Brain-DiT: A Universal Brain Scanner AI

How Diffusion Models Learn to Understand fMRI Brain Signals

Simplified version of: Brain-DiT: A Universal Multi-state fMRI Foundation Model with Metadata-Conditioned Pretraining (arXiv:2604.12683, Jun 2026)

By Junfeng Xia, Wenhao Ye, Xuanye Pan, Xinke Shen, Mo Wang, Quanying Liu — Southern University of Science and Technology, Shenzhen University

This paper presents Brain-DiT, the first fMRI foundation model trained on 349,898 brain scans from 24 datasets covering every major brain state: resting, task-focused, naturalistic viewing, disease, and sleep. Unlike prior models that only learn from resting-state data, Brain-DiT uses diffusion-based generative pretraining with metadata conditioning to learn representations that work across all brain states.

Background: What is fMRI?

Functional Magnetic Resonance Imaging (fMRI) measures brain activity by detecting changes in blood flow. When neurons fire, they consume oxygen, and blood flow increases to deliver more. fMRI captures these blood-oxygen-level-dependent (BOLD) signals, producing 4D maps of brain activity over time.

The brain is a dynamical system that continuously transitions between diverse states: resting (spontaneous activity), performing cognitive tasks, watching movies, sleeping, or affected by diseases like Alzheimer's or autism. Capturing representations that work across all these states is essential for understanding brain function.

The Foundation Model Paradigm

Inspired by success in computer vision and natural language processing, neuroimaging is shifting from task-specific models to large-scale foundation models pretrained on massive unlabeled datasets. Previous approaches include:

- Masked Autoencoding (MAE): Reconstruct missing parts of fMRI data. Can overfit to noise in low-SNR brain data.
- Joint-Embedding Predictive Architectures (JEPA): Learn representations by predicting embeddings. Requires careful design to avoid representational collapse.
- Contrastive Learning: Learn by comparing similar vs. different brain patterns.
- Diffusion Models: Learn by gradually denoising corrupted data, capturing the full data distribution.

Three Limitations of Existing Models

1. Restricted Pretraining Distribution: Most models train exclusively on resting-state data, never seeing what the brain looks like during tasks, sleep, or disease.
2. Mismatched Pretraining Tasks: Raw reconstruction overfits to noise; latent-space reconstruction risks representational collapse.
3. Underutilized Metadata: Age, sex, and diagnosis information are ignored during pretraining, preventing the model from separating intrinsic neural dynamics from population variability.

Brain-DiT: The Solution

Brain-DiT addresses all three limitations with a metadata-conditioned Diffusion Transformer pretrained on multi-state fMRI data.

Key Innovation 1: Multi-State Pretraining

Brain-DiT is trained jointly on 349,898 fMRI sessions from 24 datasets spanning five brain states:

- Resting state (9 datasets): HCP, CHCP, ABCD, PIOP1, PIOP2, ISYB, BHRC, SALD, SLIM
- Task-based (3 datasets): HCP Task, CHCP Task, N&N;
- Naturalistic viewing (8 datasets): StudyForrest, Emo Film, NSD, Things, CineBrain, HCP Movie, and more
- Disease (2 datasets): ABIDE (autism), ADNI (Alzheimer's)
- Sleep (1 dataset): Simultaneous EEG-fMRI during sleep

This is the first fMRI foundation model trained on multiple brain states under a single generative objective.

Key Innovation 2: Metadata-Conditioned Pretraining

Brain-DiT integrates demographic and clinical metadata (age, sex, diagnosis) as explicit conditioning signals during pretraining. This allows the model to:

- Separate intrinsic neural dynamics from population-level variability (e.g., distinguishing age-related changes from disease-related changes)
- Generate realistic synthetic fMRI data conditioned on specific subject profiles
- Improve downstream performance by providing contextual information about the subject

Key Innovation 3: Diffusion Transformer (DiT)

Brain-DiT uses a Diffusion Transformer architecture. The key insight: the iterative denoising process naturally provides multi-scale representations. Early denoising steps capture fine-grained local structure; later steps encode abstract global semantics. This hierarchy maps directly to what different downstream tasks need.

How Brain-DiT Works

The Diffusion Process

fMRI data is treated as a 4D signal (time x brain regions). During pretraining, the model learns to denoise corrupted versions of this signal. The forward process adds Gaussian noise progressively; the reverse process (what the model learns) removes noise step by step.

The DiT architecture processes the fMRI data as patches (like Vision Transformer patches), with each patch containing a group of brain regions over a time window. The model uses 16 transformer layers with 1024-dimensional embeddings and 8 attention heads.

