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

Reese Kneeland[1,2] **Cesar Kadir Torrico Villanueva**[3,*] **Tong Chen**[4,*]
Jordyn Ojeda[1] **Shubh Khanna**[5] **Jonathan Xu**[2,3,6] **Paul S. Scotti**[3,7,8]
Thomas Naselaris[1]

[1]University of Minnesota
[2]Alljoined
[3]Medical AI Research Center (MedARC)
[4]University of Sydney
[5]Stanford University
[6]University of Waterloo
[7]Formerly Stability AI
[8]Princeton Neuroscience Institute

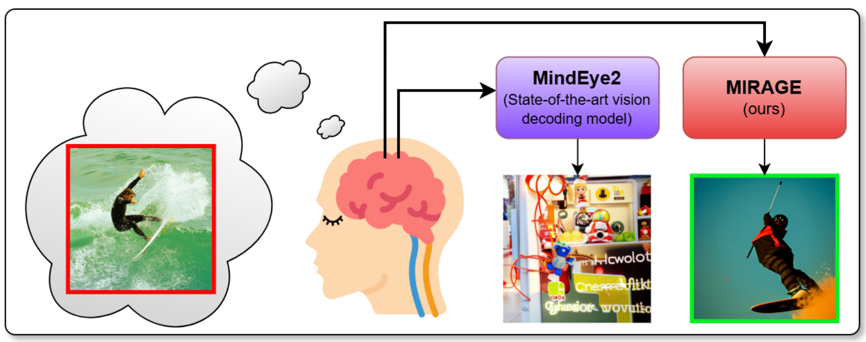

Figure 1: **MIRAGE** (ours) vs MindEye2 [1] reconstructions of an imagined image from fMRI brain activity.

Abstract

To be useful for downstream applications, vision decoding models that are trained to reconstruct seen images from human brain activity must be able to generalize to internally generated visual representations, i.e., mental images. In an analysis of the recently released NSD-Imagery dataset, we demonstrated that while some modern vision decoders can perform quite well on mental image reconstruction, some fail, and that state-of-the-art (SOTA) performance on seen image reconstruction is no guarantee of SOTA performance on mental image reconstruction. Motivated by these findings, we developed **MIRAGE**, a method explicitly designed to train on vision datasets and cross-decode mental images from brain activity. **MIRAGE** employs a linear backbone and multi-modal text and image features as input to a diffusion model. Feature metrics and human raters establish MIRAGE as SOTA for mental image reconstruction on the NSD-Imagery benchmark. With ablation analysis we show that mental image reconstruction works best when decoders use image features with relatively few dimensions and include guidance from text-based and both high- and low-level image-based features. Our work indicates that–given the right architecture–existing large-scale datasets using external stimuli

[*]Core contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body.

are viable training data for decoding mental images, and warrant optimism about the future success and utility of mental image reconstruction.

1 Introduction

The ability to decode and reconstruct mental images—internally generated visual representations not driven by sensory input—from brain activity has tremendous potential for downstream applications such as brain-computer interfaces and medical diagnostics for patients with disorders of communication or consciousness. Externalizing mental images also provides insights into cognitive processes that would be otherwise inaccessible.

Recent research on decoding has focused on developing "vision decoding models" that are trained and tested on brain responses to seen images. To utilize such vision decoding models for the downstream applications we envision, it is important to demonstrate that these vision decoding models generalize well when tested on mental imagery. Recent research on NSD-Imagery [2] has shown that this is often not the case, as decoding performance on seen images is a poor predictor of decoding performance on mental images. This indicates that making progress toward a decoder with practical utility will require careful consideration of the factors that make mental image reconstruction challenging, and of the attributes that make vision decoders more or less likely to generalize. To that end, we present the following contributions:

(1) We introduce **MIRAGE** (**M**ental **I**mage **R**econstruction using **A**dvanced **G**enerative Mod**E**ls), the first method explicitly designed for effective cross-decoding of internally generated mental images after training exclusively on vision data from NSD. **(2)** We establish **MIRAGE** as a SOTA mental imagery decoding method by comparing evaluations across a broad selection of image feature metrics and human preference ratings derived from large-scale behavioral experiments. **(3)** We conduct a detailed ablation analysis to isolate the architectural choices that facilitate generalization from seen to mental image reconstruction, and identify the specific technical reasons why MIRAGE successfully generalizes to NSD-Imagery where MindEye2 and other vision decoding models fail (Figure 1).

