Bridging Frequency Gaps in fMRI-to-Image Reconstruction with a Learnable Fourier Adaptive Filter

Junliang Ye [1] Lei Wang [2,3] Zakir Hossain [3,4]

[1]Australian National University [2] Griffith University [3] Data61/CSIRO [4] Curtin University

Motivation

- Long-standing challenge: how the human brain encodes and reconstructs visual experiences
- Recent fMRI-based methods use powerful generative models (VAEs, GANs, diffusion) to map neural activity into image latent variable.
- **Gap:** These approaches largely ignore the rich **frequency structure** of visual input
  - **Low frequencies** → global layout & semantics
  - **High frequencies** → textures & fine details
- **Our solution:** a learnable **Fourier Adaptive Filter (FAF)**
  - Dynamically modulates feature frequencies during reconstruction
  - Enhances both coarse structure and fine-grained detail

Datasets

Natural Scenes Dataset (NSD): 7 T fMRI study, participants viewed COCO images under a continuous recognition task

Subjects: 4 selected (sub1, sub2, sub5, sub7) who completed the full stimulus set

Training set:

- 8,859 unique COCO images
- 24,980 single-trial beta estimates (≤3 repeats per image, then averaged)

Test set: 982 images, 2,770 single-trial betas

Preprocessing: volumetric data masked with 1.8 mm NSD General ROI (early → higher-order visual cortex)

