FUNCTIONAL MRI TIME SERIES GENERATION VIA WAVELET-BASED IMAGE TRANSFORM AND SPECTRAL FLOW MATCHING FOR BRAIN DISORDER IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Functional Magnetic Resonance Imaging (fMRI) provides non-invasive access to dynamic brain activity by measuring blood oxygen level-dependent (BOLD) signals over time. However, the resource-intensive nature of fMRI acquisition limits the availability of high-fidelity samples required for data-driven brain analysis models. While modern generative models can synthesize fMRI data, they often remain challenging in replicating their inherent non-stationarity, intricate spatiotemporal dynamics, and physiological variations of raw BOLD signals. To address these challenges, we propose Dual-Spectral Flow Matching (DSFM), a novel fMRI generative framework that cascades dual frequency representation of BOLD signals with spectral flow matching. Specifically, our framework first converts BOLD signals into a wavelet decomposition map via a discrete wavelet transform (DWT) to capture globalized transient and multi-scale variations, and projects into the discrete cosine transform (DCT) space across brain regions and time to exploit localized energy compaction of low-frequency dominant BOLD coefficients. Subsequently, a spectral flow matching model is trained to generate class-conditioned cosine-frequency representation. The generated samples are reconstructed through inverse DCT and inverse DWT operations to recover physiologically plausible time-domain BOLD signals. This dual-transform approach imposes structured frequency priors and preserves key physiological brain dynamics. Ultimately, we demonstrate the efficacy of our approach through improved downstream fMRI-based brain network classification.

1 INTRODUCTION

Recent advances in deep generative modeling have shown promising capability in synthesizing realistic yet diverse variations of neuroimaging modalities (Yap et al., 2024). Among available modalities, functional MRI (fMRI) signals offer a non-invasive view of neuronal activity, critical for diagnosing neuropsychiatric and neurodevelopmental disorders (Noman et al., 2022; 2024). However, fMRI data collection is costly and yields limited, often imbalanced datasets (Tan et al., 2024a). These shortcomings limit the generalizability of data-driven brain analysis models, ultimately affecting the reliability of computer-aided clinical tools for neurological and psychiatric conditions (Bollmann & Barth, 2021; Ting et al., 2022). To address these challenges, generative models have been explored for fMRI signal synthesis to support data augmentation and downstream applications (Power et al., 2014; Tan et al., 2023).

Most existing approaches generate brain connectivity directly in the functional connectivity (FC) space, where BOLD signal dependencies are summarized by a single correlation matrix (Biswal & Uddin, 2025). For instance, Tan et al. (2024b) proposes a DCGAN that preserves connectomic structure and improves the performance of downstream FC classifiers. Similarly, BrainFC-CGAN jointly trains adversarial and supervised loss components to preserve the subject identity of real FC on synthetic samples (Tan et al., 2024a). However, such FC representations encode static pairwise relations into dyads and do not effectively capture transient network states within human brain networks (Shabestari et al., 2025).

Recent works have revisited time-domain modeling of fMRI as an alternative to correlation-based functional connectivity (FC). Yuan & Qiao (2024) designs diffusion-TS, a denoising diffusion probabilistic model (DDPM) for fMRI time series data generation, showing improved robustness over GANs and (Variational Autoencoder) VAE-based generative models. Hu et al. (2024) proposes FM-TS that accelerates the sampling step yet provides quality synthetic samples via a flow matching framework. While these methods shift focus from traditional FC to time-series data generation, their feasibility and effectiveness for neuroimaging tasks remain largely unexplored. We argue that limiting generative modeling to FC matrices or the raw time series is inadequate to faithfully reproduce the brain's transient state, multiscale oscillations, and cross-frequency interactions due to difficulties in disentangling physiologically driven fluctuations (e.g., cardiac pulsation, respiratory cycles, motion-induced artifacts) (Biswal & Uddin, 2025). In contrast, a time-frequency/scale representation that captures time and spectral BOLD information can fully reproduce the rich spatiotemporal dynamics of BOLD signals. Motivated by T2I-Diff and ImagenTime, both of which frame time-series signals as an image-generation task (Tew et al., 2025; Naiman et al., 2024). T2I-Diff specifically remodeled and validated the feasibility of this time-frequency image-based approach for generating BOLD signals. Crucially, the performance gains were modest due to the fixed-resolution STFT representation, which neglects fine-grained transients and attenuates frequency amplitude modulations, leading to artifacts during the image-to-signal reconstruction (Tew et al., 2025).

To address these issues, in this paper, we propose Dual-Spectral Flow Matching (DSFM), an fMRI generation framework that cascades two spectral transformations of BOLD signals and integrates a spectral flow matching for generative modeling. Our framework first decomposed BOLD signals using the discrete wavelet transform (DWT) to form multiresolution time-scale scalogram images. Subsequently, we compute a discrete cosine transform (DCT) that exploits low-frequency BOLD coefficients. These transforms produce a dual-spectral view in which local and global dynamics are jointly represented. Additionally, our framework introduces a spectral-domain flow matching for efficient and high-fidelity generation of the time-scale fMRI scalograms conditioned on subject classes. The generated time-frequency scalograms are then reverted to BOLD signals via image-to-time series transforms. Our main contributions are summarized as follows:

1. Our proposed DSFM framework is the first to jointly leverage DWT and DCT, forming a unified dual-spectral image transform to capture both global and local spatiotemporal and spectral features for fMRI BOLD signal generation and brain disorder classification.
2. We develop a spectral flow matching to model a heat dissipation process in the DCT domain to achieve efficient, coarse-to-fine generation aligned with the frequency hierarchy of the dual-spectral representation. This enables DSFM to leverage the spectral sparsity inherent in fMRI signals to effectively capture diverse brain profiles.
3. Our results show that DSFM demonstrates strong performance on unconditional and conditional spectral image synthesis, and achieves improvement in brain disorder (MDD) classification compared to recent time-series and fMRI generation baselines.

2 METHODS

2.1 DISCRETE WAVELET TRANSFORM AND ITS INVERSION

Fig. 1 provides an overview of our proposed framework. Given high-dimensional fMRI signals from S subjects, denoted as $\mathcal{X} = \{x_s\}_{s=1}^{S}$, where each subject $x_s \in \mathbb{R}^{D\times T}$ consists of D regions of interest (ROIs) recorded over T time points, our objective is to learn the underlying real data distribution $p_{\text{data}}(\mathcal{X})$ and generate a synthetic distribution $p_\theta(\mathcal{X})$ that is statistically indistinguishable from the real data. Unlike conventional time-series generative tasks that operate exclusively in the time domain, our approach transforms fMRI time series into time-scale images using the DWT, defined as follows:

$$W(k,j) = \sum_{n=1}^{N} x(n)\,\psi_{j,k}[n], \tag{1}$$

where $x_s(n)$ is the BOLD signal at local time index $n \in \{1,2,\ldots,N\}$. Here, $\psi_{j,k}[n] = 2^{-k/2}\psi[2^k n - j]$ is the dyadic wavelet basis function, where scale $k \in \{1,2,\ldots,\lfloor \log_2 N \rfloor\}$ controls the frequency resolution, and translation index $j \in \{1,2,\ldots,N/2^k\}$ determines the time location,

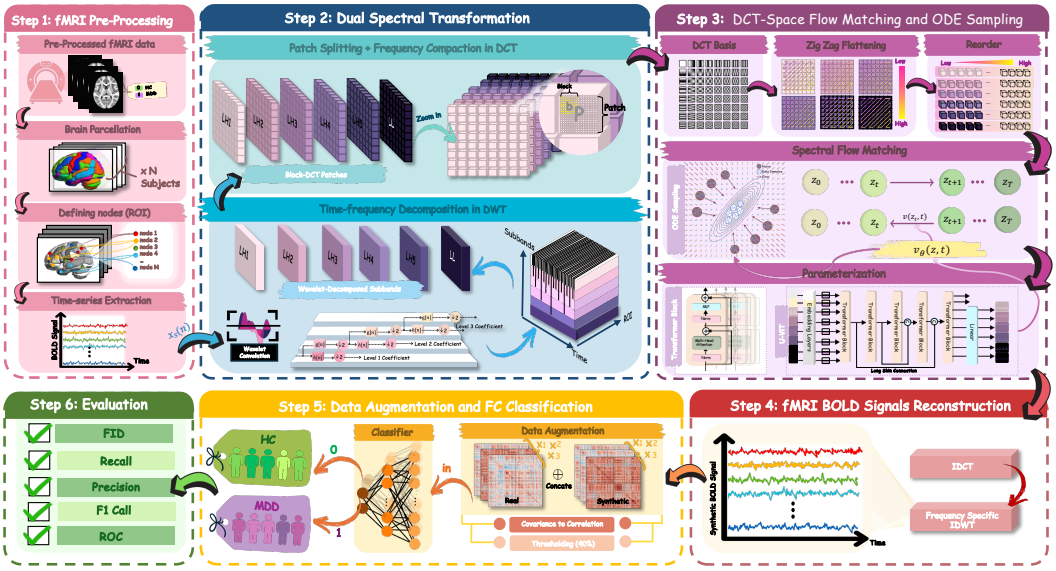

Figure 1: The pipeline of DSFM. ROI-based BOLD time series are first extracted, followed by DWT-based multiresolution decomposition and blockwise 2D DCT for localized spectral encoding. U-ViT is used to model the velocity field in the DCT domain for ODE-based sampling. The reconstructed signals (via IDCT and IDWT) are then used for data augmentation, FC matrix construction, and classification. Finally, fidelity and downstream performance are evaluated.

derived from the mother wavelet $\psi_{j,k}[n]$. To construct a wavelet decomposition map, we upsample each wavelet subband to the original time length and stack them along the scale axis, forming a multiresolution wavelet decomposition map. Thus forming a full wavelet coefficient tensor $W(i,j,k) \in \mathbb{R}^{D\times T_\psi \times C}$, where $T_\psi = N/2^C$ and $C = \lfloor \log_2 N \rfloor$, that captures both low-frequency trends and high-frequency transients in the fMRI BOLD signals. We further perform component-wise normalization to accentuate the difference between high and low coefficients over brain regions and time. As shown in Fig. 2, this allows the time-series signals to be represented as multichannel images with preserved spectral-temporal characteristics.

