	Conditional next-window diffusion prediction	Past FC, sMRI, optional video/audio	Future-window generation, autoregressive rollout, and transferable representations

BrainLM. BrainLM [8] is an ROI-level fMRI foundation model based on masked autoencoding. It represents each fMRI recording as AAL-424 ROI time series and learns to reconstruct masked brain activity tokens using a Transformer encoder–decoder. BrainLM was pretrained on 77,298 fMRI recordings from UK Biobank and HCP, corresponding to approximately 6,700 hours of fMRI data. For downstream fine-tuning, we use a batch size of 64, a learning rate of 1×10^{-5}, and 30 training epochs.

Brain-JEPA. Brain-JEPA [12] follows a joint-embedding predictive learning paradigm for fMRI dynamics. It uses 450 ROIs obtained from Schaefer-400 cortical parcels and Tian Scale-III subcortical regions. Instead of reconstructing masked signals directly, Brain-JEPA predicts target latent embeddings from visible context under spatial and temporal masking. The model is pretrained on UK Biobank data and then adapted to downstream tasks. In our setup, Brain-JEPA is trained or fine-tuned with a batch size of 16, a learning rate of 4×10^{-4}, and 50 epochs.

BrainMass. BrainMass [67] is a functional-connectivity foundation model built on Schaefer-100 brain networks. It learns network-level representations through self-supervised objectives that combine masked ROI modeling with latent representation alignment. The released model was pretrained on 26 datasets with 64,584 subjects, including UK Biobank, HCP, and ADHD-200. In our downstream experiments, BrainMass is fine-tuned with a batch size of 64, a learning rate of 2×10^{-4}, and 100 training epochs.

Brain-DiT. Brain-DiT [63] is included as a diffusion-based ROI-level foundation model baseline for downstream representation learning. It was pretrained on 349,898 fMRI sessions from 24 datasets, covering multiple brain states including resting, task-evoked, naturalistic-stimulus, disease, and sleep states. The main Brain-DiT model uses Schaefer-1000 ROI time series as input, with each fMRI window represented as an ROI–time matrix. An additional AAL424 variant is used in the original study to align with the atlas setting of BrainLM. During pretraining, Brain-DiT adds noise to an ROI–time window at a sampled diffusion timestep and trains a Diffusion Transformer to predict the v-target under a 1000-step denoising diffusion probabilistic model (DDPM) schedule. Demographic and clinical metadata, including age, sex, and diagnosis, are embedded as conditioning variables and injected into DiT blocks through AdaLN-Zero, with classifier-free condition dropout used during training.

SLIMBrain. SLIMBrain [56] is used as an efficient atlas-free voxel-level foundation model baseline. It first uses a lightweight temporal extractor to score and select informative fMRI windows, and then feeds only the selected windows into a 4D Hiera-JEPA encoder for fine-grained voxel-level representation learning. This selective-compute design reduces memory and data requirements while preserving voxel-level spatial details. We include SLIMBrain as an atlas-free representation baseline complementary to ROI-level models.

SwiFT. SwiFT [26] is a voxel-level 4D fMRI Transformer based on shifted-window attention. Unlike ROI-based methods, it takes raw 4D fMRI volumes as input and performs local spatiotemporal attention within 4D windows, making it a strong end-to-end baseline for volume-based fMRI representation learning. In our experiments, SwiFT is fine-tuned using a batch size of 16, a learning rate of 1×10^{-6}, weight decay of 0.01, and 30 training epochs.

NeuroSTORM. NeuroSTORM [54] is a voxel-level fMRI foundation model designed for direct learning from raw 4D fMRI volumes. Its backbone combines shifted-window processing with Mamba-style sequence modeling to reduce the cost of long-range spatiotemporal modeling. The model is pretrained on large-scale fMRI data from more than 50,000 subjects across UK Biobank, ABCD, and HCP. It also uses a spatiotemporal redundancy dropout strategy to prevent masked pretraining from relying on local redundancy. We include NeuroSTORM as a recent large-scale volume-based foundation model baseline.

