WORKSHOP ON BRAIN-MACHINE INTERFACES
IEEE SMC 2026
October 4-7, 2026

IEEE SMC 2026
Bellevue, WA USA



Introducing Spatiotemporal Topological Modeling Aids in Visual Decoding

Hongyi Liu, Xvxvan Xie, Yuchen Guo and Youyong Kong

School of Artificial Intelligence
Southeast University

bci
Kleve

Motivation

The Problem

- Reconstructions from fMRI lack clarity and fine-grained features; generalization is limited by scarce paired data
- Local voxel/patch-wise modeling misses global functional dependencies across distant brain regions
- Samples treated as independent (i.i.d.) — temporal context shared across continuous acquisitions is ignored

Proposed Solution

- Treat fMRI signals as time series; sliding-window averaging yields robust patch sequences
- Functional topology graph: patches as nodes, thresholded Pearson correlations as edges
- GNN encoder extracts spatiotemporal graph features, fused with MAE features to guide an LDM

Background & Related Work

Shen et al. 2019; Han et al. 2019	End-to-end fMRI-to-image reconstruction with VAE / GAN generative frameworks
Rombach et al. 2021	Latent diffusion models: denoising in latent space reduces computation while enabling high-resolution image synthesis
Chen et al. 2023 (CVPR)	Sparse masked modeling of fMRI with a conditional diffusion model (Mind-Vis); strong semantic decoding, but limited global field of view
Du et al. 2018	Full covariance matrix models inter-voxel correlations, but overlooks cross-regional collaborative patterns
Ye et al. 2022; Tan et al. 2024	Graph structures model ROI-level functional connectivity; predefined atlases limit generalizability across datasets

Gap: explicit spatiotemporal functional-connectivity modeling without predefined ROIs remains unexplored for visual decoding

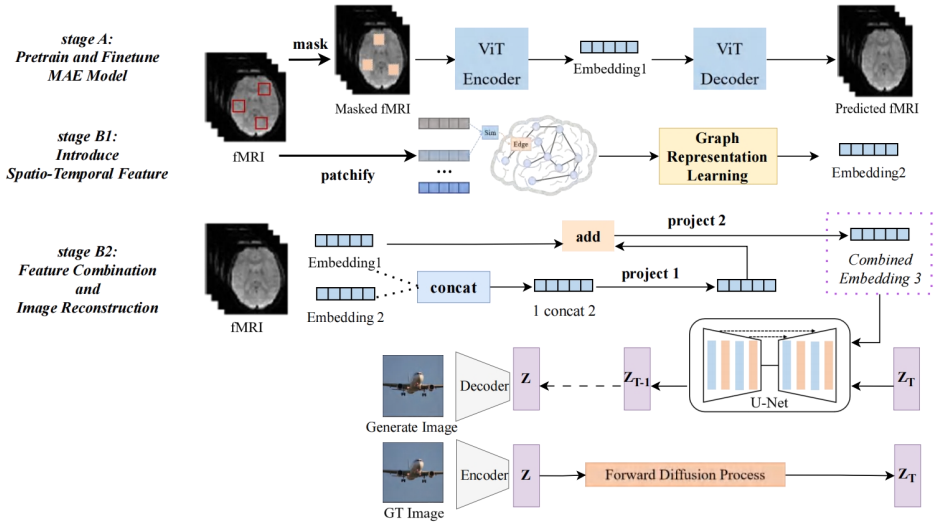
Method Overview

Two-Branch Feature Learning

- Stage A: patchify fMRI + sparse masking; pretrain & finetune a ViT-based MAE encoder
- Stage B1: sliding-window averaging across samples builds robust spatio-temporal patch features