To reconstruct the original signals from the generated scalogram representation, we first denormalize the predicted wavelet components $\hat{W}^{(i)}(j,k) \in \mathbb{R}^{T\times C}$ of each i^{th} ROI. The coefficients are then downsampled according to their corresponding dyadic scales and computed the inverse DWT (IDWT) to obtain the fMRI BOLD signals as follows:

$$\hat{x}(n) = \frac{1}{N}\sum_{k=1}^{C}\sum_{j=1}^{T} W^{(i)}(j,k)\,\psi_{j,k}[n]. \tag{2}$$

Finally, these wavelet subbands are reconstructed through a hierarchical combination of approximation and detail components across all scales to obtain the reconstructed time-domain signal $\hat{x}_s$ for each subject. This process ensures that the inherited spectral-temporal characteristics of the original fMRI BOLD signals are well-preserved.

2.2 Discrete Cosine Transform for BOLD Signals

To extract localized energy compactions of low-frequency spontaneous BOLD coefficients. We divide each subband map $\hat{W}^{(k)}(i,j) \in \mathbb{R}^{D\times T_\psi}$ of each k^{th} wavelet scale into non-overlapping 2D blocks of size $B \times B$, resulting in a set of blocks (patches):

$$W^{(k)} \equiv \left\{ W_p^{(k)}(x,y) \in \mathbb{R}^{B\times B} \right\}_{p=1}^{P}, \tag{3}$$

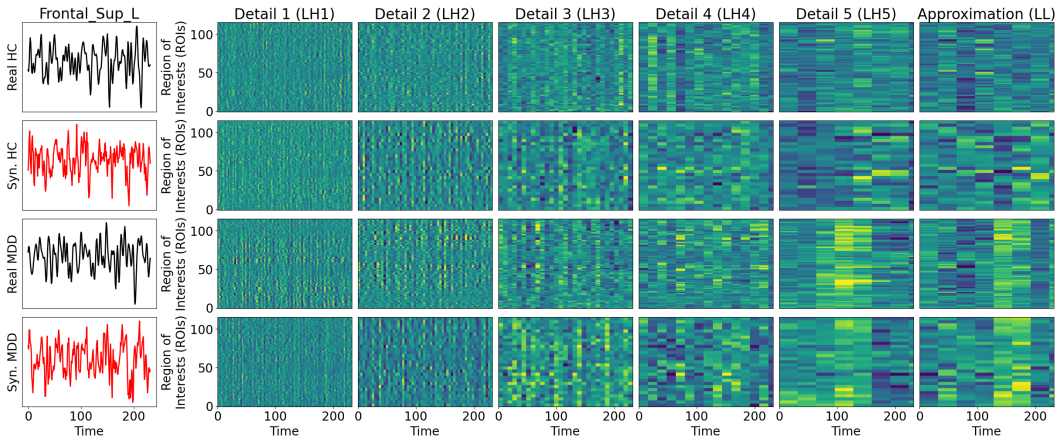

Figure 2: Original (Rows 1&3) vs. synthetic BOLD signals (Rows 2 &4) and generated normalized scalograms. Our framework generates new synthetic BOLD signals as opposed to correlation matrices or functional connectivity, with distributional statistics that closely match the original samples.

where P is the number of blocks per subband image and B is the block size. Each block is then transformed via 2D type-II DCT, as follows:

$$D^{(k)}(u,v) = \alpha(u)\,\alpha(v) \sum_{x=1}^{B} \sum_{y=1}^{B} W^{(k)}(x,y) \cos\left[\frac{(2x+1)u\pi}{2B}\right] \cos\left[\frac{(2y+1)v\pi}{2B}\right], \tag{4}$$

where $\alpha(u) = \sqrt{\frac{1}{B}}$ if $u = 0$, and $\alpha(u) = \sqrt{\frac{2}{B}}$ otherwise. The resulting $D^{(k)}(u,v) \in \mathbb{R}^{B\times B}$ at each scale k represents the DCT coefficients within each block.

To recover the full image representation, we apply the inverse 2D DCT (IDCT) to each block $D^{(k)}(u,v)$. The signal block is reconstructed via the following inverse transform:

$$\hat{W}^{(k)}(x,y) = \sum_{u=1}^{B} \sum_{v=1}^{B} \alpha(u)\,\alpha(v) D^{(k)}(u,v) \cos\left[\frac{(2x+1)u\pi}{2B}\right] \cos\left[\frac{(2y+1)v\pi}{2B}\right]. \tag{5}$$

Once all blocks have been transformed back into the original spatial domain, we stitch the patches to recover the full subband map. Since the DCT is applied to non-overlapping blocks, the reconstruction involves simply tiling the inverse-transformed blocks back into their original positions within the subband image. This blockwise DCT preserves localized low-frequency structure in the ROI–time space while discarding high-frequency noise components. The resulting set of filtered subbands can then be passed back into the IDWT to recover the time-domain fMRI BOLD signal $\hat{x}_s(n)$, ensuring global and local spectral characteristics are retained.

2.3 Spectral Flow Matching in DCT Domain

Recent studies have empirically demonstrated that pixel-based diffusion models exhibit approximate autoregressive behavior in the frequency domain (Dieleman, 2024; Falck et al., 2025). Specifically, diffusion models (Ho et al., 2020; Song et al., 2021) tend to eliminate high-frequency components early in the forward process, followed by progressively lower-frequency components as the diffusion timestep advances. While prior studies focus on the Fourier basis, this property also holds in the DCT domain (Skorokhodov et al., 2025; Ning et al., 2025), which offers practical advantages: real-valued orthogonality, energy compaction in low-frequency bands, and compatibility with block-wise architectures.

Modeling diffusion directly in the frequency domain enables the exploitation of spectral sparsity for designing frequency-aware noise schedules. However, existing frequency-domain generative models (Hoogeboom & Salimans, 2023; Rissanen et al., 2023) remain constrained to the diffusion

framework, which relies on stochastic differential equation (SDE) sampling and typically requires hundreds to thousands of steps for high-quality synthesis. In contrast, flow-matching approaches based on ordinary differential equations (ODEs) provide a deterministic alternative with significantly lower sampling complexity. In this work, we introduce a spectral flow-matching framework that extends frequency-based generative modeling beyond the diffusion paradigm. Our approach enables coarse-to-fine generation aligned with the spectral structure of natural images, while achieving high-fidelity synthesis with significantly fewer sampling steps.

To achieve this, we formulate a spectral flow-matching loss that learns the velocity field of a probability flow directly in the DCT domain. First, consider a forward-time heat dissipation process (Rissanen et al., 2023) as an alternative to the conventional isotropic diffusion, described by the following stochastic partial differential equation (SPDE):

$$dx_t(c) = \eta(t)\,\Delta_c\,x_t(c)\,dt + G(t)\,dW(t), \tag{6}$$

where $x_t : \mathbb{R}^2 \times \mathbb{R}_+ \to \mathbb{R}$ is an idealized continuous-space representation of a single image channel at time $t \in [0, 1]$, and $\Delta_c = \nabla_c \cdot \nabla_c$ is the Laplace operator with respect to the spatial image coordinates c; $\eta(t)$ and $G(t)$ are time-dependent scalar drift and matrix-valued diffusion coefficients, respectively. The corresponding reverse-time probability flow ODE (Song et al., 2021) is given by:

$$\frac{dx_t}{dt} = \eta(t)\,\Delta_c\,x_t(c) - \frac{1}{2}\,G(t)G(t)^T\,\nabla_{x_t}\log p(x_t). \tag{7}$$

Subsequently, define the forward and inverse DCT transforms formally as

$$z_t = V^T x_t = \mathrm{DCT}(x_t), \quad x_t = V z_t = \mathrm{IDCT}(z_t), \tag{8}$$

where V denotes the matrix of orthonormal DCT basis eigenvectors. It then follows that the Laplacian operator in equation 6 can be diagonalized via the eigendecomposition $\Delta_c = V\,\Lambda\,V^T$, where Λ denotes the diagonal matrix of DCT mode-specific Laplacian eigenvalues. Applying DCT to the forward-time SPDE equation 6 yields

$$dz_t = -\eta(t)\,\Lambda\,z_t\,dt + G(t)\,dW(t), \tag{9}$$

where $W(t)$ is a standard Wiener process, but in the DCT domain. Moreover, given that V is orthonormal, the following change-of-variables holds for any differentiable function f:

$$V^T\nabla_x f(x) = V^T\left(\frac{\partial z}{\partial x}\right)^T\nabla_z f(z) = \underbrace{V^TV}_{I}\,\nabla_z f(\underbrace{V^Tx}_{z}) = \nabla_z f(z).$$

By letting $f(x_t) = \log p(x_t)$, the score transforms as $V^T\nabla_{x_t}\log p(x_t) = \nabla_{z_t}\log p(z_t)$. Since Λ is diagonal and the DCT basis orthogonalizes the frequency modes, applying DCT to equation 7, the reverse-time probability flow ODE admits the following mode-wise decomposition:

$$\frac{dz_t[k]}{dt} = -\eta(t)\,\lambda_k\,z_t[k] - \frac{1}{2}\,g(t,k)^2\,\nabla_{z_t[k]}\log p(z_t), \tag{10}$$

where λ_k is the k-th diagonal entry of Λ, corresponding to the Laplacian eigenvalue of the k^{th} DCT basis component, which evolves independently under the ODE dynamics.

