Proceedings of the 3rd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2025.LD31747

An Improved Method for fMRI-based Human Visual Reconstruction Using Diffusion Models

Tongyu Fang

School of Mechanical and Electronic Control Engineering, Beijing Jiaotong University, Beijing, China
23222064@bjtu.edu.cn

Abstract. We introduce a plug-and-play framework for reconstructing natural images from functional Magnetic Resonance Imaging (fMRI) using a frozen diffusion prior. The approach decouples appearance and semantics: early-visual activity is linearly decoded to a Variational Autoencoder (VAE) latent that preserves structure, while higher-level activity is decoded to a text embedding that guides content. A lightweight diagonal ridge calibration aligns predicted semantics with the model's conditioning space. Reconstruction runs in image-to-image mode by stochastically encoding from the predicted appearance latent and denoising with U-Net conditioned on the calibrated semantics—without fine-tuning the generative model. On Natural Scenes Dataset (NSD), the method improves semantic alignment and perceptual fidelity over strong baselines, offering a simple and reproducible pathway to brain-guided image synthesis.

Keywords: fMRI, brain decoding, image reconstruction, diffusion models, diagonal ridge calibration

1. Introduction

Our visual experience is not determined directly by the objective, physical stimuli to eyes but by a complicated mechanism processing in human brains. Elucidating how human visual system works has been a significant topic in the study of neuroscience. Functional Magnetic Resonance Imaging (fMRI), as a potent non-invasive method providing the whole brain activity information, is widely used in the research of human vision mechanism. Recently, with the development of artificial intelligence, reconstructing the stimuli directly from fMRI-to-image has become feasible. However, current fMRI-to-image reconstruction methods often suffer from semantic distortions and typically require high-precision fMRI data, making high-fidelity, semantically accurate reconstruction particularly challenging.

Previous work applies Generative Adversarial Networks (GANs) to map fMRI representations into a pretrained latent space to synthesize reconstructions consistent with the stimuli [1,2]. Despite this progress, two obstacles remain prominent: (i) preserving semantic fidelity—ensuring that reconstructed images reflect the correct objects and scene relations—and (ii) reducing reliance on high-field or high-SNR acquisitions so that methods remain effective under comparatively low fMRI quality.

DOI: 10.54254/2755-2721/2025.LD31747

Diffusion models generate images by learning to reverse a gradual noising process. During training, clean images are progressively corrupted with Gaussian noise; the model is taught to predict and remove that noise at each step until form a rather high-field picture [3].

We propose a diffusion-prior decoder that jointly conditions on two fMRI-derived signals: an early-visual appearance latent z_0 (a manually drawn occipital cluster covering primary/secondary visual cortex) and a high-level semantic embedding (ventral region, ventral temporal cortex). A lightweight per-dimension diagonal ridge calibration aligns the predicted codes with the model's conditioning space, which enables effective reconstructions from routine-quality fMRI without tuning the generative backbone. On NSD, this yields higher semantic fidelity and perceptual quality than GAN-based and single-stream baselines.

2. Related work

2.1. MinD-Vis

MinD-Vis proposes a two-stage brain-to-image framework: a sparse-coded masked brain modeling (SC-MBM) pretraining on large-scale fMRI to learn high-capacity representations, followed by a double-conditioned latent diffusion model (DC-LDM) that injects fMRI features via cross-attention and time-step conditioning to enforce sampling consistency. Evaluated with n-way top-k semantic identification and FID, MinD-Vis reports clear gains over prior GAN/feature-based baselines and shows transfer to Brain, Object, Landscape Dataset [4].

2.2. Denoising Diffusion Probabilistic Models and U-Net

A Denoising Diffusion Probabilistic Model (DDPM) defines a forward Markov chain that gradually adds Gaussian noise to data, and a learned reverse Markov chain that denoises step-by-step to sample from the data distribution. Training optimizes a variational bound; with the common "ε-prediction" parameterization this becomes essentially denoising score matching, and sampling looks like Langevin-style updates. These models can produce high-quality images [3,5].

U-Net is an encoder–decoder CNN with a contracting path that captures context and a symmetric expanding path for precise localization; skip connections fuse high-resolution encoder features with decoder up-samples. With strong data augmentation it trains effectively from few images and became a powerful tool for prediction [6].