2 Related Work

Basic neuroscience has demonstrated an extensive overlap in the representation of seen and mental images [3–7]. Nonetheless, differences between vision and mental imagery make cross-decoding from vision to imagery challenging [8, 9]. Compared to vision, brain activity during mental imagery has a lower signal-to-noise ratio (SNR) [10], varies along fewer signal dimensions [11], and encodes imagined stimuli with less spatial resolution [12, 13].

Previous studies have reported successful n-way classification of mental images [14, 8, 15], retrieval of mental images of natural scenes [6, 16], and reconstruction of simple blobs, letters, or singular natural objects [17, 5, 18–21]. With the open releases of CLIP [22], Stable Diffusion [23], and large-scale fMRI datasets like NSD [24], newer vision decoding models now yield highly accurate reconstructions of natural scenes [25–35]. These methods map fMRI brain activity patterns to embeddings of pre-trained deep learning models that are used to drive a diffusion model [36–38, 1] to generate image reconstructions of the content present in visual cortex.

Tests of open-source vision decoding models [27, 28, 1, 20, 25, 26] on mental imagery activity in the NSD-Imagery dataset [2] showed that improved performance on vision decoding does not necessarily translate to mental imagery decoding. For example, MindEye2 [1] was the SOTA model on the test set of NSD, but not on the imagery trials of NSD-Imagery (Figure 1). Furthermore, in that study it was observed that methods with linear backbones, compact representations, and multimodal guidance yielded the best performance. We interpreted these observations as follows: First, [39] showed that linear mappings from CLIP to brain activity explains much of variance in activity in visual cortex. Thus, we should expect linear methods to work quite well when decoding CLIP. Second, as models increase in expressivity, models trained in a high SNR regime may overfit to the input, exhibiting unacceptably high levels of error variance when tested in a lower SNR setting [40]; this danger is especially acute in our case, given that imagery has very low SNR. Thus, our decoding pipeline prioritizes robustness over expressivity, implementing a linear ridge regression backbone (Figure 2)—in contrast to the less robust non-linear MLP backbones used more commonly in vision decoding. Third, imagery and visual activity are most aligned in brain areas with language-like representations ([8, 12, 41]). Thus, it makes sense to include high-quality language-like representations in the decoding pipeline for mental imagery. Our method drives an image generator with the CLS token of a CLIP ViT-L/14 image embedding, and incorporates multi-modal guidance from decoded CLIP ViT-bigG/14 text features (Figure 3). These ideas informed the design of MIRAGE, and our results, including detailed ablation analyses in Appendix A.2, support our reasoning, explaining why MIRAGE is SOTA for mental image reconstruction.

3 MIRAGE

3.1 Datasets

Our method is trained exclusively on the Natural Scenes Dataset (NSD) [24], and evaluated primarily on the NSD-Imagery benchmark [2]. NSD consists of between 22k and 30k fMRI-natural image pairs per subject (8 subjects). We train models for the nsdgeneral mask of subjects 1, 2, 5, and 7, as only these subjects completed the full 30k trials of the NSD experiment. Data from the other 4 NSD subjects are used as a hyperparameter tuning set, as discussed in Section 3.2. NSD-Imagery is a small extension of NSD where the same subjects completed mental imagery trials. In this separate scan session, the NSD subjects learned to associate single-letter cues with a set of shapes, natural scenes, and visual concepts (e.g., “stripes"), and imagine them when presented with the corresponding cue.