- Voxel counts per subject: sub1 15,724; sub2 14,278; sub5 13,039; sub7 12,682

Method

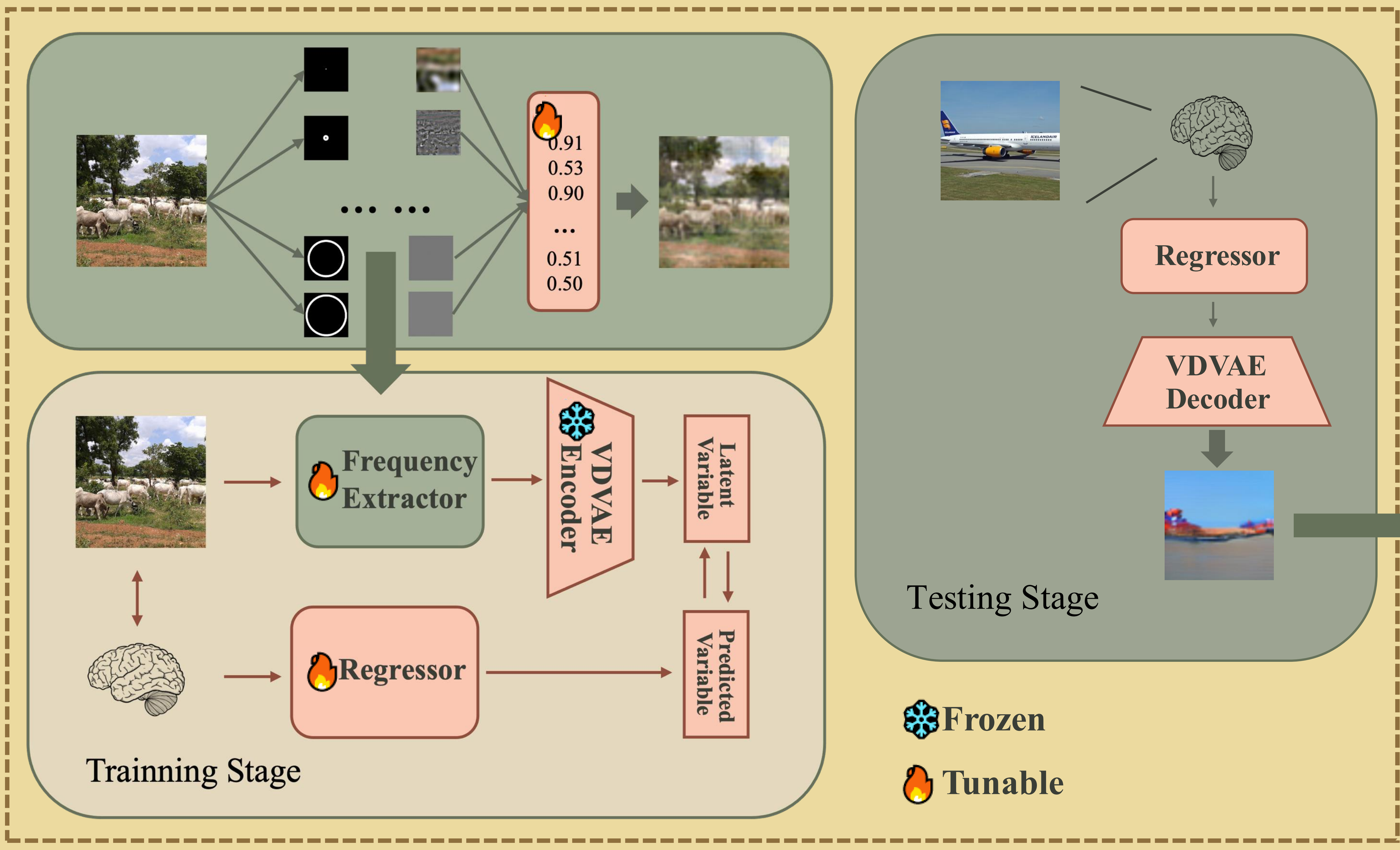


Figure 1: Adaptive Fourier–VDVAE fMRI-to-Image Reconstruction

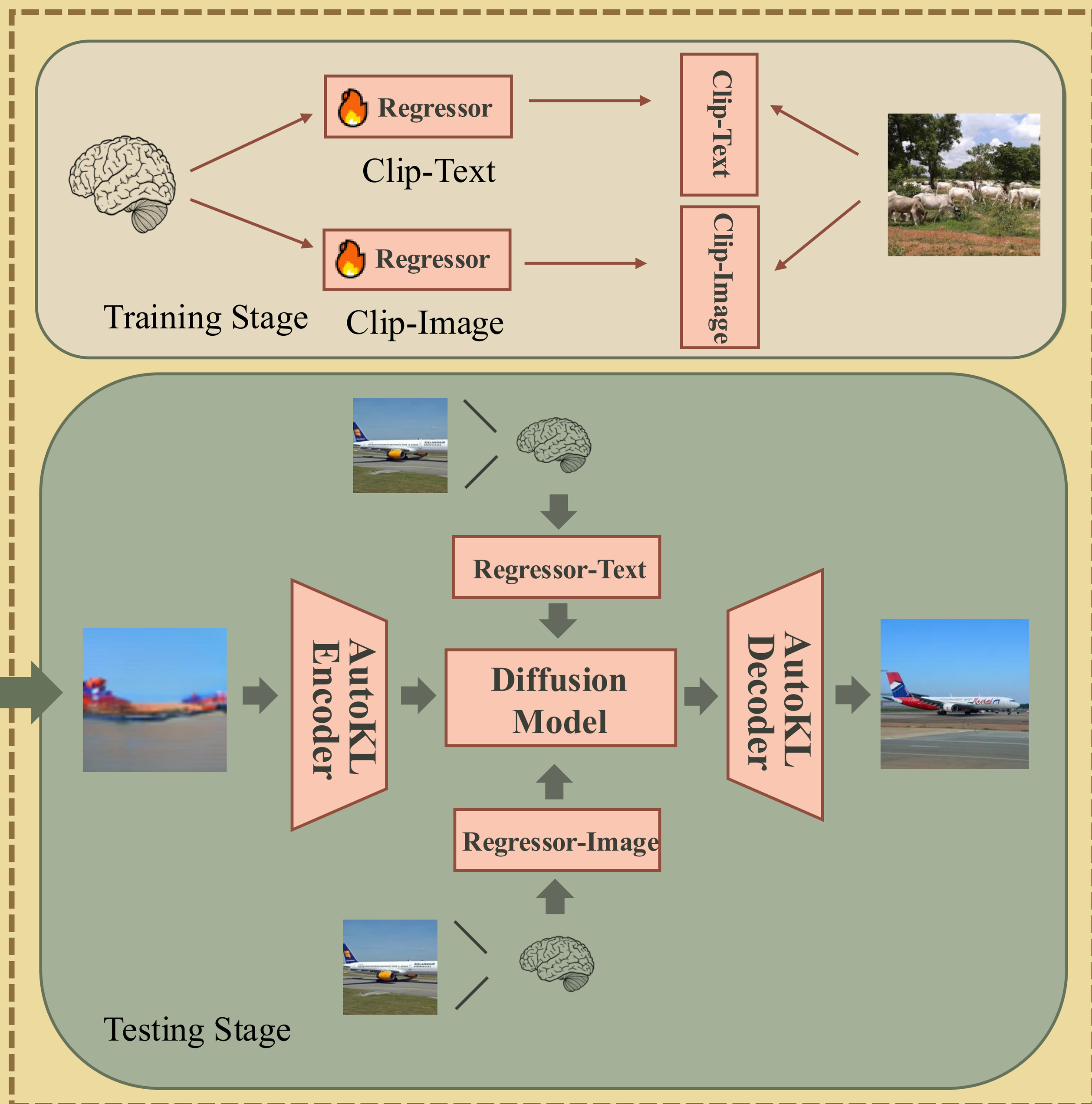


Figure 2:Diffusion Denoising Reconstruction of fMRI

Training Phase

1.Image Branch

The original scene image first passes through the **Fourier Frequency Extractor** 🔥. This module divides the image into ***N*** frequency-band channels in the Fourier domain, each corresponding to a specific frequency range, and assigns a learnable weight to each channel; a weight of 1 retains that band completely, while 0 fully filters it out. The weighted frequency-domain image is then fed into the **frozen** VDVAE encoder ❄, yielding the latent variable.

