

# Reconstructing seen images from human brain activity via guided stochastic search

**Reese Kneeland (rek@umn.edu)**

Department of Computer Science, University of Minnesota  
Minneapolis, MN 55455 USA

**Jordyn Ojeda (ojeda040@umn.edu)**

Department of Computer Science, University of Minnesota  
Minneapolis, MN 55455 USA

**Ghislain St-Yves (gstyves@umn.edu)**

Department of Neuroscience, University of Minnesota  
Minneapolis, MN 55455 USA

**Thomas Naselaris (nase0005@umn.edu)**

Department of Neuroscience, University of Minnesota  
Minneapolis, MN 55455 USA

## Abstract

Visual reconstruction algorithms are an interpretive tool that map brain activity to pixels. Past reconstruction algorithms employed brute-force search through a massive library to select candidate images that, when passed through an encoding model, accurately predict brain activity. Here, we use conditional generative diffusion models to extend and improve this search-based strategy. We decode a semantic descriptor from human brain activity (7T fMRI) in voxels across most of visual cortex, then use a diffusion model to sample a small library of images conditioned on this descriptor. We pass each sample through an encoding model, select the images that best predict brain activity, and then use these images to seed another library. We show that this process converges on high-quality reconstructions by refining low-level image details while preserving semantic content across iterations. Interestingly, the time-to-convergence differs systematically across visual cortex, suggesting a succinct new way to measure the diversity of representations across visual brain areas.

**Keywords:** decoding; vision; generative models; diffusion models; fMRI

## Introduction

Successfully reconstructing even a moderately complex image from its evoked brain activity requires a strong and appropriate prior. In (Naselaris, Prenger, Kay, Oliver, & Gallant, 2009), we used a massive image library as the prior. We used an encoding model for early areas to screen the library for images with low-level details that were consistent with brain activity in early visual areas. A high-level encoding model was then used to select images that were also consistent with brain activity in higher visual areas. This “structure before semantics” strategy was an inversion of reconstruction priorities, as it produced semantically correct reconstructions only in cases where low-level details and semantic content were highly correlated in the image library. This made the success of the method dependent on the content of the library.

We use a recently developed generative model (Rombach, Blattmann, Lorenz, Esser, & Ommer, 2022) to impose a strong prior that guarantees naturalistic reconstructions with interpretable content. We further take advantage of this model to induce a “semantics before structure” search. We search through an image library generated by the diffusion model with guidance from a semantic CLIP embedding (Radford et al., 2021) decoded from brain activity. We then use encoding models for early visual areas (St-Yves, Allen, Wu, Kay, & Naselaris, 2022) to iteratively refine the library until convergence on a final reconstruction.

## Methods

We analyze the Natural Scenes Dataset (Allen, St-Yves, & Wu, n.d.). For the subject shown here (S1), we partition the data into training ( $n=20,809$ , used for encoding/decoding

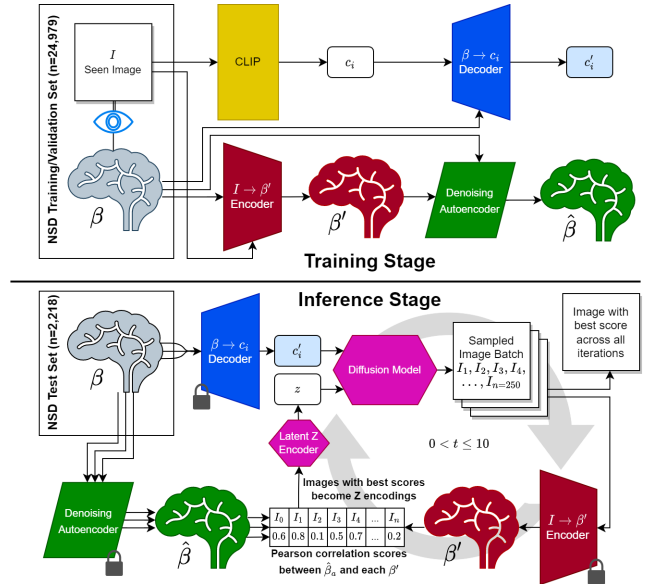


Figure 1: Pipeline diagram for our method: the top half demonstrates the Training stage for our models, and the bottom half depicts the Inference Stage that deploys our stochastic search procedure.



Figure 2: Qualitative comparison of our method against a CLIP decoding-only method, and brute-force search through a library of COCO images. (T.-Y. Lin et al., 2015)

model parameter estimation), validation ( $n=4,170$ , used for model regularization), hyperparameter ( $n=552$ , used to fine-tune the iterative search procedure), and test ( $n=2,218$ , used to assess decoding performance) sets. Hyperparameter and test sets derive from a collection of 1,000 images that were seen by all subjects in the NSD. Training and validation sets originate from the remaining data. Our training stage prepares a decoding model to extract CLIP image embeddings  $c_I$  from brain activity  $\beta$ , voxel-wise encoding models (St-Yves & Naselaris, 2017) that map images  $I$  to predicted brain activity



Figure 3: Refinement of reconstructions over search iterations (left to right)

$f(I) = \beta'$ , and a denoising autoencoder that maps measured brain activity  $\beta$  to denoised brain activity  $\hat{\beta}$ . During inference, we average  $\beta$  across three repetitions of the target image and decode to  $I_{C_I}$ , which is held constant throughout the inference. We denoise the three fMRI data repetitions individually to produce  $\hat{\beta}$ , which we use to guide the search process. We use SD to generate a small library ( $n=250$ ) conditioned on  $I_{C_I}$  and a latent  $z_{t=0}$  sampled from a uniform Gaussian. We pass each image in the library,  $I_j$ , through the encoding models to obtain predicted brain activity  $\beta'_j$ . We scored each image by computing Pearson correlation between the predicted activity pattern and  $\hat{\beta}$ . We select the images with the highest average correlation across the three repetitions, encode them as latent representations  $z_{t=1}$ , and generate a new library. With each iteration, we decrease the strength parameter of the diffusion model, which determines the relative level of guidance of the conditioning  $c_I$  and  $z_t$  variables during diffusion. For the examples shown here, we performed 10 iterations, reducing the strength parameter along a cubic decay schedule from 1.0 to 0.6. We compare our search-based reconstruction procedure to a CLIP-only decoder in which we generate a single sample conditioned on  $I_{C_I}$ , and a simple library search method in which the encoding model is used to select one image from a large (60K images) library of COCO images.