Brain Harmony. Brain Harmony [13] is included as a multimodal structure–function foundation model baseline. It separately encodes T1-weighted structural MRI and fMRI dynamics, and then fuses the modality-specific latents into compact 1D brain-hub tokens. For the functional branch, Brain Harmony uses Schaefer-400 ROI time series and introduces temporal adaptive patch embedding to handle heterogeneous TRs. For the structural branch, T1-weighted volumes are encoded with a 3D masked autoencoding framework. During multimodal fusion, the structural and functional latents are compressed into shared hub tokens, which are then used for downstream prediction. Brain Harmony was pretrained on UK Biobank and ABCD, using 64,594 T1-weighted volumes and 70,933 fMRI time series. In our experiments, we use Brain Harmony as the main representation-level F+S fusion baseline to compare against BrainWorld's generative structure-function fusion. Brain Harmony fine-tunes downstream tasks for 50 epochs using AdamW with a learning rate of 5e-4, weight decay of 0.05, layer-wise learning-rate decay of 0.65, 5 warmup epochs, and cosine warmup-decay scheduling. Its linear-probe setting uses AdamW for 100 epochs with a learning rate of 1e-3, typically with batch size 8 in the public example scripts.

Omni-fMRI. Omni-fMRI [58] is used as an atlas-free voxel-level foundation model baseline. It avoids ROI parcellation and directly models 4D fMRI volumes with a dynamic patching strategy, where background regions are pruned and foreground regions are assigned adaptive patch sizes according to local spatiotemporal complexity. This design reduces token length while preserving fine-grained voxel information. The model was pretrained on 49,497 fMRI sessions across nine datasets and benchmarked on 11 datasets. For downstream adaptation, we fine-tune Omni-fMRI for 30 epochs using AdamW with a base learning rate of 1×10^{-5}, a head learning rate of 1×10^{-4}, weight decay of 0.05, layer decay of 0.75, 2 warmup epochs, linear warmup-decay scheduling, and a total batch size of 32.

BrainIAC. BrainIAC [51] is included as a structural MRI foundation model baseline. Unlike fMRI-specific baselines that operate on BOLD dynamics or ROI time series, BrainIAC is pretrained on 3D brain MRI volumes and learns general-purpose anatomical representations through self-supervised contrastive learning. It was pretrained on 32,015 multiparametric brain MRI scans and evaluated within a larger collection of 48,965 brain MRIs spanning diverse neurological conditions and MRI sequences. The model uses a SimCLR-style objective, where augmented 3D MRI patches from the same anatomical region form positive pairs and patches from other samples serve as negatives. Based on backbone benchmarking, the original study adopted a ViT-B encoder as the final BrainIAC backbone.

In our experiments, BrainIAC is used to represent the structural MRI modality. Preprocessed T1-weighted MRI volumes are passed through the frozen BrainIAC encoder to extract subject-level anatomical embeddings, which serve either as structural-only features or as MRI conditions in BrainWorld. We include BrainIAC to provide a strong structural representation baseline and to support comparison against multimodal methods that combine structural and functional information. BrainIAC fine-tunes downstream tasks for 30 epochs using AdamW with a single learning rate of 8e-4, weight decay of 1e-4, batch size 8, and a CosineAnnealingWarmRestarts scheduler under mixed-precision GPU training. Unlike Brain-Harmony, the current BrainIAC downstream code does not expose separate backbone/head learning rates, layer-wise decay, or warmup scheduling.

MONAI-based latent diffusion baseline. We use a MONAI-based latent diffusion model as a generation baseline for 4D fMRI future-window prediction [42]. Specifically, we use MONAI GenerativeModels to implement the Stage-1 vector-quantized variational autoencoder (VQ-VAE) tokenizer, but do not use any off-the-shelf MONAI pretrained weights; all components are trained on fMRI data. This baseline provides a medical-image generative comparison under the same next-window autoregressive setting as BrainWorld. Unlike BrainWorld, whose VAE is designed for 4D fMRI latent modeling, the MONAI tokenizer is a frame-wise 3D VQ-VAE.