2.fMRI Branch

The fMRI signals recorded while the subject views the same image are input to a **tunable** regressor 🔥, which predicts the corresponding latent variable.

3.Loss & Optimization

The discrepancy between the predicted and latent variable is used as the **MSEloss** to jointly update the Fourier Frequency Extractor and the regressor, while the VDVAE encoder remains frozen.

Testing Phase

1.New fMRI signals are mapped to latent representations using only the trained regressor.

2.The frozen VDVAE decoder ❄ decodes these latent back into images, completing the reconstruction.

Training Phase

1.Text Branch

The ground-truth caption for each training image passes through the frozen CLIP-Text encoder , yielding a text embedding.

Image Branch

The original scene image passes through the frozen CLIP-Image encoder , yielding an image embedding.

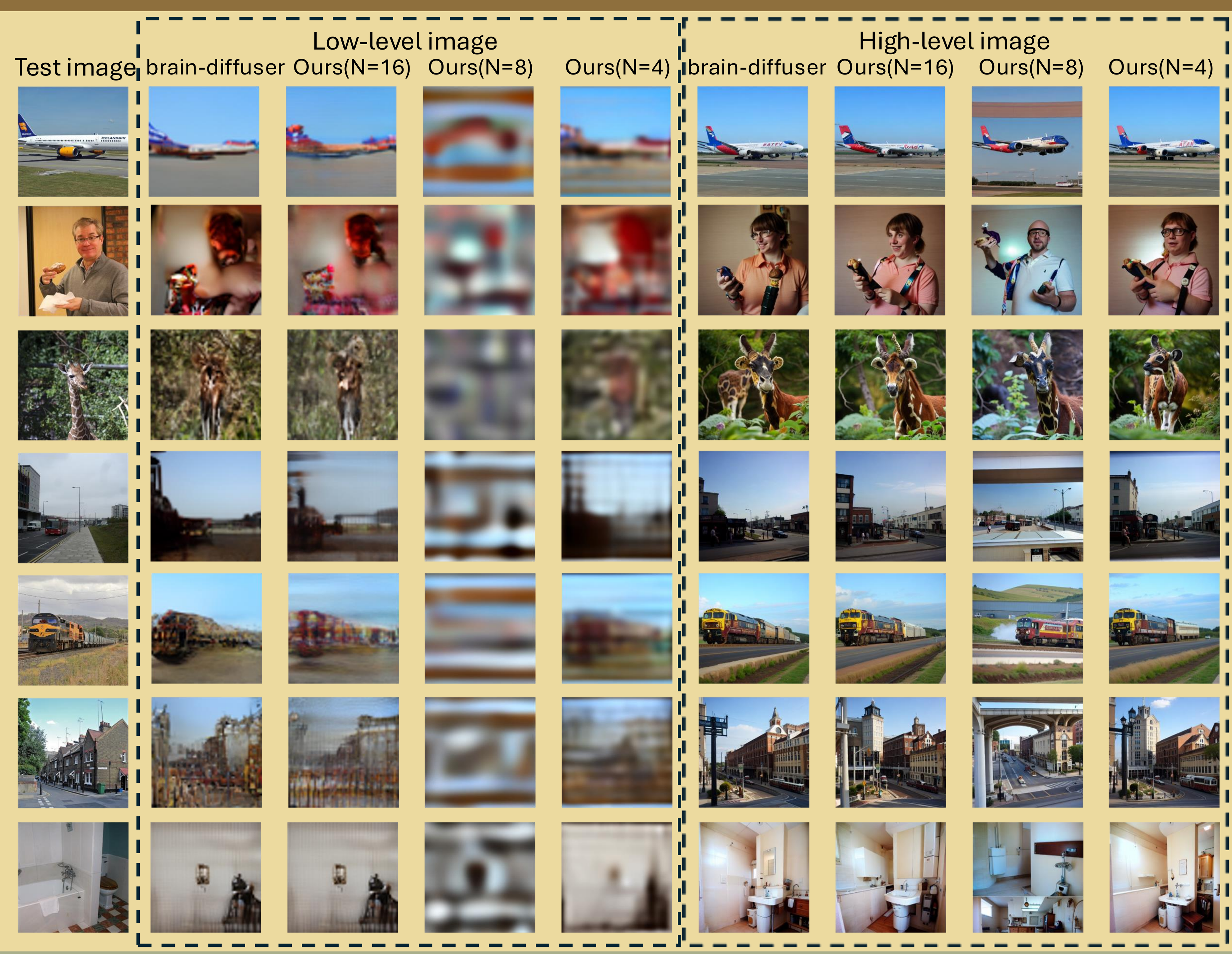
Loss & Optimization

Compute the MSE loss between predicted and true CLIP-Text embeddings, and separately between predicted and true CLIP-Image embeddings. Jointly update both regressors .

