MIRAGE: Robust multi-modal architectures translate fMRI-to-image models from vision to mental imagery

Reese Kneeland (reese@alljoined.com)
University of Minnesota
Alljoined

Cesar Kadir Torrico Villanueva
Medical AI Research Center (MedARC)

Tong Chen
University of Sydney

Jordyn Ojeda
University of Minnesota

Shubh Khanna
Stanford University

Jonathan Xu
Alljoined
Medical AI Research Center (MedARC)
University of Waterloo

Paul S. Scotti
Medical AI Research Center (MedARC)
Formerly Stability AI
Princeton Neuroscience Institute

Thomas Naselaris (nase0005@umn.edu)
University of Minnesota

Abstract

To be practically useful, vision decoding models trained on perceived images must generalize effectively to mental images—internally generated visual representations. Analysis of the NSD-Imagery dataset revealed that achieving state-of-the-art (SOTA) results on seen images does not guarantee similar performance on mental images. To address this gap, we developed MIRAGE, a model explicitly designed to generalize from visually perceived data to mental imagery decoding. MIRAGE employs a robust ridge regression approach, utilizes multi-modal conditioning with text and small-dimension image embeddings, and leverages the Stable Cascade diffusion model. Extensive human evaluations and image metrics establish MIRAGE as the SOTA method for mental image reconstruction on NSD-Imagery. Our ablation studies emphasize that mental imagery decoding benefits from simple architectures robust to low signal-to-noise conditions, explicit low-level guidance, multi-modal semantic features, and lower-dimensional embeddings compared to typical vision decoders. These findings highlight the potential of existing visual datasets for training models capable of effective mental imagery decoding.

Keywords: fMRI; mental imagery; decoding

Introduction

Decoding mental imagery—visual experiences generated internally without direct sensory input—from brain activity is crucial for advancing brain-computer interfaces and clinical tools for communication-impaired individuals. Although recent developments in generative AI and deep learning have significantly improved decoding visual perceptions from brain signals (Takagi & Nishimoto, 2023; Scotti et al., 2024, 2023), these advances do not reliably extend to decoding internally generated mental images (Kneeland et al., 2025). The recently released NSD-Imagery dataset highlighted that state-of-the-art vision decoders such as MindEye2 (Scotti et al., 2024)—trained on visually perceived images—often perform poorly when applied to mental imagery, underscoring the unique neurobiological and representational challenges associated with mental imagery (Favila, Kuhl, & Winawer, 2019; Friston, 2017).

We present MIRAGE (Mental Image Reconstruction using Advanced Generative ModEls), designed specifically for mental imagery decoding. MIRAGE leverages architectural features inspired by frameworks that are observed to be more robust on mental images (Ozcelik & VanRullen, 2023), and demonstrates superior performance through critical architectural choices that enable successful generalization from seen to imagined images.

MIRAGE

Datasets

We train MIRAGE exclusively on the Natural Scenes Dataset (NSD) (Allen et al., 2022), using subjects 1, 2, 5, and 7, each completing approximately 30k fMRI-image pairs. Data preprocessing included applying the provided nsdgeneral voxel mask at 1.8 mm resolution, capturing activity from early to higher-level visual cortex areas.

Evaluation is conducted on the NSD-Imagery benchmark (Kneeland et al., 2025), where participants imagined prelearned visual stimuli upon cue presentation. The experimental conditions closely follow NSD, enabling direct comparisons.

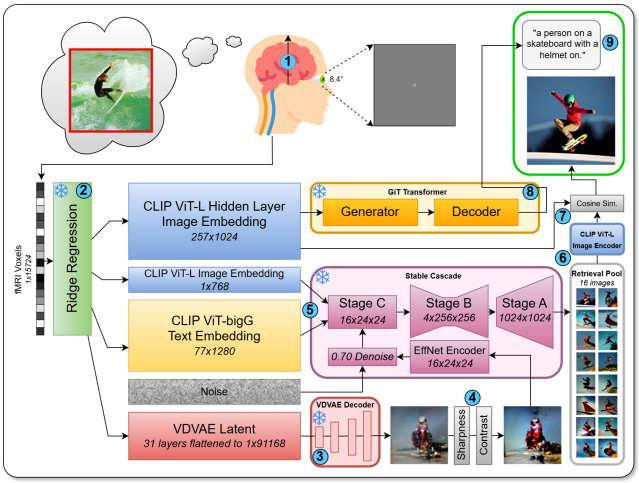

Figure 1: **MIRAGE** inference pipeline. (1) Subjects imagine visual stimuli cued by letters during 7T fMRI scanning. (2) Ridge regression decodes fMRI data into embeddings. (3) VDVAE (Child, 2021) latents reconstruct a low-level image, (4) which is filtered to enhance structure. (5) Stable Cascade (Pernias et al., 2024) uses the filtered image and decoded embeddings to generate 16 candidate images. (6) Candidates are encoded using CLIP ViT-L/14 (Radford et al., 2021), (7) and the final reconstruction is selected based on cosine similarity to the decoded embedding. (8) The selected embedding drives a GiT (Wang et al., 2022) captioning model to (9) produce a caption alongside the final image.