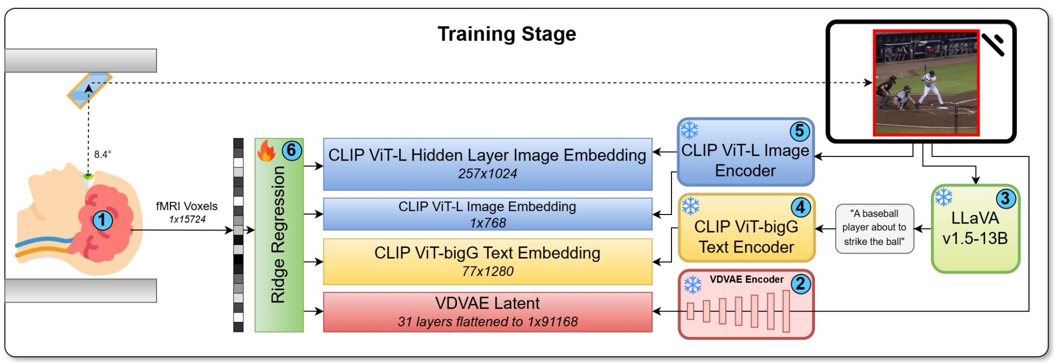

Figure 2: **MIRAGE** training pipeline. (1) Brain activity (7T fMRI) acquired as NSD subjects view $> 10K$ stimuli. (2) Stimuli are passed to VDVAE encoder [36] yielding (1×91168) latents (3) LLaVA v1.5-13B [42, 43] generates synthetic captions. (4) Captions are encoded into CLIP ViT-bigG/14 text embeddings (77×1280) [44]. (5) Stimuli are also passed through the CLIP ViT-L/14 image encoder [22] to generate both CLS token (1×768) and hidden layer (257×1024) image embeddings. (6) Parallel ridge regression modules are trained from the measured fMRI brain activity to the various feature spaces.

3.2 Architectural Components

Ridge Regression Backbone: To map preprocessed fMRI data to feature representations in our decoding pipeline, we employ parallel L_2 regularized ridge regression models trained on each individual feature set. In keeping with our first constraint, ridge regression is chosen because it is known to be effective in low signal-to-noise regimes, in contrast to MLPs that offer advantages for capturing nonlinear relationships, but can also be fragile in brain decoding contexts when the input data are of low SNR. For each set of features, we train a parallel ridge regression model to predict the feature value from our fMRI responses. Details on our hyperparameter search are in Appendix A.3.

Low-Level Module: To begin the reconstruction process, we decode and reconstruct a “low-level" image that captures the structural layout of the target image. Inspired by the Brain Diffuser method [27], our approach utilizes a Very Deep Variational Autoencoder (VDVAE) [36] for this task. Our model predicts the first 31 layers of the VDVAE model, which are concatenated into a 1×91168 dimensional feature vector and passed through the VDVAE decoder to generate reconstructed images at 64×64 pixel resolution. We also apply a set of image filters to boost the sharpness and contrast of the low-level images. Details about our implementation of the VDVAE are in Appendix A.4.

Image Features: Noting that the large high-dimensional hidden layer CLIP ViT-bigG/14 image embeddings used in the MindEye2 architecture (257×1664) fail to robustly drive image generators when decoded from mental images, we use only the CLS token of the CLIP ViT-L/14 model (1×768)

for our image embeddings. We hypothesize that the removal of the spatial patch tokens used in the hidden layer of ViT embeddings increases the alignment of our features with the encoding of mental images in the brain.

Text Features: To accommodate any potentially “language-like" properties of mental images encoded in brain activity, we incorporate multi-modal guidance into our method through the addition of an unpooled CLIP ViT-bigG/14 text embedding (77×1280). To increase the quality of our text features, we replace the COCO captions for the training set with synthetic captions generated by LLaVA v1.5-13B that are longer and more descriptive.