The following proposition bridges between this DCT mode-wise probability flow ODE and the conditional velocity (vector) field in flow matching (Lipman et al., 2023).

Proposition 1. *A mode-wise conditional perturbation kernel (isotropic in each DCT mode) is*

$$p(z_t[k] \mid z_0[k]) = \mathcal{N}\big(\mu(t,k)\,z_0[k], \sigma(t,k)^2\big), \tag{11}$$

with mean and standard deviation (std) schedules

$$\mu(t,k) = \alpha(t)\,\omega(t,k), \quad \omega(t,k) = e^{-\lambda_k\tau(t)}, \quad \sigma(t,k)^2 = 1 - \mu(t,k)^2, \tag{12}$$

where $\alpha(t)$ is the mean schedule of a variance-preserving (VP) diffusion process and $\tau(t) = \int_0^t \eta(s)\,ds$, satisfies the heat dissipation process (6) and (9). The mode-wise diffusion coefficients are then given by

$$g(t,k)^2 = 2\,\sigma(t,k)\,\big(\dot\sigma(t,k) - f(t,k)\,\sigma(t,k)\big), \tag{13}$$

where $f(t,k) = \frac{\dot\alpha(t)}{\alpha(t)} - \eta(t)\,\lambda_k$, and $\dot\mu(t,k)$, $\dot\sigma(t,k)$ denote time-derivatives of the mean and std schedules in (12).

Table 1: Comparison of unconditional Netsim dataset generation across SOTA and proposed model.

	CoT-GAN	DiffTime	DiffWave	TimeVAE	TimeGAN	Diffusion-TS	T2I-Diff	DSFM (Ours)
cFID ↓	7.813±.550	0.340±.015	0.244±.018	14.449±.969	0.126±.002	0.105±.006	1.384±.107	0.193±.017
Corr. ↓	26.824±.449	1.501±.048	3.927±.049	17.296±.526	23.502±.039	1.411±.042	4.121±.094	4.552±.041
Disc. ↓	0.492±.018	0.245±.051	0.402±.029	0.476±.044	0.484±.042	0.167±.023	0.400±.059	0.497±.001
Pred. ↓	0.185±.003	0.100±.000	0.101±.000	0.113±.003	0.126±.002	0.099±.000	0.102±.001	0.104±.000

Proof. Refer to the Supplementary Material. □

Proposition 2. *A mode-wise conditional velocity field*

$$\left.\frac{dz_t[k]}{dt}\right|_{z_0[k]} = v(z_t|z_0;t,k) = \dot{\mu}(t,k)\, z_0[k] + \dot{\sigma}(t,k)\,\epsilon, \tag{14}$$

where $\epsilon \sim \mathcal{N}(0,1)$*, is equivalent to the conditional probability flow ODE*

$$\left.\frac{dz_t[k]}{dt}\right|_{z_0[k]} = -\eta(t)\,\lambda_k\, z_t[k] + \frac{1}{2}\, g(t)^2\, \nabla_{z_t[k]} \log p(z_t|z_0). \tag{15}$$

Furthermore, it follows that the marginal velocity field

$$\frac{dz_t[k]}{dt} = v(z_t;t,k) = \mathbb{E}_{p_{data}(z_0|z_t)}\big[v(z_t|z_0;t,k)\,|\,z_t\big], \tag{16}$$

given by the law of the unconscious statistician (Lipman et al., 2024), satisfies the marginal mode-wise probability flow ODE (10).

Proof. Refer to the Supplementary Material. □

Given this correspondence between the probability flow ODE from diffusion models and flow matching, we parameterize the velocity field v_θ using a deep neural network (U-ViT (Bao et al., 2023)) and train it via the following conditional spectral flow matching (CSFM) loss:

$$\mathcal{L}^{\text{CSFM}}(\theta) = \mathbb{E}_{t,p(z_t\,|\,z_0)p_{\text{data}}(z_0)} \big\| v_\theta(z_t;t,k) - v(z_t|z_0;t,k) \big\|^2, \tag{17}$$

where $v(z_t|z_0;t,k)$ is the conditional velocity field in (14), with z_t sampled from the per-mode conditional perturbation kernel (11), and $t \sim \mathcal{U}(0,1)$ is uniformly sampled. Notably, this CSFM loss recovers the standard flow matching loss under the OT-CFM schedules $\mu(t) = 1 - t$ and $\sigma(t) = t$, where the time convention adopted here is the reverse of that in (Lipman et al., 2023). Hence, our framework generalizes flow matching to a heat dissipation process in the DCT domain. In our experiments, we use $\alpha(t)$ from the variance-preserving (VP) cosine schedule and set $\tau(t) = \sigma_{\max} \sin^2\left(\frac{\pi}{2}t\right)$ following (Hoogeboom & Salimans, 2023), which observes optimal performance with $\sigma_{\max} = 20$.

To enable class-conditioned generation, we employ classifier-free guidance (Ho & Salimans, 2021) by conditioning the velocity model on the class label c, i.e., $v_\theta(z_t;t,k,c)$ and set $c = \varnothing$ for the unconditional model. The conditional and unconditional models are trained jointly by randomly replacing the class label c with the null token $\varnothing$ with probability $p_\varnothing$. During sampling, the classifier-free guided velocity is obtained as a weighted combination of the model outputs (Zheng et al., 2023). Finally, DCT samples are generated by numerically integrating the learned flow velocity using an adaptive ODE solver.

3 Experiment

3.1 Settings

Data Acquisition and Pre-processing. We preprocessed the resting-state fMRI (rs-fMRI) dataset from the REST-meta-MDD Consortium database (Yan et al., 2019) using the Data Processing Assistant for Resting-State fMRI (DPARSF) (Yan & Zang, 2010). This dataset comprises 250 Healthy

Table 2: The best MDD dataset generation quality and classification performance of different classifiers trained on ground-truth data augmented at three levels. - refers to FC-based generation *Full results.* Refer to the Supplementary Material.

	W/O Aug.	2D-DCGAN	WGAN-GP	TimeGAN	Diffusion TS	T2I-Diff	DSFM (Ours)
Metric	Real (R.)	R. + Synth. 3×	R. + Synth. 2×	R. + Synth. 1×	R. + Synth. 1×	R. + Synth. 1×	R. + Synth. 1×
cFID	–	–	–	4.98 ± 0.65	2.06 ± 0.21	7.45 ± 0.42	1.51 ± 0.41
Corr.	–	–	–	197.05 ± 17.75	64.16 ± 3.92	62.32 ± 1.04	57.30 ± 2.89
Accuracy	58.90 ± 2.98	58.86 ± 2.24	65.04 ± 2.02	66.78 ± 0.02	67.29 ± 0.02	66.87 ± 3.22	70.84 ± 5.89
Recall	58.90 ± 2.98	58.86 ± 2.24	65.04 ± 2.02	66.78 ± 0.02	67.29 ± 0.02	66.87 ± 3.22	70.84 ± 5.89
Precision	59.56 ± 2.74	59.91 ± 2.57	66.35 ± 2.13	67.14 ± 0.02	67.55 ± 0.02	67.06 ± 3.34	70.99 ± 5.80
F1-Score	58.39 ± 3.09	57.64 ± 1.96	64.12 ± 2.10	66.48 ± 0.02	67.21 ± 0.02	66.83 ± 3.21	70.77 ± 5.97
ROC	59.00 ± 2.56	58.57 ± 2.05	64.74 ± 2.08	67.26 ± 0.03	64.57 ± 0.03	67.26 ± 6.00	71.49 ± 5.73

Table 3: Generation quality and classification accuracy with Schaefer parcellation on ABIDE dataset.