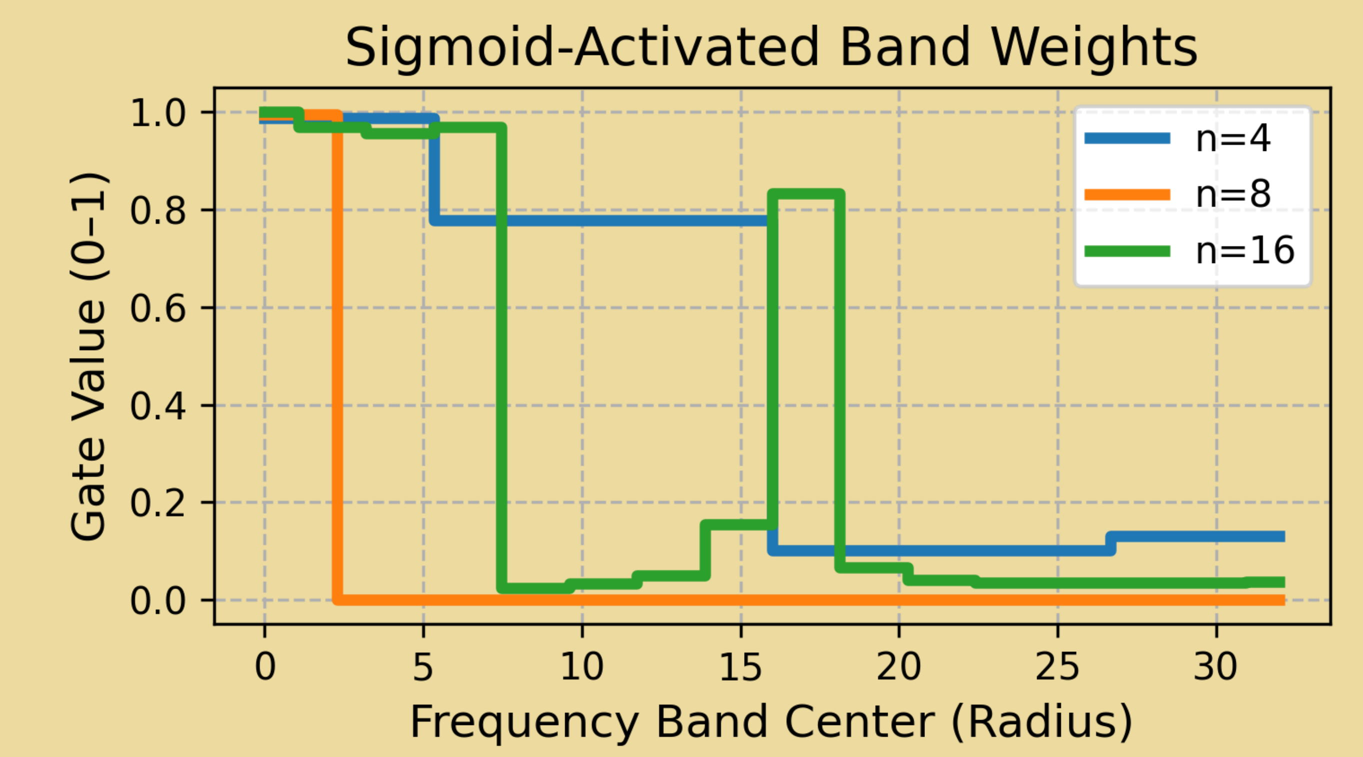
Testing Phase

New fMRI signals are mapped to text and image embeddings via the trained Regressor-Text and Regressor-Image. The pictures in Figure 1 is denoised by the diffusion model, **conditioned** on these two predicted CLIP embeddings.

Visual result



Channel weight



Evaluation

	Low level				High level			
Method	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
Brain-Diffuser*	**0.304**	0.293	**96.84%**	**97.48%**	**88.6%**	92.5%	**0.761**	**0.41**
Ours (N=4)	0.2902	0.2914	94.94%	96.80%	88.11%	91.85%	0.7725	0.4176
Ours (N=8)	0.0738	0.2633	86.82%	93.13%	86.00%	91.81%	0.7986	0.452
Ours (N=16)	0.2734	**0.2961**	96.25%	97.46%	88.36%	**92.65%**	0.7672	0.4141