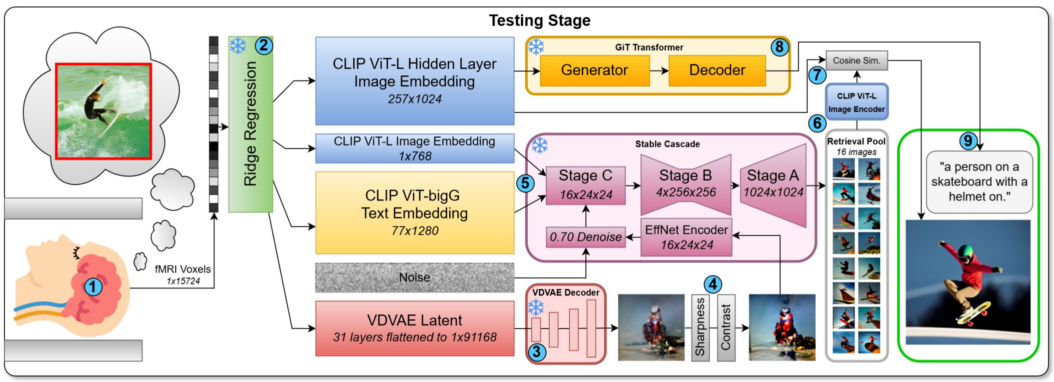

Figure 3: **MIRAGE** inference pipeline. (1) The NSD subjects imagine stimuli from letter cues under 7T fMRI. (2) A set of feature embeddings is predicted by passing the measured fMRI brain activity through our frozen ridge regression models. (3) The VDVAE [36] latents are reconstructed into a low-level image. (4) The image is filtered to boost its structure. (5) The filtered low-level image, decoded image embedding, and decoded text embedding are used as input to Stable Cascade [45] to generate a retrieval pool of 16 candidate reconstructions. (6) Each of the candidate reconstructions is encoded into a CLIP ViT-L/14 hidden layer image embedding [22]. (7) The final reconstruction is automatically selected as the image with the highest cosine similarity to the decoded CLIP ViT-L/14 hidden layer image embedding. (8) The same hidden layer image embedding is passed through the GiT [46] captioning model. (9) A decoded caption is generated along with our final reconstruction.

Diffusion Architecture: To generate the final reconstructions, we utilize Stable Cascade, one of the latest multi-modal diffusion models by Stability AI. Stable Cascade uses Würstchen [45], a text-to-image architecture that also has support for CLIP Image guidance. This architecture accepts both our small 1×768 CLIP ViT-L/14 image embeddings, 77×1280 CLIP ViT-bigG/14 text embeddings, and a partially diffused input image for img2img mode [47]. The model’s multi-modal guidance capabilities, native img2img support, and robust multi-stage architecture make it well-suited for our mental imagery reconstruction task.

Retrieval Pooling: Because diffusion models are stochastic generators, the quality of the reconstructions they yield will vary from sample to sample. Our pipeline therefore includes a retrieval step to select the best reconstruction from among a small pool samples output by the diffusion model. We decode a CLIP ViT-L/14 hidden layer image embedding (257×1024) that has been unit normalized to approximate a cosine similarity loss in the regression training stage. We then generate 16 candidate reconstructions to form a retrieval pool, pass each candidate through the CLIP ViT-L/14 image encoder to get hidden layer image embeddings, and select only the highest-scoring reconstruction via cosine similarity with our decoded retrieval embedding as our output.

Caption Decoding: We implement a caption decoding module that provides a text description of the visual content being decoded. To decode captions, we reuse the CLIP ViT-L/14 hidden layer image features decoded in the retrieval module and pass them through a frozen GiT [46] transformer (which contains a generator and a decoder) to generate a predicted caption, an approach originally pioneered by Ferrante et al. [48] and MindEye2 [1].

4 Results

Some mental image reconstructions produced by **MIRAGE** can be seen in Figure 4. For simple stimuli, the overall structure and orientation of the ground truth images are approximated. For complex stimuli, we successfully reconstruct images that contain similar categories and themes—such as

Figure 4: **MIRAGE** best reconstructions of imagined stimuli from NSD-Imagery.

donuts, a man riding a surfboard, and a bird-like animal with a beak. For conceptual stimuli, the images reconstructed clearly reflect content that is related or identical to the corresponding concept word, with very clear "stripes", a "zebra", and a recognizable banana for the "banana" prompt. Vision reconstructions for simple and complex stimuli (Appendix A.5) also perform well, despite vision decoding performance not being our primary target. Complete reconstructions can be seen in Appendices A.5, A.6, and A.7. Quantitative benchmarks across image-feature metrics are provided in Appendix A.8.

4.1 Human ratings of reconstruction quality

We conducted several large-scale online behavioral experiments in which human raters (n=500) assessed the quality of the reconstructions. (See Appendix A.17).

Experiment 1 Human raters performed a 2-alternative forced choice (2AFC) judgment about whether a reconstruction was more similar to the ground truth image than a randomly selected reconstruction of a different stimulus sampled from the same stimulus type, method, and NSD subject. Results (Table 5A) confirm **MIRAGE** as SOTA for every stimulus type.

Experiment 2 Human raters viewed a ground truth image, a reconstruction of that stimulus from a vision trial, and a reconstruction of the same stimulus from an imagery trial. Raters provided continuous measures of similarity between each reconstruction and the ground truth image. The rating provided a direct comparison of the similarity between vision and imagery reconstructions. **MIRAGE** shows SOTA generalization from vison to imagery ([39]).