In Stage 1, the VQ-VAE is trained on individual 3D fMRI frames. MONAI uses the same preprocessed fMRI inputs as BrainWorld: each complete fMRI time series undergoes background-aware global Z-scoring before being segmented into 40-frame windows, with normalization statistics computed over non-background brain voxels and background regions kept fixed. Each resulting window has shape $96 \times 96 \times 96 \times 40$ and is reordered from `xyzt` to `txyz`. The trained VQ-VAE compresses each 3D frame into a latent grid of size $4 \times 24 \times 24 \times 24$. For a 40-frame clip, the resulting frame-wise latents form a clip-level latent tensor of shape $[40, 4, 24, 24, 24]$. For the Stage-2 3D DiT, we merge the temporal and latent-channel dimensions, yielding

$$x_0 \in \mathbb{R}^{160\times24\times24\times24}.$$

In Stage 2, MONAI uses a conditional 3D DiT to generate the next-window latent. For each target window, the immediately preceding latent window from the same subject, session, and sequence is used as the anchor condition. The noisy target latent is divided into non-overlapping $4 \times 4 \times 4$ patches, producing 216 spatial tokens. The DiT uses width 768, depth 12, 12 attention heads, and a multilayer perceptron (MLP) ratio of 4.0. To follow the same autoregressive prediction protocol as BrainWorld, the model is conditioned on the anchor latent, an available dataset/state label, and a past-window FC embedding computed from the anchor fMRI window. The anchor latent and FC embedding are encoded into both global vectors for AdaLN-Zero modulation and condition tokens for cross-attention, allowing temporal context and FC-based functional information to influence the denoising process.

The diffusion objective follows v-prediction with a 1000-step DDPM schedule and a scaled-linear beta schedule. We train MONAI with AdamW using learning rate 1×10^{-4}, weight decay 0.01, and exponential moving average (EMA) decay 0.9999. Condition-label dropout is applied for robustness, and samples without FC features are handled using an explicit `has_fc` indicator and a learned null FC condition. During inference, MONAI starts from Gaussian noise, generates the target latent conditioned on the anchor latent, available label condition, and FC embedding, and then decodes the generated latent back to 4D fMRI space using the trained VQ-VAE decoder.

C.5 Configurations

This subsection summarizes the configuration details required to reproduce BrainWorld and the comparison baselines, including architectural parameters, condition encoders, training schedules, rollout settings, and downstream fine-tuning and linear-probing protocols. Tables S7 and S8 summarize the main hyperparameter settings used in BrainWorld, including the conditional latent DiT backbone, the default downstream probing protocol, and the VAE variants.

C.6 Computational Resources

All experiments were conducted on Linux GPU servers using PyTorch with mixed-precision training. BrainWorld VAE pretraining and conditional latent DiT training were performed on NVIDIA L40 graphics processing units (GPUs) with 48 GB memory using distributed data parallelism when applicable. The Stage-I VAE contains 3.56M parameters and was trained with a per-GPU batch size of 2 on 6 L40 GPUs, giving a total batch size of 12. The Stage-II conditional latent DiT contains 116.5M parameters and was trained with a total batch size of 32 on 8 L40 GPUs using distributed data parallelism, as summarized in Table S7. Stage-II DiT pretraining took approximately 2 days, corresponding to roughly 384 GPU-hours. Unless otherwise specified, mixed precision was enabled to reduce memory usage and improve training efficiency.

Downstream full fine-tuning was conducted using the protocol in Table S7, with a total batch size of 160 across 5 GPUs. Linear probing and baseline fine-tuning were substantially cheaper and were run with smaller GPU allocations depending on the memory requirement of each model.

Table S7: Key hyperparameter settings of the Conditional Latent DiT and its full-finetuning protocol. BS denotes batch size.