Methods	PixCorr	SSIM	CLIP
Stochastic Search	.215 ± 0.041	.295 ± 0.042	85.1 ± 9.1 %
CLIP Decoding	.067 ± 0.038	.268 ± 0.041	82.1 ± 8.9 %
Best COCO Image	.220 ± 0.043	.277 ± 0.067	82.0 ± 9.0 %

Table 1: Quantitative comparison of methods. PixCorr: pixel-wise correlation metric. SSIM: structural similarity index measure. CLIP: two-way identification experiment comparing CLIP from reconstruction to CLIP from ground truth target and a random reconstruction of the same type.

## Results

Preliminary experiments with our “semantics before structure” search converge on reconstructions that depict the right objects at roughly the right location and scale (Figure 1) by refining low-level detail across search iterations (Figure 3), outperforming simpler approaches on several metrics (Table 1).

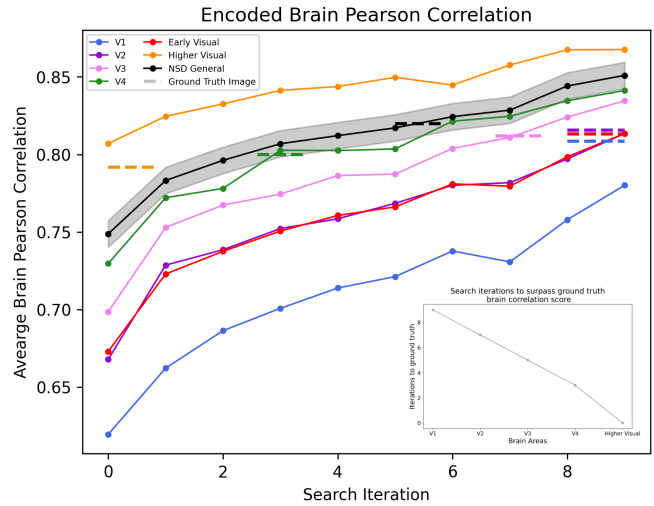


Figure 4: Correlation of predicted and actual brain activity (“score”) for the top image at each iteration for different ROIs (curves). Inset: the number of iterations for each ROI to cross score for ground truth image (dashed lines).

We tracked the score for the best image at each iteration for different brain areas relative to the score for the ground truth target image. The number of search iterations required to achieve the score of the ground truth image decreased monotonically with progression from V1-V4 and into “high-level” visual cortex. This time-to-convergence provides a succinct measure of the level of representational invariance across visual cortex.

## Conclusion

Our effort is one of several (Takagi & Nishimoto, 2023; S. Lin, Sprague, & Singh, 2022; Ozelik & VanRullen, 2023; Gu, Jamison, Kuceyeski, & Sabuncu, 2023; St-Yves & Naselaris, 2019) that use recent developments in AI to achieve impressive reconstruction quality. Although further improvement is warranted, these efforts suggest that reconstructing seen images from human brain activity in controlled settings may soon be a solved problem. For better or worse, this brings promise of using brain decoding as a novel modality for communicating internal states incrementally closer to fulfillment.

## Acknowledgments

Stability AI for providing code and pre-trained models; Professor Paul Schrater for guidance. This work was supported by NIH R01EY023384 (T.N.). Collection of the NSD dataset was supported by NSF IIS-1822683 and NSF IIS-1822929.

## References

- Allen, E., St-Yves, G., & Wu, Y. e. a. (n.d.). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. In *Nat neurosci*.
- Gu, Z., Jamison, K., Kuceyeski, A., & Sabuncu, M. (2023). *Decoding natural image stimuli from fmri data with a surface-based convolutional network*.
- Lin, S., Sprague, T., & Singh, A. K. (2022). *Mind reader: Reconstructing complex images from brain activities*.
- Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., . . . Dollár, P. (2015). *Microsoft coco: Common objects in context*.
- Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6), 902-915. doi: <https://doi.org/10.1016/j.neuron.2009.09.006>
- Ozcelik, F., & VanRullen, R. (2023). *Brain-diffuser: Natural scene reconstruction from fmri signals using generative latent diffusion*.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., . . . Sutskever, I. (2021). *Learning transferable visual models from natural language supervision*.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). *High-resolution image synthesis with latent diffusion models*.
- St-Yves, G., Allen, E. J., Wu, Y., Kay, K., & Naselaris, T. (2022). Brain-optimized neural networks learn non-hierarchical models of representation in human visual cortex. *bioRxiv*. doi: 10.1101/2022.01.21.477293
- St-Yves, G., & Naselaris, T. (2017). The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. doi: 10.1101/126318
- St-Yves, G., & Naselaris, T. (2019, January 16). Generative adversarial networks conditioned on brain activity reconstruct seen images. In (pp. 1054–1061). Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/SMC.2018.00187
- Takagi, Y., & Nishimoto, S. (2023). High-resolution image reconstruction with latent diffusion models from human brain activity. *bioRxiv*. doi: 10.1101/2022.11.18.517004