2.3. Stable Diffusion reconstruction from fMRI

Different from Chen et al.'s work, Takagi & Nishimoto's CVPR 2023 method uses the Stable Diffusion model as fixed latent diffusion prior and learns only simple linear mappings from fMRI to two internal representations of the model. While decoding early visual cortex to image latent z_0, the text prompt c_0 is decoded from the ventral cortex. Then, by condition denoising an image with improved semantic meaning is produced. This work establishes latent diffusion models as strong priors for the brain-to-image reconstruction while offering innovative insights into how different latent components map onto visual cortex [7].

DOI: 10.54254/2755-2721/2025.LD31747

3. Methods

3.1. Overview

The overview of our method is shown in Fig. 1.

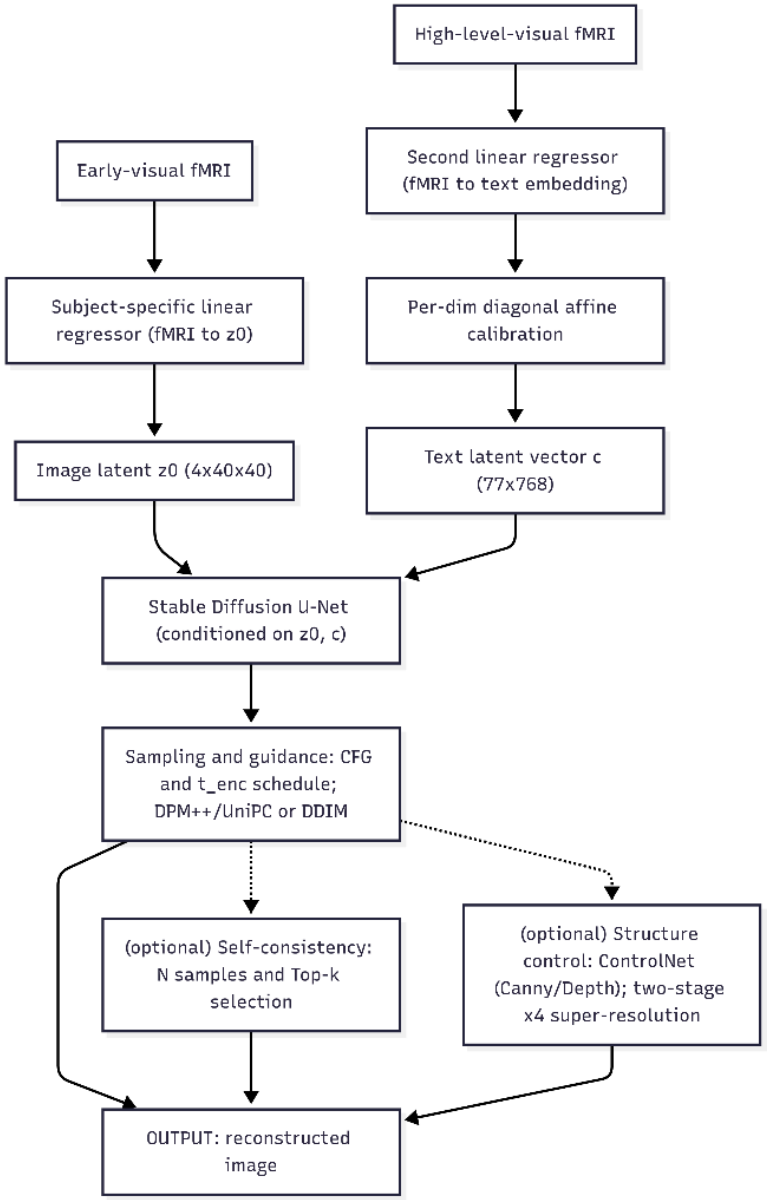


Figure 1. Overview of our pipeline. Early-visual fMRI is linearly decoded to an image latent z (structure), while ventral-stream fMRI is decoded to a text embedding $\hat{c}$, which is calibrated to $\hat{c}^{cal}$ via per-dimension diagonal ridge. Stable Diffusion runs in image-to-image mode, denoising a noisy version of $\hat{z}$, under conditioning on $\hat{c}^{cal}$

3.2. Dataset

In this project, Natural Scenes Dataset (NSD) is chosen. NSD provides whole-brain 7T fMRI at 1.8-mm isotropic resolution from eight participants, and each subject viewed approximately 9,000–

DOI: 10.54254/2755-2721/2025.LD31747

10,000 color natural scenes across ≈30–40 sessions; in total the cohort covers 70,566 distinct images. For Subject 1, we load single-trial fMRI beta estimates (betas_fithrf_GLMdenoise_RR, 1.8-mm in volume space) and restrict analyses to the region of interest (ROI) masks from the stream's atlas (e.g., early visual cortex for appearance and ventral stream for semantics). Visual stimuli are the 70,566 natural images from Microsoft's Common Objects in Context (COCO) used in NSD. Following the standard NSD protocol, we split data into data for training (non-shared images) and data for test (the ≈1k shared images across sessions), averaging fMRI responses across repeats for test items [8].