Metadata Conditioning

Subject metadata (age, sex, diagnosis) is encoded into a conditioning vector and injected into every transformer layer via Adaptive Layer Normalization (AdaLN). During training, the conditioning is randomly dropped 10% of the time to enable both conditional and unconditional generation.

Multi-Scale Feature Aggregation

For downstream tasks, Brain-DiT doesn't just use the final output. Instead, it extracts features from multiple denoising timesteps and multiple transformer layers, then fuses them using a learnable query-based aggregator. This captures:

- Early timesteps + early layers: Fine-grained local brain structure
- Late timesteps + deep layers: Abstract global brain semantics

The aggregator learns which combination of scales works best for each task.

Training Details

The model was trained on 6 NVIDIA A800 (80GB) GPUs for 150 epochs using AdamW optimizer. The fMRI data was resampled to 2mm isotropic resolution, cropped to 96x96x96 voxels, and parcellated into 1000 brain regions using the Schaefer atlas.

Results: How Well Does It Work?

Generative Fidelity

Brain-DiT can generate realistic synthetic fMRI signals conditioned on specific subject metadata. When generating data for 200 virtual autism subjects, the synthetic functional connectivity patterns closely match real patient data. The metadata-conditioned generation outperforms both unconditional generation and random conditioning.

Downstream Task Performance

Brain-DiT was evaluated on 7 downstream tasks across in-distribution (ID) and out-of-distribution (OOD) datasets:

Task	Brain-DiT	Best Baseline	Improvement
HCP (rest classification)	82.7%	70.0%	+12.7 pp
ABIDE (age prediction)	0.75 MSE	0.98 MSE	23% better
SALD (age prediction)	0.33 MSE	0.68 MSE	51% better
PPMI (Parkinson's)	61.2%	58.7%	+2.5 pp
ADHD-200	62.5%	58.4%	+4.1 pp
NKI-EDU (education)	69.9%	62.4%	+7.5 pp
NKI-AGE (age)	0.38 MSE	0.50 MSE	24% better

Brain-DiT significantly outperforms all baselines (BrainLM, Brain-JEPA, BrainMass) on nearly every task. The improvements are statistically significant ($p < 0.05$).

Why Brain-DiT Works: Analysis

State Diversity Drives Gains

The authors ran an ablation study fixing the pretraining budget at 30,000 sessions and varying only the brain state mixture. Performance improved consistently as state coverage increased:

- Resting only: baseline
- +Disease: small improvement
- +Task: larger improvement
- +Naturalistic: further improvement
- All states: best performance

This suggests the gains come from broader brain-state coverage, not just more data.

Global vs. Local Representations

Different downstream tasks prefer different representational scales:

- ADNI (Alzheimer's classification): Strongest preference for later denoising timesteps (coarse, network-level structure). This makes sense — Alzheimer's causes system-level network disruptions.
- Age prediction: More evenly distributed across timesteps, consistent with age effects spanning both global and local connectivity.
- Gender prediction: Falls between the two.

This confirms that the multi-scale representation from diffusion denoising captures genuinely different types of information at different scales.

Representation Quality

Even with a frozen backbone (no fine-tuning), Brain-DiT outperforms voxel-level baselines. For example, 62.80% on NKI-EDU vs. 60.59% for SlimBrainmae. This means the pretrained representations are genuinely better, not just adapted better during fine-tuning.

Comparison with Existing Models

Model	Training Data	Method	Key Limitation
BrainLM	Resting only	MAE	Limited brain states
Brain-JEPA	Resting only	JEPA	Representational collapse risk
BrainMass	Resting + some task	Contrastive	Limited metadata use
SlimBrainmae	Resting only	MAE	No multi-state training
SlimBrainjepa	Resting only	JEPA	No metadata conditioning
Brain-DiT	349K sessions, 24 datasets, 5 states	Diffusion + Metadata	None (this paper)

The key difference: Brain-DiT is the only model that trains on multiple brain states AND uses metadata conditioning AND uses diffusion-based generative pretraining. All three innovations contribute to its superior performance.

Why Diffusion Beats Reconstruction

Previous models use reconstruction (MAE) or alignment (JEPA) as their pretraining objective. Brain-DiT uses diffusion. The key insight: diffusion models learn the full data distribution, not just how to reconstruct. This means:

- Richer features: The model learns what realistic brain data looks like, not just how to fill in blanks.
- Multi-scale representations: Early denoising = local structure, late denoising = global semantics.
- Conditional generation: Can generate brain data for specific subject profiles.

What This Means for Neuroscience

1. Universal Brain Representations

Brain-DiT shows that a single model can learn representations that work across all major brain states. This is a step toward universal brain encoders that could replace task-specific models.