	GAT	GatedGCN	GPS	BrainNetCNN	ContrastPool	BNTF	DSFM (Ours)
Metric	W/O Aug.	W/O Aug.	W/O Aug.	W/O Aug.	W/O Aug. 1×	W/O Aug. 1×	R. + Synth. 1×
cFID	–	–	–	–	–	–	0.07 ± 0.01
Corr.	–	–	–	–	–	–	13.05 ± 1.69
Accuracy	60.10 ± 4.13	61.66 ± 3.36	63.04 ± 3.36	65.75 ± 3.24	65.01 ± 3.84	63.70 ± 4.84	71.30 ± 0.03
Recall	55.11 ± 7.89	55.74 ± 11.58	68.75 ± 11.22	61.25 ± 5.66	61.45 ± 5.43	70.19 ± 8.66	71.40 ± 0.03
Precision	59.57 ± 4.63	61.65 ± 4.12	59.97 ± 5.36	63.98 ± 3.47	63.56 ± 3.62	60.34 ± 5.40	72.60 ± 0.03
F1-Score	56.96 ± 5.16	58.05 ± 8.20	63.48 ± 5.98	62.39 ± 3.13	62.28 ± 2.81	64.64 ± 5.65	71.40 ± 0.03
ROC	60.43 ± 3.88	62.31 ± 4.32	63.34 ± 5.15	64.78 ± 2.52	64.52 ± 2.30	64.15 ± 5.42	71.50 ± 0.07

Controls (HC) subjects and 227 individuals diagnosed with Major Depressive Disorder (MDD). All scans were acquired using a Siemens Tim Trio 3T scanner TR/TE = 2000/30 ms, and a slice thickness of 3mm. The brain was parcellated into 116 ROIs, covering cortical and subcortical areas, and the mean BOLD signal for each ROI was extracted across 232 time points using the Automated Anatomical Labeling (AAL) atlas. The Autism Brain Imaging Data Exchange (ABIDE) initiative provides rs-fMRI data curated from multiple international sites to advance research on Autism Spectrum Disorder (ASD) Di Martino et al. (2014). Our analysis includes 488 ASD patients and 537 normal controls (NC) from the ABIDE database. Lastly, we incorporate the NetSim dataset, a widely used benchmark for evaluating causal discovery algorithms in neuroimaging. NetSim offers biologically realistic simulations of blood-oxygen-level–dependent (BOLD) time series, we chose simulation 4 with 50 features from the original dataset Smith et al. (2011). **Quality Metrics.** Our proposed DSFM model is first assessed in the unconditional setting using standard metrics used by ImagenTime and T2I-Diff (Naiman et al., 2024; Tew et al., 2025), against seven time-series and time-frequency generative model baselines such as CoT-GAN (Xu et al., 2020), DiffTime (Coletta et al., 2023), DiffWave (Kong et al., 2020), TimeVAE (Desai et al., 2021), TimeGAN (Yoon et al., 2019), Diffusion-TS (Yuan & Qiao, 2024), and T2I-Diff (Tew et al., 2025). In the conditional setting, we computed the image-domain FID score on subject-specific DCT and DWT image representations (Heusel et al., 2017). To ensure image-to-signal reconstruction quality, we evaluate the time-domain using the context-FID (cFID) score (Jeha et al., 2022). **Classification Metrics.** We further evaluate the downstream performance with a specially designed classifier for brain connectivity (Kawahara et al., 2017). The baselines include GANs and diffusion models such as Vanilla-GAN (Goodfellow et al., 2020), 1D-DCGAN (Radford et al., 2015), 2D-DCGAN (Tan et al., 2024b), WGAN-GP (Gulrajani et al., 2017), and T2I-Diff (Tew et al., 2025).

3.2 Implementation Details

Connectivity Network Construction. The subject-specific functional connectivity is derived using the Ledoit-Wolf (LDW) regularized shrinkage covariance estimator to preserve the strongest $\tau = 40\%$ connections, resulting in a sparse 116×116 FCs with all other connections set to zero. **DSFM Training.** The proposed DSFM framework generates the fMRI signals corresponding to the subjects' condition (HC and MDD). The classifiers then discriminate between the HC and MDD subjects. We train the DSFM using an AdamW optimizer with a learning rate of $2e^{-4}$ over 300k iterations. All experiments employ a Haar wavelet with a 5-level basis, yielding a real-valued property of 116×232 image size, and we compare numbers of function evaluations (NFE) of 20,50, and

Table 4: Ablation analysis of frequency-specific FC classification by incorporating individual and grouped wavelet subbands.

	Wavelet Subbands						Accuracy		Precision		F1-Score		ROC	
Setting	LH1	LH2	LH3	LH4	LH5	LL	Value	Drop (%)	Value	Drop (%)	Value	Drop (%)	Value	Drop (%)
Full-band	✓	✓	✓	✓	✓	✓	70.84	–	70.99	–	70.77	–	71.49	–
Low-pass	✗	✗	✓	✓	✓	✓	66.89	-5.58	66.96	-5.68	66.77	-5.65	65.79	-7.97
Mid-pass	✓	✓	✗	✗	✓	✓	63.30	-10.64	63.74	-10.21	63.05	-10.91	60.41	-15.50
High-pass	✓	✓	✓	✓	✗	✗	65.40	-7.68	65.55	-7.66	65.18	-7.90	63.66	-11.0
Band-pass 1	✗	✓	✓	✓	✓	✓	66.45	-6.20	66.53	-6.28	66.39	-6.19	68.38	-4.35
Band-pass 2	✓	✓	✓	✓	✓	✗	66.66	-5.90	66.88	-5.79	66.60	-5.89	66.74	-6.64
Band-pass 3	✗	✓	✓	✓	✓	✗	66.88	-5.59	66.76	-5.96	67.06	-5.24	67.77	-5.20

100 steps, and signal-to-noise ratios (SNRs) of 1.0 and 2.0. **Data Augmentation and Classifier Training.** The trained DSFM is used to augment real fMRI signals by factors of $1\times$, $2\times$, and $3\times$. For our classifier, the L2 regularization weight decay is from 10^{-8} to 10^{-2}, the scheduler learning rate reduction factor is from 0.1 to 0.9, and the batch size is from 5 to 16, the same as in (Tan et al., 2022). All hyperparameters are selected based on a 5-fold stratified cross-validation.

3.3 Overall Performance

We first trained our constrained spectral flow matching models to produce similar outputs in Table 1. Then, we followed the standard setting for the quality evaluation of the time-series generation.

Classification Score. To validate the fidelity of the generated samples, we evaluate the classification performance of BrainNetCNN (Kawahara et al., 2017), comparing DSFM to GAN and diffusion-based baselines on our fMRI dataset. Here, we use the parameter setting of NFE = 100 and SNR = 1.0 in subsequent downstream analyses, as supported by the quality metrics of distinguishing HC and MDD subjects in Table 7. Table 3 reports the classification results on the 5-fold cross-validation test set. Notably, DSFM achieves the highest accuracy under a $1\times$ data augmentation setting. Moreover, our model exhibits lower variance across increased augmentation levels, indicating strong generalization and robustness. These results confirm that DSFM not only enriches sample diversity but also preserves discriminative structural and functional patterns critical for clinical tasks. Figure 3 further demonstrates that our proposed DSFM model excels in generating class-conditioned synthetic data whose statistical distribution closely matches that of the original samples.

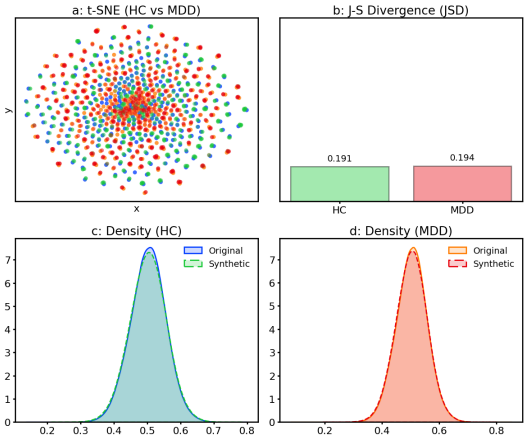

Figure 3: We plot the 2D t-SNE embedding of HC and MDD synthetic data generated with our method (top & bottom). Then, we compare with the Jensen-Shannon Divergence and probability density functions (top right).

3.4 Ablation Studies

We conducted ablation studies on six wavelet detail bands, i.e., LH1: 0.125 - 0.250Hz, LH2: 0.0625 - 0.125Hz, LH3: 0.03125 - 0.0625Hz, LH4: 0.015625 - 0.03125Hz, LH5: 0.007825 - 0.015625Hz, and a coarse approximation LL: 0 - 0.007825Hz, contrasting each setting with the full 0 to 0.25Hz spectrum. Table 4 assesses the impact of different wavelet subbands on model performance. The steepest decline occurred when the mid-frequency LH3–LH4 pair was removed, highlighting the pivotal role of 0.01–0.06 Hz oscillations to capture disease-specific interactions due to insufficient contextual information. Suppressing either the highest (LH1–LH2) or the very lowest components (LL and LH5) produced a comparable, still significant degradation (5–8%), indicating that both rapid fluctuations and slow drifts provide complementary cues. Conversely, removing individual

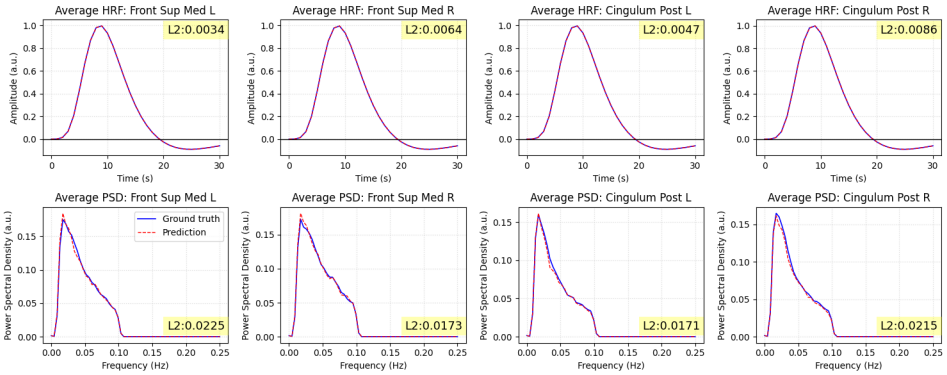

Figure 4: Visualization of the average resting-state hemodynamic response function (rsHRF) and power spectral density (PSD) of real and synthetic BOLD in the Medial Prefrontal Cortex (mPFC) and Posterior Cingulate Cortex (PCC) region of the Default Mode Network (DMN). Highlighted L2 norm quantifies the generation and synthetic results closely resemble the real physiological profiles.

bands such as LH1 and LL also reduced performance by 5–7%, indicating that long-range and slow drifts carry global synchrony patterns essential for classification. Interestingly, we observe that although BOLD fluctuations predominantly lie in the low-frequency band, removing any subbands impaired performance, indicating that disease-related features are distributed across the entire frequency spectrum.