3.3. Denoising Diffusion Probabilistic Models

DDPM is a Markov Chain trained by variational inference. In the chain, noise is gradually adding to the information by the equation:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\varepsilon_t \tag{1}$$

Where x_t is the new information which noisier than the input x_0, ε_t is Gaussian noise and α_t is the amount of noise that is added to the original information. The reverse model is a Markov chain and in practice the simplified denoising loss can be expressed as:

$$L_{simple} = E\left\{t \sim U\{1 \ldots T\}, x^0 \sim p_{data}, \varepsilon \sim N\right\} \| \varepsilon - \varepsilon_{\theta(x_t,t)} \|_2^2 \tag{2}$$

For conditional situation, the generation $\varepsilon_\theta(x_t, t, c)$ is conditioned on c. In this project, c is derived from the information of ventral area of the fMRI image, and the x_t is derived from the early area information.

3.4. Per-dimension diagonal ridge calibration

We apply a post-hoc diagonal ridge calibration to the fMRI-predicted text embedding $\hat{c}$. Instead of a full matrix, we fit independent 1-D ridge regressions per embedding dimension to align $\hat{c}$ with the ground-truth text embeddings $\tilde{c}$ on the training set. This yields a diagonal scale vector ω and an offset b, and we calibrate test-time predictions by an element-wise affine map [9]:

$$\tilde{c} = reshape\left(w \odot vec\left(\hat{c}\right) + b\right) \quad w, b \in R^{77\times768} \tag{3}$$

Here 77 is the fixed token length and 768 is the hidden size of Contrastive Language–Image Pretraining (CLIP) text encoder.

3.5. The utilization of DDPM & U-Net in fMRI-reconstruction

We pass each stimulus image through the Stable Diffusion VAE encoder (no model fine-tuning) at 320 px to obtain a 4×40×40 latent. This latent z_0 captures low-level structure. Early-visual ROI fMRI (early cortex which locates in V1/V2/V3 around the calcarine sulcus) is paired with as its target.

DOI: 10.54254/2755-2721/2025.LD31747

We feed the image's captions into the Stable Diffusion text encoder to get a conditioning embedding c (averaged over the available captions). High-level ROI fMRI (ventral occipito-temporal cortex) is paired with. Then, a per-dimension diagonal ridge calibration aligns predicted c to the text-encoder space.

For each ROI, we train ridge regressors that map fMRI betas to the model's internal codes:

$$\hat{z}_0 = W_z^\top fMRI_{early} + b_z \quad \text{and} \quad \hat{c} = W_c^\top fMRI_{ventral} + b_c \tag{4}$$

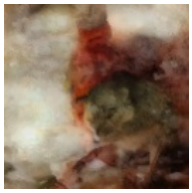
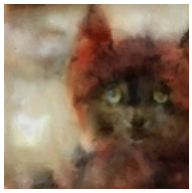
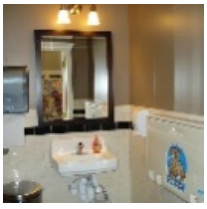
We keep the diffusion backbone frozen and run Stable Diffusion in image-to-image mode: initialize the latent with our predicted z_0 , encode it to timestep t_{enc} (set by the strength parameter). Then, the reverse diffusion is executed where the U-Net denoiser iteratively removes noise while being conditioned on the predicted semantic embedding $\hat{c}$ via cross-attention and classifier-free guidance. Finally, we decode the denoised latent with the VAE decoder to obtain $\hat{x}$.

3.6. Evaluation

We evaluate on the NSD shared-test set (≈982 images per subject; voxel responses averaged across repeats) and report subject-wise means with 95% confidence intervals obtained by bootstrapping images. Unless noted otherwise, we generate five reconstructions per image and report both mean-of-5 and best-of-5 scores (the latter upper-bounds stochastic sampling). Statistical comparisons use paired t-tests across images with Holm–Bonferroni correction [10].