A

Human Identification Accuracy

Method	All Stimuli ↑	Simple ↑	Complex ↑	Conceptual ↑
Mental Imagery Reconstructions				
MIRAGE (ours)	**78.30%**	**73.93%**	**83.19%**	**77.68%**
MindEye1	73.00%	71.01%	82.28%	65.68%
Brain Diffuser	73.95%	68.20%	82.70%	71.01%
iCNN	66.15%	66.81%	70.04%	61.60%
MindEye2	56.96%	50.21%	64.83%	55.74%
Vision Reconstructions				
MIRAGE (ours)	84.22%	**80.77%**	87.66%	*N/A*
MindEye1	**84.29%**	79.32%	89.32%	*N/A*
Brain Diffuser	80.13%	71.79%	88.46%	*N/A*
iCNN	74.52%	75.85%	73.19%	*N/A*
MindEye2	83.05%	75.54%	**90.56%**	*N/A*

B

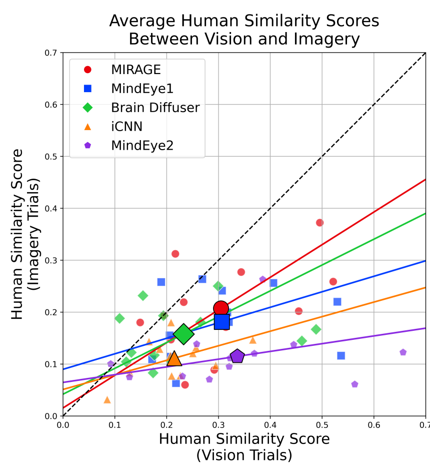

Figure 5: **(A)** Human identification accuracy scores for the vision and mental imagery trials of the NSD-Imagery benchmark. Scores are provided for each method and stimulus type; best values are bolded and second best underlined (chance = 50%, all $p < 0.001$). **(B)** Human similarity scores for simple and complex stimuli: X-axis = vision, Y-axis = imagery; each point is the mean over 12 samples (larger bold points are the overall means), colored/shaped by method. PCA-fit slopes closer to unity indicate tighter imagery–vision correspondence; dashed unity line shown.

5 Discussion

MIRAGE improves mental imagery reconstruction over vision reconstruction pipelines shown to be SOTA on NSD shared1000 test samples (i.e., MindEye2 [1]) and on the NSD-imagery dataset (i.e., BrainDiffuser [27]). Ablation studies revealed that the keys to the success of MIRAGE are (1) a simple linear decoding backbone, (2) a reduction in the dimension of latent image representations, and (3) the inclusion of high-quality multi-modal guidance to the diffusion model.

Our work is a necessary step towards applications of mental image reconstruction, including diagnostic instruments for psychiatric conditions [49] and disorders of consciousness [50–52], and alternative communication methods for patients with traumatic brain injuries [53], amyotrophic lateral sclerosis (ALS) [54], and locked-in syndrome [55].

Of course, the development of this technology raises concerns about the potential for misuse [56]. We propose that when deployed in a clinical setting, brain decoding should be defined as an invasive medical procedure that yields private health information and should therefore be subject to all relevant laws pertaining to patient consent, risk / benefit assessment, and the protection of privacy. In all other settings, it seems obvious that laws governing brain decoding should require informed consent and, where necessary, parental guidance.

References

[1] Paul S. Scotti, Mihir Tripathy, Cesare Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq Mathew Abraham. Mindeye2: shared-subject models enable fmri-to-image with 1 hour of data. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[2] Reese Kneeland, Paul S. Scotti, Ghislain St-Yves, Jesse Breedlove, Kendrick Kay, and Thomas Naselaris. 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)*, pages 28852–28862, June 2025.

[3] Stephen M Kosslyn, William L Thompson, and Giorgio Ganis. *The case for mental imagery*. Oxford University Press, 2006.

[4] Mark Stokes, Russell Thompson, Rhodri Cusack, and John Duncan. Top-Down Activation of Shape-Specific Population Codes in Visual Cortex during Mental Imagery. *Journal of Neuroscience*, 29(5):1565–1572, February 2009. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.4657-08.2009. URL https://www.jneurosci.org/content/29/5/1565. Publisher: Society for Neuroscience Section: Articles.

[5] Rainer Goebel, Rick van Hoof, Salil Bhat, Michael Lührs, and Mario Senden. Reading Imagined Letter Shapes from the Mind's Eye Using Real-time 7 Tesla fMRI. In *2022 10th International Winter Conference on Brain-Computer Interface (BCI)*, pages 1–3, February 2022. doi: 10.1109/BCI53720.2022.9735031. ISSN: 2572-7672.