Table 5 presents ablation analyses of different configurations evaluating normalization strategies, block sizes, and wavelet bases influence the generative quality of our dual-spectral representation. In particular, (1,2) shows a comparison of MinMax normalization (MM) with the Entropy-Consistent Scaling (ECS). Notably, MinMax scales each wavelet coefficient independently, broadening the distribution of high-frequency coefficients, which results in slower training and reduced performance. In contrast, ECS preserves the global spectral coefficient by normalizing DCT frequency components using a percentile-trimmed bound derived from the lowest frequency component, providing better cFID and correlation scores by maintaining the original coefficient distribution. Both experiments of smaller and larger block size B in (3,4) achieve comparable generation performance, with a tradeoff of a smaller B will lead to slower training, larger B leads to the loss of fine-grained local dependencies. Finally, ablation of (4,5) with different mother wavelet produces similar generation results on the MDD dataset. This further exemplifies that the underlying fMRI signals do not exhibit strong wavelet-specific bases sensitivity.

Table 5: Ablation of block Size, wavelet basis, and different normalization strategy.

Configurations	cFID↓	Corr↓
1) $B = 4$, MM	1.505±0.41	57.3±2.89
2) $B = 4$, ECS	0.098±0.01	18.2±1.41
3) $B = 2$, Haar	0.121±0.03	19.7±3.03
4) $B = 4$, Haar	0.098±0.01	18.2±1.41
5) $B = 4$, dB-4	0.199±0.10	20.7±3.25

Table 6: Similarity between synthetic and real FC networks across FC edges, node strength, and edge betweenness centrality. Higher values indicate better preservation of real FC topology.

Metric	Vanilla-GAN	1D-DCGAN	2D-DCGAN	WGAN	WGAN-GP	DSFM (Ours)
FC Edges	0.53 ± 0.06	0.10 ± 0.11	0.54 ± 0.49	0.51 ± 0.47	0.52 ± 0.17	0.99 ± 0.00
Node Strength	0.67 ± 0.08	0.30 ± 0.16	0.53 ± 0.08	0.64 ± 0.09	0.62 ± 0.03	0.99 ± 0.00
Edge Betweenness Centrality	0.11 ± 0.02	0.06 ± 0.02	0.14 ± 0.02	0.14 ± 0.02	0.15 ± 0.02	0.77 ± 0.09

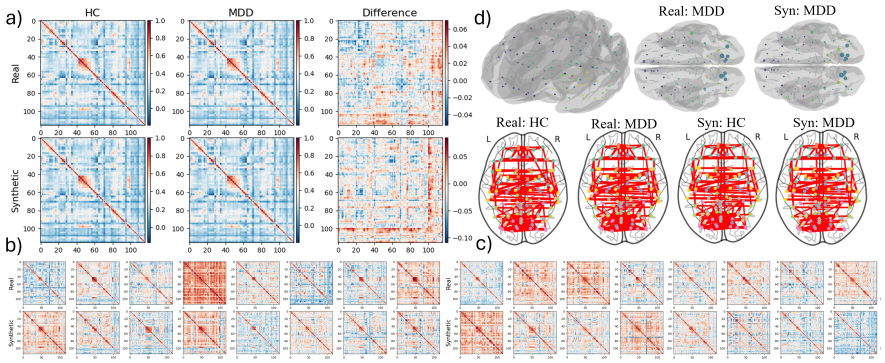

Figure 5: (a) Group-averaged connectivity patterns of real and synthetic HC/MDD connectivity patterns and their differences. (b,c) Subject-level connectivity patterns of real and synthetic from HC and MDD, respectively. (d) 3D cortical surface and brain networks visualizations showing node strength (top) and network organization for real and synthetic HC/MDD (bottom).

4 Neurophysiological Plausibility Analysis

To assess the neurophysiological plausibility of our proposed DSFM-generated BOLD signals, Figure 4 presents the qualitative and quantitative comparisons of resting-state hemodynamic response function (rsHRF) and power spectral density (PSD) between real and synthetic signals in two key hubs of the Default Mode Network (DMN). As shown in Figure 4, the near-perfect overlap of the HRF plots indicates that DSFM preserves the canonical temporal dynamics of the hemodynamic process rather than merely matching marginal statistics. Likewise, the close alignment of the PSD curves indicates that the synthetic samples exhibit meaningful fMRI-like spectral characteristics, accurately capturing the dominant low-frequency peaks and the spectral decay from low-frequency and high-frequency components. The low L2 error across both HRF and PSD provides evidence that DSFM can learn underlying spectral-temporal dynamics of the BOLD signals. Overall, these analyses suggest that our model is able to generate synthetic signals that have the neurophysiological plausibility for different downstream tasks. This is further verified by the classification performance in Table 3, where models trained with DSFM-generated data perform better on unseen samples.

5 Functional Connectivity (FC) Analysis and Visualization

Table 6 further evaluates the fidelity of the generated data FC matrices derived from real and synthetic fMRI BOLD signals. Across all graph similarity metrics, DSFM shows higher Pearson correlation with the real data than other GAN-based models, indicating more realistic synthesis of FC networks in both connectivity edges and network topology. These results demonstrate that DSFM not only reproduces plausible pairwise connectivity patterns but also faithfully captures higher-order network topology, reflecting more coherent interdependencies among FC edges than existing GAN-based generative models. Figure 5 visualizes group-averaged connectivity, thresholded at 0.6 to highlight significant edge connections. Our analysis reveals that the synthetic FC closely aligns with the functional changes observed in the real FC distribution. Furthermore, the HC and MDD connectograms between both real and synthetic FC indicate a reduction in intra-network connectivity within the left superior frontal gyrus (FrontalSupL) and weakened coupling between the left middle frontal gyrus (FrontalMidL) and the anterior cingulate cortex (CingulumAntL). The results suggest impaired cognitive functions associated with difficulties in decision-making and emotion regulation, indicating the biological plausibility of the generated data.

6 Conclusions and Future Work

In this paper, we propose DSFM, which effectively captures both temporal dynamics and spectral evolution underlying the ground-truth data distribution for accurate brain signal generation. For future work, we aim to further validate MDD classification using graph-based deep learning models.

REFERENCES

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22669–22679, June 2023.

Bharat B Biswal and Lucina Q Uddin. The history and future of resting-state functional magnetic resonance imaging. *Nature*, 641(8065):1121–1131, 2025.

Saskia Bollmann and Markus Barth. New acquisition techniques and their prospects for the achievable resolution of fmri. *Progress in neurobiology*, 207:101936, 2021.

Andrea Coletta, Sriram Gopalakrishnan, Daniel Borrajo, and Svitlana Vyetrenko. On the constrained time-series generation problem. *Advances in Neural Information Processing Systems*, 36:61048–61059, 2023.

Abhyuday Desai, Cynthia Freeman, Zuhui Wang, and Ian Beaver. Timevae: A variational auto-encoder for multivariate time series generation. *arXiv preprint arXiv:2111.08095*, 2021.

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.

Sander Dieleman. Diffusion is spectral autoregression, 2024. URL https://sander.ai/2024/09/02/spectral-autoregression.html.

Fabian Falck, Teodora Pandeva, Kiarash Zahirnia, Rachel Lawrence, Richard E. Turner, Edward Meeds, Javier Zazo, and Sushrut Karmalkar. A fourier space perspective on diffusion models, 2025. URL https://arxiv.org/abs/2505.11278.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Emiel Hoogeboom and Tim Salimans. Blurring diffusion models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=OjDkC57x5sz.

Yang Hu, Xiao Wang, Lirong Wu, Huatian Zhang, Stan Z Li, Sheng Wang, and Tianlong Chen. Fm-ts: Flow matching for time series generation. 2024.

Paul Jeha, Michael Bohlke-Schneider, Pedro Mercado, Shubham Kapoor, Rajbir Singh Nirwan, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Psa-gan: Progressive self attention gans for synthetic time series. In *The tenth international conference on learning representations*, 2022.

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

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

Shujian Liao, Hao Ni, Lukasz Szpruch, Magnus Wiese, Marc Sabate-Vidales, and Baoren Xiao. Conditional sig-wasserstein gans for time series generation. *arXiv preprint arXiv:2006.05421*, 2020.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code, 2024. URL https://arxiv.org/abs/2412.06264.

Md Mahmuddun Nabi Murad, Mehmet Aktukmak, and Yasin Yilmaz. Wpmixer: Efficient multi-resolution mixing for long-term time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19581–19588, 2025.