4. Results

Table 1. Visual resultsr

Original	Reconstructed		
	a	b	c
			
			
			

DOI: 10.54254/2755-2721/2025.LD31747

Table 2. Correlation analysis

Method	Pearson r	SSIM
GAN	0.363	0.156
Calibration	0.172	0.388

As the correlation analysis shown in the Table 2, even though we get a lower score in pixel correlation (Pearson r), our method achieves a higher score in perceptual similarity (SSIM), with consistent gains on Subject 1; improvements over ablations are significant ($p < 0.01$) [10,11].

5. Discussion

Overall, the trend matched our expectations, but the improvement was smaller than we had anticipated. The per-dimension diagonal affine calibration did reduce semantic drift and produced modest quality gains, yet its impact was limited for several reasons: (1) the transform is diagonal and linear, so it cannot correct more critical subspace rotations or nonlinear biases in the text-embedding space; (2) given the fMRI signal-to-noise ratio and dataset size, the predicted c still contains systematic noise that simple scaling/shift cannot fully remove; (3) errors in the structural channel z_0 impose a geometric ceiling, and under weak signals the diffusion prior tends to "snap back" to generic appearances. Consequently, we observe stable but modest gains.

However, results are currently demonstrated on a constrained subset (e.g., one subject / limited repeats). Cross-subject performance remains an open question. Subject-agnostic encoders with small subject adapters (feature-wise linear modulation, bias terms) and few-shot calibration on new subjects are promising directions. Report variance across seeds and images to quantify robustness.

6. Conclusion

We presented a simple, plug-and-play framework for fMRI-to-image reconstruction that decouples appearance and semantics and injects them into a frozen diffusion (DDPM) prior. A lightweight diagonal ridge calibration is utilized to align $\hat{c}$ to the text-encoder space. Across NSD, this design improves perceptual quality and semantic alignment without fine-tuning the generator, offering a data-efficient and reproducible path to brain-guided image synthesis. While semantic errors are not fully eliminated, the method establishes a strong, modular baseline that future work can extend with stronger decoders and joint training to further reduce semantic drift.

References

[1] T. Fang, Y. Qi, and G. Pan, "Reconstructing Perceptive Images from Brain Activity by Shape-Semantic GAN, " arXiv (Cornell University), vol. 33, pp. 13038–13048, Jan. 2020.

[2] I. J. Goodfellow et al., "Generative Adversarial Networks, " Arxiv, Jun. 2014, doi: https: //doi.org/10.48550/arxiv.1406.2661.

[3] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models, " arxiv.org, Jun. 2020, doi: https: //doi.org/10.48550/arXiv.2006.11239.

[4] Z. Chen, J. Qing, T. Xiang, W. L. Yue, and J. H. Zhou, "Seeing Beyond the Brain: Conditional Diffusion Model with Sparse Masked Modeling for Vision Decoding, " Nov. 2022, doi: https: //doi.org/10.48550/arxiv.2211.06956.

[5] C. J. Penington, B. D. Hughes, and K. A. Landman, "Building macroscale models from microscale probabilistic models: A general probabilistic approach for nonlinear diffusion and multispecies phenomena, " Physical Review E, vol. 84, no. 4, Oct. 2011, doi: https: //doi.org/10.1103/physreve.84.041120.

DOI: 10.54254/2755-2721/2025.LD31747

[6] M. Baldeon-Calisto and S. K. Lai-Yuen, “AdaResU-Net: Multiobjective Adaptive Convolutional Neural Network for Medical Image Segmentation, ” Neurocomputing, Apr. 2019, doi: https: //doi.org/10.1016/j.neucom.2019.01.110.
[7] Y. Takagi and S. Nishimoto, “High-resolution image reconstruction with latent diffusion models from human brain activity, ” Jun. 2023, doi: https: //doi.org/10.1109/cvpr52729.2023.01389.
[8] E. J. Allen et al., “A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence, ” Nature Neuroscience, vol. 25, no. 1, pp. 116–126, Dec. 2021, doi: https: //doi.org/10.1038/s41593-021-00962-x.
[9] Matthias Minderer et al., “Revisiting the Calibration of Modern Neural Networks, ” Neural Information Processing Systems, vol. 34, Dec. 2021.
[10] S. Holm, “A Simple Sequentially Rejective Multiple Test Procedure, ” Scandinavian Journal of Statistics, vol. 6, pp. 65–70, Jan. 1979, doi: https: //doi.org/10.2307/4615733.
[11] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image Quality Assessment: From Error Visibility to Structural Similarity, ” IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: https: //doi.org/10.1109/tip.2003.819861.