[6] Thomas Naselaris, Cheryl A. Olman, Dustin E. Stansbury, Kamil Ugurbil, and Jack L. Gallant. A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *NeuroImage*, 105:215–228, January 2015. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2014.10.018. URL https://www.sciencedirect.com/science/article/pii/S1053811914008428.

[7] Leila Reddy, Naotsugu Tsuchiya, and Thomas Serre. Reading the mind's eye: Decoding category information during mental imagery. *NeuroImage*, 50(2):818–825, April 2010. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2009.11.084. URL https://www.sciencedirect.com/science/article/pii/S1053811909012701.

[8] Sue-Hyun Lee, Dwight J Kravitz, and Chris I Baker. Disentangling visual imagery and perception of real-world objects. *Neuroimage*, 59(4):4064–4073, 2012.

[9] Joel Pearson. The human imagination: the cognitive neuroscience of visual mental imagery. *Nature reviews neuroscience*, 20(10):624–634, 2019.

[10] Tiasha Saha Roy, Jesse Breedlove, Ghislain St-Yves, Kendrick Kay, and Thomas Naselaris. Comparison of signal to noise in vision and imagery for qualitatively different kinds of stimuli. *Journal of Vision*, 23(9):5961, 2023. ISSN 1534-7362. doi: 10.1167/jov.23.9.5961. URL https://doi.org/10.1167/jov.23.9.5961.

[11] Tiasha Saha Roy, Jesse Breedlove, Ghislain St-Yves, Kendrick Kay, and Thomas Naselaris. Mental imagery: Weak vision or compressed vision? In *Conference on Cognitive Computational Neuroscience*, 2023. doi: 10.32470/CCN.2023.1693-0. URL https://2023.ccneuro.org/view_paper4eea.html?PaperNum=1693.

[12] Jesse L. Breedlove, Ghislain St-Yves, Cheryl A. Olman, and Thomas Naselaris. Generative feedback explains distinct brain activity codes for seen and mental images. *Current Biology*, 30 (12):2211–2224.e6, 2020. ISSN 0960-9822. doi: https://doi.org/10.1016/j.cub.2020.04.014. URL https://www.sciencedirect.com/science/article/pii/S0960982220304942.

[13] Serra E Favila, Brice A Kuhl, and Jonathan Winawer. Spatial perception and memory have distinct activation profiles in human visual cortex. *BioRxiv*, page 811331, 2019.

[14] Radoslaw M Cichy, Jakob Heinzle, and John-Dylan Haynes. Imagery and perception share cortical representations of content and location. *Cerebral cortex*, 22(2):372–380, 2012.

[15] Anke Marit Albers, Peter Kok, Ivan Toni, H Chris Dijkerman, and Floris P De Lange. Shared representations for working memory and mental imagery in early visual cortex. *Current Biology*, 23(15):1427–1431, 2013.

[16] Ghislain St-Yves, Jesse Breedlove, Kendrick Kay, and Thomas Naselaris. Do better models of fmri visual response better predict mental imagery responses? In *Conference on Cognitive Computational Neuroscience*, 2023. doi: 10.32470/CCN.2023.1644-0. URL https://2023.ccneuro.org/view_paper37c6.html?PaperNum=1644.

[17] Bertrand Thirion, Edouard Duchesnay, Edward Hubbard, Jessica Dubois, Jean-Baptiste Poline, Denis Lebihan, and Stanislas Dehaene. Inverse retinotopy: Inferring the visual content of images from brain activation patterns. *NeuroImage*, 33(4):1104–1116, December 2006. ISSN 10538119. doi: 10.1016/j.neuroimage.2006.06.062. URL https://linkinghub.elsevier.com/retrieve/pii/S1053811906007373.

[18] Mario Senden, Thomas C. Emmerling, Rick van Hoof, Martin A. Frost, and Rainer Goebel. Reconstructing imagined letters from early visual cortex reveals tight topographic correspondence between visual mental imagery and perception. *Brain Structure and Function*, 224(3): 1167–1183, Jan 2019. doi: 10.1007/s00429-019-01828-6.

[19] Hongmi Lee and Brice A. Kuhl. Reconstructing perceived and retrieved faces from activity patterns in lateral parietal cortex. *Journal of Neuroscience*, 36(22):6069–6082, 2016. Publisher: Soc Neuroscience.