Ilan Naiman, Nimrod Berman, Itai Pemper, Idan Arbiv, Gal Fadlon, and Omri Azencot. Utilizing image transforms and diffusion models for generative modeling of short and long time series. *Advances in Neural Information Processing Systems*, 37:121699–121730, 2024.

Mang Ning, Mingxiao Li, Jianlin Su, Jia Haozhe, Lanmiao Liu, Martin Benes, Wenshuo Chen, Albert Ali Salah, and Itir Onal Ertugrul. DCTdiff: Intriguing properties of image generative modeling in the DCT space. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=vDoAA8xKXL.

Fuad Noman, Sin-Yee Yap, Raphaël C.-W. Phan, Hernando Ombao, and Chee-Ming Ting. Graph autoencoder-based embedded learning in dynamic brain networks for autism spectrum disorder identification. In *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 2891–2895, 2022.

Fuad Noman, Chee-Ming Ting, Hakmook Kang, Raphaël C.-W. Phan, and Hernando Ombao. Graph autoencoders for embedding learning in brain networks and major depressive disorder identification. *IEEE Journal of Biomedical and Health Informatics*, 28(3):1644–1655, 2024.

Jonathan D Power, Anish Mitra, Timothy O Laumann, Abraham Z Snyder, Bradley L Schlaggar, and Steven E Petersen. Methods to detect, characterize, and remove motion artifact in resting state fmri. *neuroimage*, 84:320–341, 2014.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=4PJUBT9f2Ol.

Payam S Shabestari, Harry H Behjat, Dimitri Van De Ville, Christopher R Cederroth, Niklas K Edvall, Adrian Naas, Tobias Kleinjung, and Patrick Neff. Frequency-specific resting-state meg network characteristics of tinnitus patients revealed by graph learning. *bioRxiv*, pp. 2025–03, 2025.

Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=2hEDcA7xy4.

Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for fmri. *Neuroimage*, 54(2):875–891, 2011.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Yee-Fan Tan, Chee-Ming Ting, Fuad Noman, Raphaël C-W Phan, and Hernando Ombao. Graph-regularized manifold-aware conditional wasserstein gan for brain functional connectivity generation. *arXiv preprint arXiv:2212.05316*, 2022.

Yee-Fan Tan, Chee-Ming Ting, Fuad Noman, Raphaël C.-W. Phan, and Hernando Ombao. A unified framework for static and dynamic functional connectivity augmentation for multi-domain brain disorder classification. In *Proceedings of the 2023 IEEE International Conference on Image Processing (ICIP)*, pp. 635–639, 2023.

Yee-Fan Tan, Junn-Yong Loo, Chee-Ming Ting, Fuad Noman, Raphaël C-W Phan, and Hernando Ombao. Brainfc-cgan: A conditional generative adversarial network for brain functional connectivity augmentation and aging synthesis. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1511–1515. IEEE, 2024a.

Yee-Fan Tan, Chee-Ming Ting, Fuad Noman, Raphaël C-W Phan, and Hernando Ombao. fmri functional connectivity augmentation using convolutional generative adversarial networks for brain disorder classification. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2024b.

Hwa Hui Tew, Junn Yong Loo, Yee-Fan Tan, Xinyu Tang, Hernando Ombao, Fuad Noman, Raphaël C.-W. Phan, and Chee-Ming Ting. T2i-diff: fmri signal generation via time-frequency image transform and classifier-free denoising diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 640–650. Springer, 2025.

Chee-Ming Ting, Jeremy I. Skipper, Fuad Noman, Steven L. Small, and Hernando Ombao. Separating stimulus-induced and background components of dynamic functional connectivity in naturalistic fmri. *IEEE Transactions on Medical Imaging*, 41(6):1431–1442, 2022.

Tianlin Xu, Li Kevin Wenliang, Michael Munn, and Beatrice Acciaio. Cot-gan: Generating sequential data via causal optimal transport. *Advances in neural information processing systems*, 33: 8798–8809, 2020.

Chao-Gan Yan, Xiao Chen, Le Li, Francisco Xavier Castellanos, Tong-Jian Bai, Qi-Jing Bo, Jun Cao, Guan-Mao Chen, Ning-Xuan Chen, Wei Chen, et al. Reduced default mode network functional connectivity in patients with recurrent major depressive disorder. *Proceedings of the National Academy of Sciences*, 116(18):9078–9083, 2019.

Chaogan Yan and Yufeng Zang. Dparsf: a matlab toolbox for" pipeline" data analysis of resting-state fmri. *Frontiers in systems neuroscience*, 4:1377, 2010.

Sin-Yee Yap, Junn Yong Loo, Chee-Ming Ting, Fuad Noman, Raphaël C.-W. Phan, Adeel Razi, and David L. Dowe. A deep probabilistic spatiotemporal framework for dynamic graph representation learning with application to brain disorder identification. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 5353–5361, 2024.

Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar. Time-series generative adversarial networks. *Advances in neural information processing systems*, 32, 2019.

Xinyu Yuan and Yan Qiao. Diffusion-ts: Interpretable diffusion for general time series generation. *arXiv preprint arXiv:2403.01742*, 2024.

Qinqing Zheng, Matt Le, Neta Shaul, Yaron Lipman, Aditya Grover, and Ricky T. Q. Chen. Guided flows for generative modeling and decision making, 2023. URL https://arxiv.org/abs/2311.13443.

A Appendix

This appendix provides self-contained additional material for the submission titled *"Functional MRI Time Series Generation via Wavelet-Based Image Transform and Spectral Flow Matching for Brain Disorder Identification"*. It includes a detailed about related works, proofs and derivations, the evaluation metrics, full experimental results, limitations, reproducibility statement, as well as the use of large language models (LLMs).

B Related Works

B.1 Generative Modeling of fMRI Time Series.

Synthesizing fMRI BOLD signals is challenging due to the complex spatiotemporal dependencies, non-stationarity, and interferences arising from physiological fluctuations. Existing time-series generation is principally based on generative adversarial networks (GANs), variational autoencoders (VAEs), and diffusion-based frameworks.

GAN-based approaches: Yoon et al. (2019) proposes TimeGAN by extending GAN framework with an embedding function and a supervised loss to better capture temporal dynamics, successfully preserving both the static and dynamic characteristics of synthetic time-series data. COT-GAN introduces a causality-aware optimal transport cost, further aligning real and synthetic samples over time and reducing time-dependent discrepancy between them (Xu et al., 2020).

VAE-based approaches: TimeVAE incorporates temporal components into its encoder-decoder network, improving the interpretability of generated time series. Furthermore, it demonstrates success in reducing overall training time compared to adversarial methods (Desai et al., 2021).

Diffusion-based approaches: DiffTime improves time-series generation by applying hard constraints to enforce fixed points and global minima; alongside soft constraints introduce penalties to guide the model towards desired temporal trends (Coletta et al., 2023). DiffWave achieves high-fidelity time-series generation by replacing autoregressive dependencies with a diffusion denoising chain (Kong et al., 2020). More recently, ImagenTime and T2I-Diff demonstrated its capability in modelling long-term time-series benchmarks by converting signals into short-time Fourier transform (STFT) as the image representation, offering an alternative for modelling longer continuous signals using spectral components (Naiman et al., 2024; Tew et al., 2025). In contrast, the wavelet transform provides multi-resolution bands by using adaptive windows that narrow at high frequencies and widen at low frequencies. These adaptive methods make the wavelet transform better at capturing short transients in continuous signals while still capturing slower trends (Murad et al., 2025).

C Proofs and Derivations

C.1 Proof of Proposition 1

Proof. The forward-time SPDE (9) in the DCT domain admits the following mode-wise decomposition:

$$dz_t[k] = \eta(t)\,\lambda_k\,z_t[k]\,dt + g(t,k)\,dW_t[k] \tag{18}$$

where $W_t[k]$ is the per-mode standard Wiener process. Subsequently, introduce the variance-preserving (VP) scaling

$$z_t[k] = \alpha(t)\,\tilde{z}_t[k] \tag{19}$$

where $\alpha(t)$ is a scalar applied equally to every mode, and the DCT basis remains unchanged, i.e., the scaled z_t still obeys the heat dissipation SPDE. Substituting this into (18) and applying Itô's lemma gives

$$dz_t[k] = f(t,k)\,z_t[k]\,dt + g(t,k)\,dW_t[k] \tag{20}$$

where we have defined

$$f(t,k) = \frac{\dot{\alpha}(t)}{\alpha(t)} - \eta(t)\,\lambda_k \tag{21}$$

Taking the conditional expectation of the drift term in (20) and integrating with respect to time yields

$$\begin{aligned} \frac{d}{dt}\,\mathbb{E}\big[z_t[k] \mid z_0[k]\big] &= f(t,k)\,\mathbb{E}\big[z_t[k] \mid z_0[k]\big] \\ \mathbb{E}\big[z_t[k] \mid z_0[k]\big] &= \int_0^t f(t,k)\,\mu(t,k)\,dt = \alpha(t)\,e^{-\lambda_k \tau(t)} = \mu(t,k) \end{aligned} \tag{22}$$

which is exactly the mean schedule defined in (12). From (24), we also have

$$\dot{\mu}(t,k) = f(t,k)\,\mu(t,k) \tag{23}$$

which we will use to derive the standard deviation.