Methodology

MIRAGE incorporates two essential design principles for effective generalization from vision decoding to mental imagery:

Reduced model complexity: MIRAGE uses linear ridge regression instead of complex non-linear MLPs, enhancing robustness in low-SNR settings typical of mental imagery (Hoerl & Kennard, 1970).

Representational alignment: Leveraging the semantic overlap between vision and imagery in higher-level visual cortex (Breedlove, St-Yves, Olman, & Naselaris, 2020), and area known to encode information in language-like formats, MIRAGE incorporates multi-modal guidance from CLIP embeddings derived from both images and text.

Results

The mental image reconstructions produced by **MIRAGE** (Figure 2) are qualitatively quite faithful to the ground truth images the subjects were instructed to imagine. To validate this property, we conducted a large-scale behavioral experiment where

Figure 2: **MIRAGE** reconstructions of imagined stimuli from NSD-Imagery.

human raters (n=500) were asked to perform a two-alternative forced choice judgement between matched and unmatched reconstructions of the target stimuli. Human raters identified the correct mental image reconstructions 78.30% of the time ($p < 0.001$) for our method (Table 1).

Human Identification Accuracy – Mental Imagery Reconstructions				
Method	All Stimuli ↑	Simple ↑	Complex ↑	Conceptual ↑
MIRAGE (ours)	**78.30%**	**73.93%**	**83.19%**	**77.68%**
MindEye1	73.00%	<u>71.01%</u>	82.28%	65.68%
Brain Diffuser	<u>73.95%</u>	68.20%	<u>82.70%</u>	<u>71.01%</u>
iCNN	66.15%	66.81%	70.04%	61.60%
MindEye2	56.96%	50.21%	64.83%	55.74%

Table 1: Human identification accuracy (chance = 50%). Best values bold; second best underlined.

Ablation Study

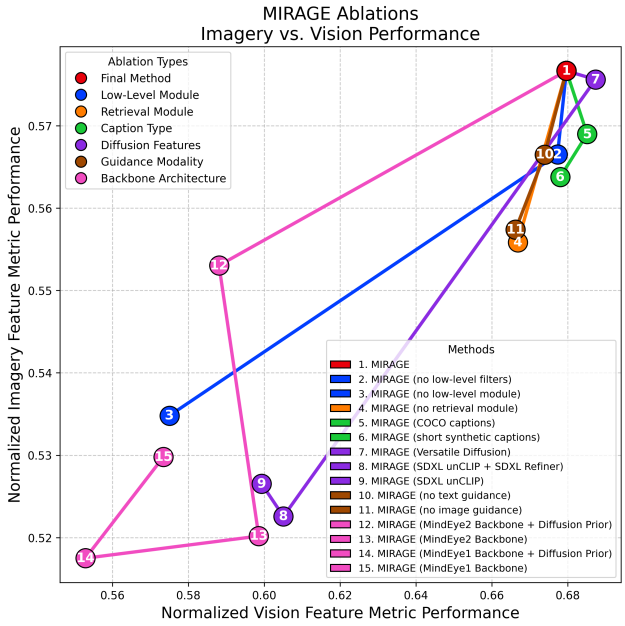

Figure 3: Ablation analyses. Model variants (numbered circles) under each ablation type (color) are assessed via a normalized average of a set of image feature metrics for vision (x-axis) and imagery (y-axis).

To understand why MIRAGE succeeds where other methods struggle, we conducted extensive ablations examining several critical architectural choices (Figure 3). The linear ridge regression backbone selected in our model (1) consistently outperformed the more complex MLP and diffusion prior architectures from MindEye models (12–15), emphasizing that simpler, more robust models generalize better to the inherently lower signal-to-noise ratio of mental imagery. Furthermore, incorporating multimodal guidance—combining both image and text features (1)—yielded substantial performance improvements over image-only (10) or text-only (11) guidance alone. The retrieval module (4), which leverages high-dimensional embeddings, significantly enhanced reconstruction quality. Additional benefits were observed from implementing image filtering techniques (2) and employing longer synthetic captions (1), demonstrating these elements effectively mitigate structural and representational differences between perceived visual stimuli and internally generated mental images.