Config	*Value*
Conditional Latent DiT Pretraining	
optimizer	AdamW
learning rate / weight decay	1×10^{-4} / 0.02
training epochs	50
total batch size	8 GPUs $\times$ 4 BS $= 32$
diffusion steps / schedule	1000 / cosine
prediction target	v-prediction, with `x0_aux_weight` $= 0.05$
latent input shape	$[40, 16, 10, 12, 10]$
spatial patch size	$[2, 4, 2]$
tokens per window	$40 \times 5 \times 3 \times 5 = 3000$
transformer depth / heads / dim	16 / 8 / 512
total parameters	116,516,160
Full-finetuning Protocol	
optimizer	AdamW
DiT / aggregator / head learning rates	1×10^{-5} / 1×10^{-4} / 1×10^{-4}
DiT / aggregator / head weight decay	1×10^{-4} / 0 / 0
training epochs / patience	40 / 10
total batch size	5 GPUs $\times$ 32 BS $= 160$
gradient clipping	global norm 1.0
embedding timesteps	$[0, 50, 100, 150]$
captured transformer layers	$[-6, -4, -2, -1]$ layers
aggregator heads / dropout	4 / 0.1

For generation evaluation, future 4D fMRI windows were sampled autoregressively in latent space and decoded by the frozen VAE decoder. A 400-frame trajectory requires 10 sequential 40-frame rollout steps. For the full 400-frame evaluation pipeline, including autoregressive latent prediction, decoding, and ROI-level metric computation, the end-to-end runtime was approximately 95.5 hours.

C.7 Evaluation Details

Generative forecasting. For generative forecasting, we compare BrainWorld with two categories of baselines. The first category includes fMRI foundation models, including BrainMAE, BrainLM, and NeuroSTORM. Since these models are not originally designed for long-horizon generation, we adapt them with an autoregressive mask-then-fill protocol, where predicted frames are iteratively reused as context for subsequent prediction steps. We further finetune BrainMAE, BrainLM, and NeuroSTORM on HCP and CineBrain under the same forecasting protocol to ensure fair comparison. The second category is a MONAI-based latent diffusion baseline implemented with a conditional 3D DiT backbone. Following Seo et al. [46], we stack frame-wise latent features along the channel dimension to adapt the 3D latent diffusion framework to 4D fMRI future-window prediction.

For fair comparison with ROI-level forecasting baselines, BrainWorld's voxel-level predictions are parcellated into ROI time series using the corresponding atlas before metric computation. MSE is computed over the ROI-by-time matrices, and Pearson correlation is computed as the average temporal correlation across ROIs. For voxel-level visualization, predictions are decoded back to 4D fMRI volumes using the frozen VAE decoder and compared with ground-truth volumes at matched horizons.

Generated-example augmentation. For generated-example (GE) augmentation, the real training set is kept fixed and generated samples are added according to the GE ratio. $\mathrm{GE} = 0$ denotes training only with real samples, whereas $\mathrm{GE} = 1.0$ adds the same number of generated samples as real training samples. Generated examples inherit subject-level labels and sMRI conditions from their conditioning real samples, and validation and test sets contain only real samples. To avoid confounding augmentation with additional optimization, all GE ratios use the same number of optimization steps and the same validation-based early stopping protocol.

Table S8: Hyperparameter settings of the BrainWorld VAE configuration. BS denotes batch size.

Config	*Value*
VAE	
input shape convention	$[B, 1, T, D, H, W]$
canonical crop size	spatial crop $[80, 96, 80]$, input $T = 40$
stage channels	$[16, 32, 64, 96]$
latent dimension	16
residual blocks per stage	2
spatial downsampling stages	3
spatial-only frame chunk size	8
spatial kernel size	5
temporal kernel size	5, defined but unused in the bypassed path
dropout	0.0
temporal residual scale	0.1
wavelet energy channels	8
wavelet fusion stages	`enc2, dec2`
log-variance clamp	$[-6.0, 4.0]$
resulting latent shape	$[B, 40, 16, 10, 12, 10]$
optimizer	AdamW
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$
learning rate schedule	constant learning rate, no explicit scheduler
learning rate	3×10^{-4}
weight decay	1×10^{-3}
gradient clipping	global norm 1.0
mixed precision	`bf16` AMP
per-GPU batch size	2
total batch size	6 GPUs $\times$ 2 BS $= 12$
training epochs	5
reconstruction loss weights	foreground $= 1.0$, background $= 0.05$, bg-zero $= 0.02$
auxiliary loss weights	wavelet $= 0.05$, temporal $= 0.0$, $\beta_{\mathrm{KL}} = 1 \times 10^{-4}$
early stopping	enabled, patience 8, min delta $= 1 \times 10^{-4}$
total parameters	3,557,829