Applying Itô's lemma once again to the square of (20), and taking conditional expectations yields

$$\frac{d}{dt}\,\mathbb{E}[z_t[k]^2] = 2\,f(t,k)\,\mathbb{E}\big[z_t[k]^2\big] + g(t,k)^2 \tag{24}$$

Additionally, taking the time-derivative

$$\sigma(t,k)^2 = \mathrm{Var}\big[z_t[k] \mid z_0[k]\big] = \mathbb{E}\big[z_t[k]^2\big] - \mu(t,k)^2 \tag{25}$$

and substituting $\dot{\mu} = f(t,k)\,\mu$ from (23), we have

$$\dot{\sigma}^2 = 2\,f(t,k)\,\sigma^2 + g(t,k)^2 \tag{26}$$

where we use the shorthand notations μ, σ and $\dot{\mu}$, $\dot{\sigma}$ for brevity. Since the conditional perturbation kernel is variance–preserving, we also have

$$\sigma(t,k)^2 = 1 - \mu(t,k)^2 \tag{27}$$

Differentiating this gives

$$\dot{\sigma}^2 = -2\,\mu\,\dot{\mu} = -2\,f(t,k)\,\mu^2 = -2\,f(t,k)\,(1-\sigma^2) \tag{28}$$

Equating (26) and (28) gives

$$g(t,k)^2 = 2\,\sigma(t,k)\,\big(\dot{\sigma}(t,k) - f(t,k)\,\sigma(t,k)\big) \tag{29}$$

which is exactly (13). This completes the proof. □

C.2 Proof of Proposition 2

Proof. The Gaussian reparameterization trick

$$z_t[k]|_{z_0[k]} = \mu(t,k)\,z_0[k] + \sigma(t,k)\,\epsilon \tag{30}$$

follows from the mode-wise conditional perturbation kernel (11) , and its time-derivative gives the conditional vector field (14). Using the results (21), (23) and (29) from the proof of Proposition 1, and substituting (30), we can reformulate the conditional vector field (14) as follows:

$$\begin{aligned} \left.\frac{dz_t[k]}{dt}\right|_{z_0[k]} &= v(z_t \mid z_0; t, k) \\ &= \dot{\mu}\,z_0[k] + \dot{\sigma}\,\epsilon \\ &= \frac{\dot{\mu}}{\mu}\,\big(z_t[k] - \sigma\,\epsilon\big) + \dot{\sigma}\,\epsilon \\ &= f(t,k)\,\big(z_t[k]|_{z_0[k]} - \sigma\,\epsilon\big) + \dot{\sigma}\,\epsilon \\ &= f(t,k)\,z_t[k]|_{z_0[k]} + \big(\dot{\sigma} - f(t,k)\,\sigma\big)\,\epsilon \\ &= f(t,k)\,z_t[k]|_{z_0[k]} + \frac{1}{2}\,g(t,k)^2\,\frac{\epsilon}{\sigma} \\ &= f(t,k)\,z_t[k]|_{z_0[k]} + \frac{1}{2}\,g(t,k)^2\,\nabla_{z_t[k]} \log p(z_t[k] \mid z_0[k]) \end{aligned} \tag{31}$$

which arrives at the conditional probability flow ODE (15). Here, we again use the shorthand notations for brevity.

Applying the law of the unconscious statistician from (16)

$$\mathbb{E}_{p_{\text{data}}(z_0|z_t)}\big[v(z_t|z_0;t,k)\,|\,z_t\big] \tag{32}$$

to the score $\nabla_{z_t}\log p(z_t\mid z_0)$, we have

$$\begin{aligned}
&\int_{\mathbb{R}} \nabla_{z_t}\log p(z_t\mid z_0)\, p_{\text{data}}(z_0|z_t)\; dz_0 \\
&= \int_{\mathbb{R}} \nabla_{z_t}\log p(z_t\mid z_0)\, \frac{p(z_t\mid z_0)\, p_{\text{data}}(z_0)}{\int_{\mathbb{R}} p(z_t\mid z_0)\, p_{\text{data}}(z_0)\; dz_0}\; dz_0 \\
&= \int_{\mathbb{R}} \frac{\nabla_{z_t} p(z_t\mid z_0)}{p(z_t\mid z_0)}\, \frac{p(z_t\mid z_0)\, p_{\text{data}}(z_0)}{p(z_t)}\; dz_0 \\
&= \frac{1}{p(z_t)}\nabla_{z_t}\int_{\mathbb{R}} p(z_t\mid z_0)\, p_{\text{data}}(z_0)\; dz_0 \\
&= \frac{1}{p(z_t)}\nabla_{z_t} p(z_t) = \nabla_{z_t}\log p(z_t)
\end{aligned} \tag{33}$$

where we have repeatedly apply the log-derivative trick $\frac{1}{p(z)}\nabla p(z) = \nabla \log p(z)$. This gives us the marginal score and the same applies to the drift term $f(t,k)\; z_t[k]|_{z_0[k]}$ in (31), thus completing the proof. □

D Evaluation Protocol

D.1 Time-Series Metrics.

We extend the standardized time-series generation metrics from Naiman et al. (2024) to broaden their applicability. We employ the following four metrics and provide their mathematical formulations to ensure comparable evaluation across multiple aspects:

Discriminative (Disc.) & Predictive score (Pred.). We adopt the same experimental setup of (Yoon et al., 2019) for both the discriminative and predictive scores. Both the classifier and sequence-prediction model use a two-layer GRU-based architecture. The discriminative score is computed as $|\text{accuracy} - 0.5|$, where lower scores indicate better indistinguishability, and higher scores reflect greater divergence. The predictive score is the mean absolute error (MAE) of the one-step-ahead predictions and the ground-truth values.

Context-FID score (cFID). Context-FID score is a time-series adaptation of the image-based Frechet Inception Distance (FID) that measures how close in distribution synthetic data is to the real data in a learned embedding space (Jeha et al., 2022). Instead of image features, it uses a trained encoder called TS2Vec to capture temporal context. Lower scores indicate higher fidelity and have been shown to correlate with better downstream tasks.

Correlational score (Corr.). Following (Liao et al., 2020), we first estimate the covariance of the *i*th and *j*th feature of time series as follows:

$$\text{cov}_{i,j} = \frac{1}{T}\sum_{t=1}^{T} X_i^t X_j^t - \left(\frac{1}{T}\sum_{t=1}^{T} X_i^t\right)\left(\frac{1}{T}\sum_{t=1}^{T} X_j^t\right) \tag{34}$$

Then, the correlation score is defined as the average absolute difference between corresponding pairwise correlations in the real and synthetic data:

$$\text{Corr} = \frac{1}{10}\sum_{i,j}\left|\frac{\text{cov}^r_{i,j}}{\sqrt{\text{cov}^r_{i,i}\,\text{cov}^r_{j,j}}} - \frac{\text{cov}^s_{i,j}}{\sqrt{\text{cov}^s_{i,i}\,\text{cov}^s_{j,j}}}\right| \tag{35}$$

D.2 Classification Metrics.

We quantify classification performance using accuracy, precision, recall, F1-score, and the area under the ROC curve, with larger values indicating better performance; their definitions are given in

equation 36-equation 40.

$$\text{Accuracy (ACC)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{36}$$

$$\text{Precision (PRE)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{37}$$

$$\text{Recall (REC)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{38}$$

$$\text{F1-score} = 2 \cdot \frac{\text{PRE} \times \text{REC}}{\text{PRE} + \text{REC}} \tag{39}$$

$$\text{ROC} = \int_0^1 \text{TPR}(\tau)\, d\big(\text{FPR}(\tau)\big) \tag{40}$$

E Full Experimental Results

Table 7: Evaluation of our proposed DSFM with class-conditional (HC vs MDD) generation under varying NFE and SNR.

NFE	NS	Discrete Cosine Transform (DCT)		Discrete Wavelet Transform (DWT)		Signal Transform (Time)	
		HC ↓	MDD ↓	HC ↓	MDD ↓	HC ↓	MDD ↓
20	SNR 1.0	5.303±1.890	5.352±2.126	4.552±0.150	4.569±0.203	1.703±0.302	1.428±0.785
	SNR 2.0	5.735±1.120	5.296±1.583	4.632±0.156	4.623±0.200	1.441±0.859	1.702±1.036
50	SNR 1.0	5.481±1.892	5.580±1.528	4.567±0.303	4.624±0.100	1.327±0.208	1.312±0.186
	SNR 2.0	5.637±2.131	5.811±1.702	4.574±0.295	4.626±0.095	1.378±0.521	1.603±0.538
100	SNR 1.0	5.079±1.507	5.386±1.448	4.520±0.191	4.569±0.134	1.237±0.519	1.255±0.471
	SNR 2.0	5.502±1.975	5.701±2.058	4.463±0.221	4.860±0.116	1.913±0.462	1.428±0.919

E.1 fMRI Signal Generation Quality.

Table 7 compares the generative fidelity of our DSFM framework across three representations: frequency (DCT), time-scale (DWT), and the raw time-series domains. Overall, DSFM demonstrates competitive performance in the DWT domain by achieving the lowest FID across HC and MDD subjects with hyperparameter settings of NFE = 100, SNR = 1.0, indicating precise reconstruction of scale-specific BOLD dynamics. Consistently low cFID values in the time domain further confirm that the synthetic signals remain well aligned with in-distribution temporal patterns, outlining that the model is complementary with additional spectral features. By raising the noise level to 2.0, we observe increased variances and occasionally worsened FID scores, reporting that higher diffusion noise scales impede fine-grained generation quality. In contrast, we also observe that increasing the number of NFE from 20 to 100 consistently reduces error across subjects. These results validate DSFM as an effective time-series-to-image framework for synthesizing biologically plausible, frequency-aligned fMRI signals across representations.