2. Clinical Applications

The model's ability to distinguish intrinsic neural dynamics from population variability (via metadata conditioning) has direct clinical implications. For example, separating age-related brain changes from Alzheimer's-related changes could improve early diagnosis.

3. Synthetic Brain Data

Brain-DiT can generate realistic synthetic fMRI data conditioned on specific subject profiles. This could help:

- Augment small clinical datasets for training diagnostic models
- Simulate brain activity under conditions that are hard to measure (e.g., specific disease stages)
- Test hypotheses about brain-behavior relationships

4. Transfer Learning

Even with a frozen backbone, Brain-DiT outperforms baselines that were fine-tuned. This means the pretrained representations are genuinely useful for downstream tasks, reducing the need for large labeled datasets.

5. Multi-Scale Brain Analysis

The finding that different tasks prefer different representational scales (global vs. local) provides a new lens for understanding brain function. Alzheimer's affects global network structure; age affects both local and global connectivity.

The 24 Datasets Behind Brain-DiT

Brain-DiT was pretrained on an unprecedented collection of fMRI data spanning every major brain state. Here's the full breakdown:

Brain State	# Datasets	Examples	What It Captures
Resting	9	HCP, CHCP, ABCD, SALD, SLIM	Spontaneous brain activity without tasks
Task-based	3	HCP Task, CHCP Task, N&N	Brain activity during cognitive challenges
Naturalistic	8	StudyForrest, Emo Film, NSD, CineBrain	Brain activity during movies and natural viewing
Disease	2	ABIDE (autism), ADNI (Alzheimer's)	Brain changes caused by neurological conditions
Sleep	1	Simultaneous EEG-fMRI during sleep	Brain activity during different sleep stages

The In-Distribution (ID) set includes 22 datasets used for pretraining. The Out-of-Distribution (OOD) set includes NKI, ADHD-200, and PPMI — datasets the model never saw during pretraining. Strong OOD performance proves the model learned generalizable representations, not just memorized specific datasets.

Metadata Conditioning in Practice

During pretraining, each fMRI session comes with subject metadata: age (continuous), sex (binary), and diagnosis (categorical). This metadata is encoded into a conditioning vector and injected into every transformer layer. The key design choices:

- CFG-style regularization: 10% of the time, the metadata is replaced with an unconditional embedding. This enables both conditional and unconditional generation.
- Missing data handling: Samples with missing metadata are either dropped or treated as unconditional, depending on the missing-label policy.
- AdaLN injection: The conditioning vector is injected via Adaptive Layer Normalization, which modulates the transformer's internal representations based on the subject's profile.

Quick Reference

Key Numbers to Remember:

Metric	Value
Total fMRI Sessions	349,898 from 24 datasets
Brain States Covered	5 (resting, task, naturalistic, disease, sleep)
Architecture	Diffusion Transformer (DiT), 16 layers, 1024-dim
HCP Classification	82.7% (vs. 70.0% best baseline)
ADHD-200 Accuracy	62.5% (vs. 58.4% best baseline)
NKI-EDU Accuracy	69.9% (vs. 62.4% best baseline)
Training Compute	6x NVIDIA A800 GPUs, 150 epochs
Key Innovation	First fMRI model with multi-state + metadata conditioning
Code Available	Yes (public release)

One-Line Summary:

Brain-DiT is the first fMRI foundation model trained on 349K brain scans spanning all major brain states, using diffusion-based generative pretraining with metadata conditioning to learn representations that outperform all existing models on 7 downstream tasks.

Limitations

ROI-Level Analysis: Brain-DiT operates on parcellated ROI time series, not raw voxel data. While this is computationally efficient, it may lose fine-grained spatial information.

Training Compute: Training on 349K sessions requires significant compute (6x A800 GPUs for 150 epochs). This may limit accessibility for some research groups.

Resolution: The model operates at 2mm isotropic resolution with TR=0.72s. Higher-resolution data could reveal additional patterns.

Single Atlas: The primary analysis uses the Schaefer-1000 atlas. Different parcellations might yield different results.

Future Directions

Extending Brain-DiT with richer metadata, broader state coverage, and improved scaling efficiency for longer sequences and higher-resolution inputs. The code and parameters are publicly available for the research community.

Conclusion

Brain-DiT demonstrates that combining multi-state pretraining, metadata conditioning, and diffusion-based generative modeling produces an fMRI foundation model that outperforms all existing baselines on both in-distribution and out-of-distribution tasks. The model's ability to generate realistic synthetic brain data and learn multi-scale representations makes it a versatile tool for both basic neuroscience research and clinical applications. As fMRI datasets continue to grow in size and diversity, foundation models like Brain-DiT will become increasingly important for understanding the human brain.