Downstream transfer. For downstream transfer, we evaluate both full fine-tuning and frozen-backbone linear probing. Full fine-tuning updates the DiT backbone, feature aggregator, and downstream head, whereas linear probing keeps the backbone frozen and trains only the lightweight downstream readout. We compare functional-only (F), structural-only (S), and multimodal (F+S) baselines. To isolate the effect of the fusion mechanism, we additionally include BrainWorld$_{\mathrm{F+S\text{-}concat}}$, a same-backbone representation-level fusion baseline that concatenates fMRI features and independently extracted sMRI embeddings before the downstream head. BrainWorld$_{\mathrm{F+S\text{-}prior}}$ instead injects sMRI through the DiT structural-prior pathway.

For classification tasks, we report accuracy and F1 score. For regression tasks, we report MSE and Pearson correlation. Age labels are z-score normalized using statistics from the training split, and MSE is reported in the normalized label space.

Statistical reporting. All downstream results are reported as mean $\pm$ standard deviation over three runs. Given the limited number of runs, we report Cohen's d as an effect-size measure rather than relying on significance testing. Asterisks in the main tables denote a large favorable effect size with Cohen's $d \geq 0.8$, computed in the favorable direction against the best external baseline for the same task and metric.

C.8 Additional Generation Analyses

Functional Connectivity Preservation. We further evaluate the correspondence between generated and ground-truth functional connectivity at the Yeo-7 [68] and AAL424 parcellation scales across different rollout horizons.

Table S9: Functional connectivity similarity across rollout horizons. Results are reported as Pearson correlations ($r \uparrow$) between generated and ground-truth FC. Values correspond to 40 / 100 / 200 / 400 frames, respectively. Bold indicates the better result at each horizon.

Dataset	**Model**	**Yeo-7 FC** $r \uparrow$	**AAL424 FC** $r \uparrow$
HCP-REST	BrainWorld	**.411 / .385 / .451 / .526**	**.261 / .327 / .331 / .346**
	MONAI	.383 / .373 / .421 / .465	.253 / .303 / .313 / .328
SALD	BrainWorld	.327 / **.194 / .249 / .233**	.280 / **.298 / .267 / .251**
	MONAI	**.379** / .120 / .100 / .129	**.310** / .121 / .079 / .077

BrainWorld outperforms MONAI on both FC metrics at the 400-frame horizon across HCP-REST and SALD, suggesting better preservation of functional connectivity structure during long-horizon generation.

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 main claims in the abstract and introduction are supported by the generation, augmentation, and downstream transfer experiments reported in the paper.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 5 explicitly discusses limitations of long-horizon rollout, condition control, external embeddings, and biological interpretation.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The architecture, preprocessing, datasets, configurations, compute resources, and evaluation protocols are documented in the main paper and Supplementary Section C.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: The public code repository is linked in the abstract, and implementation and training details are documented in Supplementary Section C.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Data splits, preprocessing, model configurations, training protocols, and evaluation details are provided in Section 4.1 and Supplementary Section C.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: The paper reports mean and standard deviation over repeated runs for the main quantitative experiments.

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Compute resources, including GPU settings and training cost, are reported in the supplementary material.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work uses de-identified neuroimaging datasets under their original data-use terms and follows the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: This is methodological research on de-identified neuroimaging data and does not propose a deployed clinical or automated decision-making system.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: The paper cites the original datasets and software assets and follows their corresponding licenses, access conditions, and data-use terms.

    Guidelines:

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

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

13. **New assets**

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

    Answer: [Yes]

    Justification: Code and documents are provided.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or newly recruited human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The study uses previously collected, de-identified datasets and performs no new participant recruitment, intervention, or data collection; the original datasets remain subject to their respective approvals and data-use terms.

    Guidelines:

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

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs were used only for writing and editing, not as a component of the research methodology.

    Guidelines:

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