E.2 Full Results of Classification Performance.

Table 8 presents the complete MDD classification results across three augmentation levels. The performance gains at each level indicate that the synthesized FCs accurately capture brain connectivity patterns and that the data augmentation strategy significantly improves classifier generalizability to unseen samples.

F Additional Visualization

F.1 Spectral Image Transformations

Figure 6 illustrates the forward and inverse processes of ImagenTime/T2I-Diff and DSFM applied to our proposed fMRI signals. The top row shows an univariate (Short-Time Fourier Transform) STFT

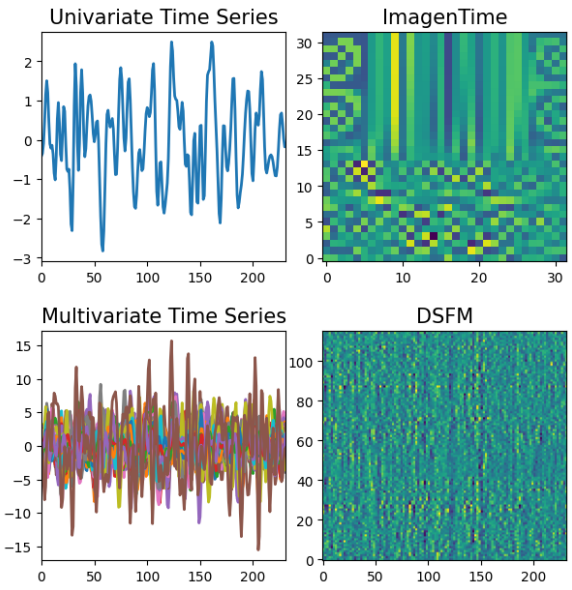

Figure 6: Comparison of univariate and multivariate spectral representations: ImagenTime/T2I-Diff and our proposed DSFM.

Method	Train Set	Accuracy	Recall	Precision	F1-Score	ROC
W/O Augmentation	Real	58.90 ± 2.98	58.90 ± 2.98	59.56 ± 2.74	58.39 ± 3.09	59.00 ± 2.56
Vanila-GAN	Real + Synth 1×	56.90 ± 1.66	56.90 ± 1.66	56.40 ± 2.86	53.68 ± 3.31	56.29 ± 1.92
	Real + Synth 2×	50.71 ± 3.69	50.71 ± 3.69	48.60 ± 7.23	46.74 ± 5.18	50.81 ± 4.13
	Real + Synth 3×	58.86 ± 2.24	58.86 ± 2.24	59.91 ± 2.57	57.64 ± 1.96	58.57 ± 2.05
1D-DCGAN	Real + Synth 1×	62.94 ± 2.01	62.94 ± 2.01	63.43 ± 2.20	62.23 ± 2.68	62.71 ± 2.26
	Real + Synth 2×	65.04 ± 2.02	65.04 ± 2.02	66.35 ± 2.13	64.12 ± 2.10	64.74 ± 2.08
	Real + Synth 3×	58.21 ± 2.98	58.21 ± 2.98	55.70 ± 6.58	52.86 ± 4.14	57.38 ± 3.11
2D-DCGAN	Real + Synth 1×	60.78 ± 4.98	60.78 ± 4.98	61.30 ± 5.34	60.01 ± 5.00	60.33 ± 5.03
	Real + Synth 2×	61.41 ± 2.59	61.41 ± 2.59	61.99 ± 3.73	62.18 ± 3.29	61.04 ± 2.79
	Real + Synth 3×	62.88 ± 4.99	62.88 ± 4.99	63.12 ± 5.02	62.48 ± 5.25	62.67 ± 5.15
WGAN	Real + Synth 1×	64.98 ± 5.54	64.98 ± 5.54	65.19 ± 5.34	64.86 ± 5.61	64.95 ± 5.39
	Real + Synth 2×	60.59 ± 1.81	60.59 ± 1.81	60.89 ± 1.96	60.35 ± 1.78	60.53 ± 1.84
	Real + Synth 3×	61.83 ± 3.03	61.83 ± 3.03	62.27 ± 3.29	61.44 ± 2.70	61.58 ± 2.73
WGAN-GP	Real + Synth 1×	66.02 ± 4.25	66.02 ± 4.25	66.22 ± 4.24	65.93 ± 4.20	65.95 ± 4.13
	Real + Synth 2×	64.76 ± 4.25	64.76 ± 4.25	65.67 ± 4.08	64.23 ± 4.52	64.73 ± 4.14
	Real + Synth 3×	64.56 ± 3.18	64.56 ± 3.18	64.78 ± 3.17	64.38 ± 3.15	64.41 ± 3.08
T2I-Diff	Real + Synth 1×	66.87 ± 3.22	66.87 ± 3.22	67.06 ± 3.34	66.83 ± 3.21	67.26 ± 6.00
	Real + Synth 2×	65.41 ± 2.37	65.41 ± 2.37	66.30 ± 1.67	64.73 ± 2.80	65.75 ± 3.22
	Real + Synth 3×	66.03 ± 1.75	66.03 ± 1.75	66.50 ± 1.32	65.85 ± 1.82	66.58 ± 5.33
DSFM (Ours)	Real + Synth 1×	70.84 ± 5.89	70.84 ± 5.89	70.99 ± 5.80	70.77 ± 5.97	71.49 ± 5.73
	Real + Synth 2×	69.58 ± 3.89	69.58 ± 3.89	69.75 ± 3.72	69.43 ± 3.86	69.91 ± 4.23
	Real + Synth 3×	69.80 ± 3.13	69.80 ± 3.13	69.61 ± 3.02	69.80 ± 3.13	69.00 ± 4.29

Table 8: Classification performance of different classifiers trained on the ground-truth data and an increasing amount of augmented time series data using our proposed model.

spectrogram, and the bottom row presents a multivariate DWT coefficient map. Our framework directly transforms multivariate BOLD signals into a single image representation.

F.2 Frequency-Specific FC Analysis

Figure 7 compares the HC and MDD FC matrices against the ground-truth data correlation across different wavelet subbands. Consistent with the full-band correlation, removing the highest-

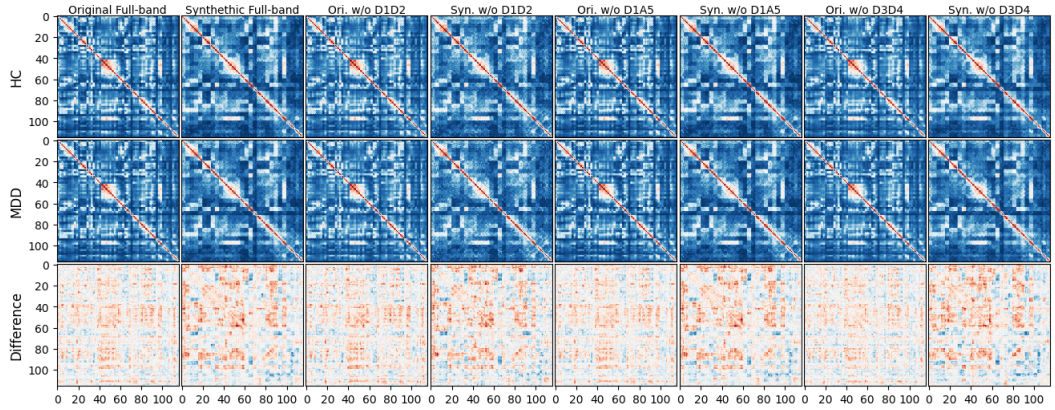

Figure 7: Frequency-specific functional connectivity (FC) matrices for healthy controls (HC) and patients with major depressive disorder (MDD), alongside their differences. The FCs are shown under four different conditions: full-band; removal of the highest-frequency subbands (D1 + D2) and the lowest-frequency component (A5), which both yield the two highest classification scores; and removal of the mid-band subbands (D3 + D4), which produces the greatest deviations and the lowest score.

frequency subbands (D1 and D2), or combining D1 with the lowest band (A5) preserves dense edge connections near the main diagonal. In contrast, removing the mid-frequency subbands (D3 and D4) results in sparser connectivity, particularly in the lower-right region of the matrices.

G Computational cost

The training required 22 hours, 40 minutes, and 52.698 seconds of wall-clock time, while inference for generating the full samples took 48 minutes and 48.98 seconds with 1x A100 GPU. The model contains 130,844,352 parameters.

H Limitations

Currently, DSFM is specially designed for the generation of resting-state fMRI signals. This opens a valuable opportunity to expand our work to other human brain activity signals, such as electroencephalography (EEG), functional near-infrared spectroscopy (fNIRS), and magnetoencephalography (MEG). Our spectral flow matching framework offers flexibility to capture spectral-temporal dynamics of other neural signals with frequency-specific representation.

I Reproducibility statement

We provide the datasets, source code, and configurations for all key experiments, including instructions on how to preprocess data and train the models at https://anonymous.4open.science/r/DSFM-123C.

J The Use of Large Language Models (LLMs)

We used LLMs solely for grammar correction. All ideas, analyses, and results are by the authors.