Graph-Guided Reconstruction

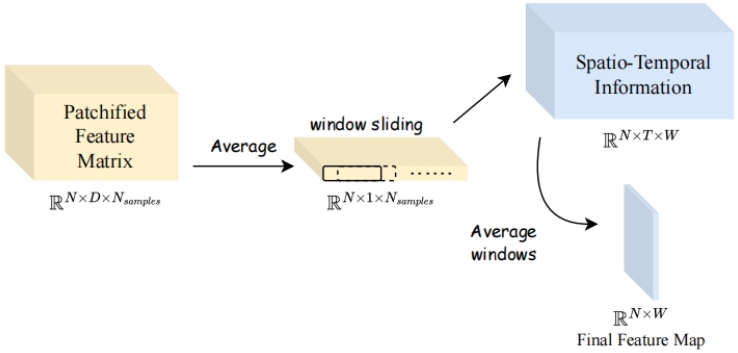
- Stage B1: thresholded Pearson correlations define a functional topology graph, encoded by GNN
- Stage B2: graph embedding fused with MAE features as conditional encoding guiding a pretrained LDM



Stage A & B1: Feature Extraction

Stage A — MAE Pretraining

- Flatten fMRI to a 1-D voxel sequence, zero-pad, patchify via 1-D convolution
- Random sparse masking (ratio 0.75); ViT encoder–decoder reconstructs masked patches



Stage B1 — Spatio-Temporal Graph

- **No predefined ROIs — each local patch is a micro-functional node**
- Window of fixed size W slides over samples; averaging yields $F \in R^{N\times W}$
- Feature dimension = W regardless of patch size → high generalizability
- Pearson correlation matrix + threshold τ → adjacency matrix A
- GNN graph convolution extracts the topological embedding T_{final}

Stage B2: Feature Combination and Image Reconstruction

Fuse the MAE representation with the spatio-temporal graph embedding, then use the fused code to condition the latent diffusion model.

1. Align features

$$\mathbf{C}_{\mathbf{cond}} = f_2\{\mathbf{Z}_{\mathbf{final}} + f_1[\mathrm{concat}(\mathbf{Z}_{\mathbf{final}}, \mathbf{T}_{\mathbf{final}})]\}$$

Project the graph features onto the same dimension as the MAE representation, then add them to the sequence features to preserve as much local information as possible

2. Build condition

$$\phi_i(x_t) = \phi_i(x_t, \sigma(\mathbf{C}_{\mathbf{cond}}))$$

This is then remapped to the condition space required by LDM to obtain the final condition encoding

3. Reconstruct in latent space

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}t}}\ \mathrm{UNet}(x_t, t, \mathbf{C}_{\mathbf{cond}}))$$

$$\mathrm{Image} = \mathscr{D}(x_0^{denoised})$$

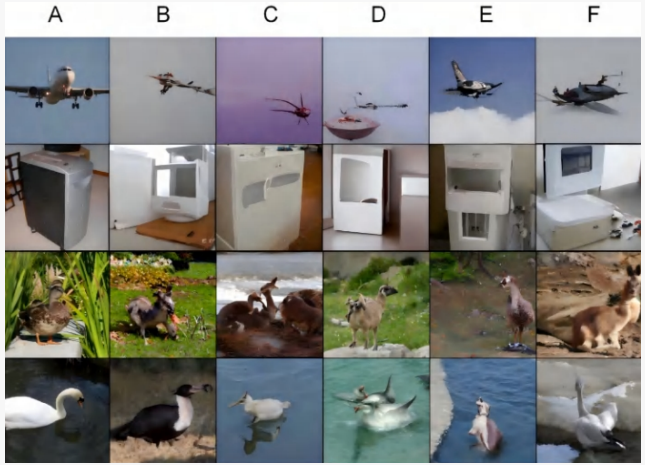
Inject condition encoding at each denoising step, then decode the final latent variable to generate the reconstructed image