Discussion

MIRAGE introduces a powerful decoding pipeline that significantly advances mental imagery reconstruction beyond existing state-of-the-art methods (Scotti et al., 2024; Ozcelik & VanRullen, 2023). Its simplified linear decoding backbone, dimensionality reduction of latent representations, and multimodal integration are pivotal for its success. MIRAGE's effective training solely on vision datasets greatly simplifies data requirements, addressing the scarcity of mental imagery datasets. However, limitations remain, including dataset availability and computational requirements, which pose challenges for practical deployment.

Potential applications for MIRAGE span consumer-focused brain-computer interfaces, memory visualization tools, and clinical diagnostics for psychiatric disorders (Holmes & Mathews, 2010), disorders of consciousness (Edlow et al., 2017), or severe motor disabilities (Canny, Vansteensel, van der Salm, Müller-Putz, & Berezutskaya, 2023).

Ethical Considerations

Brain decoding technologies raise significant ethical issues concerning privacy, consent, and misuse. Clear ethical and legal frameworks are vital for responsible usage. Clinical applications must adhere strictly to medical guidelines and patient privacy regulations, while non-clinical uses should enforce informed consent and additional protective measures to safeguard individual rights.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... Kay, K. (2022, January). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*(1), 116–126. Retrieved 2022-11-01, from https://www.nature.com/articles/s41593-021-00962-x doi: 10.1038/s41593-021-00962-x

Breedlove, J. L., St-Yves, G., Olman, C. A., & Naselaris, T. (2020). Generative feedback explains distinct brain activity codes for seen and mental images. *Current Biology*, *30*(12), 2211-2224.e6. Retrieved from https://www.sciencedirect.com/science/article/pii/S0960982220304942 doi: https://doi.org/10.1016/j.cub.2020.04.014

Canny, E., Vansteensel, M. J., van der Salm, S. M., Müller-Putz, G. R., & Berezutskaya, J. (2023, Nov). Boosting brain–computer interfaces with functional electrical stimulation: Potential applications in people with locked-in syndrome. *Journal of NeuroEngineering and Rehabilitation*, *20*(1). doi: 10.1186/s12984-023-01272-y

Child, R. (2021). Very deep {vae}s generalize autoregressive models and can outperform them on images. In *International conference on learning representations.* Retrieved from https://openreview.net/forum?id=RLRXCV6DbEJ

Edlow, B. L., Chatelle, C., Spencer, C. A., Chu, C. J., Bodien, Y. G., O'Connor, K. L., ... et al. (2017). Early detection of consciousness in patients with acute severe traumatic brain injury. *Brain*, *140*(9), 2399–2414. doi: 10.1093/brain/awx176

Favila, S. E., Kuhl, B. A., & Winawer, J. (2019). Spatial perception and memory have distinct activation profiles in human visual cortex. *BioRxiv*, 811331.

Friston, K. (2017). Distinct top-down and bottom-up brain connectivity during visual perception and imagery. *Scientific reports*, *7*(1), 5677.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. Retrieved 2025-01-30, from http://www.jstor.org/stable/1267351

Holmes, E. A., & Mathews, A. (2010). Mental imagery in emotion and emotional disorders. *Clinical psychology review*, *30*(3), 349–362.

Kneeland, R., Scotti, P. S., St-Yves, G., Breedlove, J., Kay, K., & Naselaris, T. (2025, June). Nsd-imagery: A benchmark dataset for extending fmri vision decoding methods to mental imagery. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr).*

Ozcelik, F., & VanRullen, R. (2023). Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, *13*. Retrieved from https://api.semanticscholar.org/CorpusID:260439960

Pernias, P., Rampas, D., Richter, M. L., Pal, C., & Aubreville, M. (2024). Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In *The twelfth international conference on learning representations.* Retrieved from https://openreview.net/forum?id=gU58d5QeGv

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021, 18–24 Jul). Learning transferable visual models from natural language supervision. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th international conference on machine learning* (Vol. 139, pp. 8748–8763). PMLR. Retrieved from https://proceedings.mlr.press/v139/radford21a.html

Scotti, P. S., Banerjee, A., Goode, J., Shabalin, S., Nguyen, A., Ethan, C., ... Abraham, T. M. (2023). Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. In *Thirty-seventh conference on neural information processing systems.* Retrieved from https://openreview.net/forum?id=rwrblCYb2A

Scotti, P. S., Tripathy, M., Torrico, C., Kneeland, R., Chen, T., Narang, A., ... Abraham, T. M. (2024). Mindeye2: Shared-subject models enable fMRI-to-image with 1 hour of data. In *Iclr 2024 workshop on representational alignment.* Retrieved from https://openreview.net/forum?id=paqqd1O0Dl

Takagi, Y., & Nishimoto, S. (2023). High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 14453–14463).

Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., ... Wang, L. (2022). GIT: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research*. Retrieved from https://openreview.net/forum?id=b4tMhpN0JC