Experimental Setup

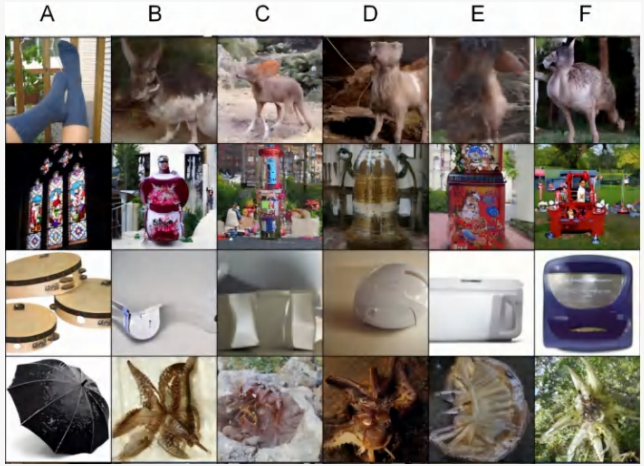
Datasets	GOD (1,250 pairs, 200 categories) & BOLD5000 (5,254 pairs)
fMRI Encoder	1024-dim embedding, 24 ViT encoder layers
Masking Ratio	0.75
Training	500-epoch pretraining + end-to-end finetuning with pretrained LDM
Generation	256×256 images, PLMS stride 250, 5 generations averaged
Graph Construction	Correlation threshold $\tau = 0.75$
Evaluation Metrics	PCC、SSIM、Top-1 Accuracy (ViT-H-14 classifier)

Reconstruction Results

Reconstruction results (Fig. 3)



Failure cases (Fig. 4)



Column A: original images; columns B–F: five reconstruction rounds.
Failures concentrate on complex colors, occlusions and blurry textures.

Quantitative Results

25.90

Ours — Top-1 Acc (max)

PCC 0.570 · SSIM 0.543

23.87

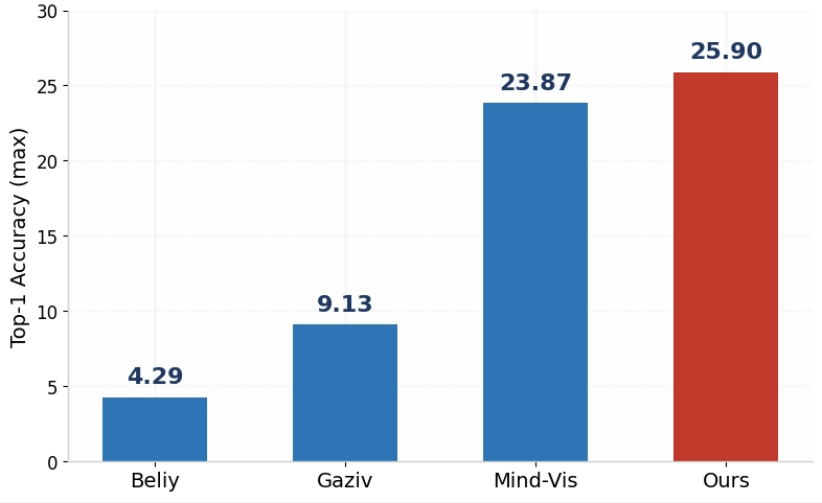
Mind-Vis

PCC 0.564 · SSIM 0.534

+2.03

Improvement

Best semantic accuracy



Ablation Study

Best Configuration	Emb 1024 · Mask 0.75 · Patch 16 → Top-1 Acc 25.90
Patch Size	32 → 20.1 · 48 → 19.8 (patch 16 best)
Masking Ratio	0.65 → 21.7 · 0.85 → 16.1 (0.75 best)
Temporal Modeling	w/o temporal info → 22.4 vs 25.9 (with)

Findings

- Balanced hyperparameters are essential
- spatio-temporal topological information brings a clear gain (+3.5 Top-1)

Conclusion & Discussion

Objective:

Proposed a functional topology graph-based framework: patch-level spatiotemporal graphs + GNN features fused with MAE features to guide LDM reconstruction
Outperforms baselines on GOD and BOLD5000 (Top-1 Acc 25.90 vs 23.87 of Mind-Vis), with clearer, semantically faithful reconstruction

Future Goals:

Improve robustness on complex color combinations, occlusions and blurry textures
Reduce color distortion and spatial confusion to further raise fidelity
Account for noise from imagery-related brain activity during visual perception
Validate on more datasets and extend the framework

Thanks for Listening