[20] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLOS Computational Biology*, 15(1):e1006633, January 2019. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006633. URL https://dx.plos.org/10.1371/journal.pcbi.1006633.

[21] Naoko Koide-Majima, Shinji Nishimoto, and Kei Majima. Mental image reconstruction from human brain activity: Neural decoding of mental imagery via deep neural network-based bayesian estimation. *Neural Networks*, 170:349–363, 2024. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2023.11.024. URL https://www.sciencedirect.com/science/article/pii/S0893608023006470.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752, 2021. URL https://arxiv.org/abs/2112.10752.

[24] Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, January 2022. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-021-00962-x. URL https://www.nature.com/articles/s41593-021-00962-x.

[25] Yu Takagi and Shinji Nishimoto. 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*, pages 14453–14463, 2023.

[26] Yu Takagi and Shinji Nishimoto. Improving visual image reconstruction from human brain activity using latent diffusion models via multiple decoded inputs, 2023.

[27] Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, 13, 2023. URL https://api.semanticscholar.org/CorpusID:260439960.

[28] Paul Steven Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Cohen Ethan, Aidan James Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, and Tanishq Mathew Abraham. Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=rwrblCYb2A.

[29] Reese Kneeland, Jordyn Ojeda, Ghislain St-Yves, and Thomas Naselaris. Reconstructing seen images from human brain activity via guided stochastic search. In *Conference on Cognitive Computational Neuroscience*, 2023. doi: 10.32470/CCN.2023.1672-0. URL https://2023.ccneuro.org/view_paper1337.html?PaperNum=1672.

[30] Reese Kneeland, Jordyn Ojeda, Ghislain St-Yves, and Thomas Naselaris. Second Sight: Using brain-optimized encoding models to align image distributions with human brain activity, June 2023. URL http://arxiv.org/abs/2306.00927. arXiv:2306.00927 [cs, q-bio].

[31] Reese Kneeland, Jordyn Ojeda, Ghislain St-Yves, and Thomas Naselaris. Brain-optimized inference improves reconstructions of fMRI brain activity, December 2023. URL http://arxiv.org/abs/2312.07705. arXiv:2312.07705 [cs, q-bio].

[32] Matteo Ferrante, Tommaso Boccato, Furkan Ozcelik, Rufin VanRullen, and Nicola Toschi. Through their eyes: multi-subject brain decoding with simple alignment techniques. *Imaging Neuroscience*, 2, 04 2024. doi: 10.1162/imag_a_00170.

[33] Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22710–22720, 2022. URL https://api.semanticscholar.org/CorpusID:253510456.

[34] Jingyuan Sun, Mingxiao Li, Zijiao Chen, Yunhao Zhang, Shaonan Wang, and Marie-Francine Moens. Contrast, Attend and Diffuse to Decode High-Resolution Images from Brain Activities, December 2023. URL http://arxiv.org/abs/2305.17214. arXiv:2305.17214 [cs].

[35] Weijian Mai and Zhijun Zhang. UniBrain: Unify Image Reconstruction and Captioning All in One Diffusion Model from Human Brain Activity, August 2023. URL http://arxiv.org/abs/2308.07428. arXiv:2308.07428 [cs].

[36] Rewon Child. Very deep {vae}s generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=RLRXCV6DbEJ.

[37] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

[38] Xingqian Xu, Zhangyang Wang, Eric Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7720–7731, 2022. URL https://api.semanticscholar.org/CorpusID:253523371.

[39] Aria Y. Wang, Kendrick Kay, Thomas Naselaris, Michael J. Tarr, and Leila Wehbe. Incorporating natural language into vision models improves prediction and understanding of higher visual cortex, September 2022. URL https://www.biorxiv.org/content/10.1101/2022.09.27.508760v1. Pages: 2022.09.27.508760 Section: New Results.

[40] Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, Jan 1992. doi: 10.1162/neco.1992.4.1.1.

[41] Aria Y. Wang, Kendrick Kay, Thomas Naselaris, Michael J. Tarr, and Leila Wehbe. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, 5(12):1415–1426, December 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00753-y. Publisher Copyright: © 2023, The Author(s), under exclusive licence to Springer Nature Limited.

[42] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[43] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, June 2024.

[44] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

[45] Pablo Pernias, Dominic Rampas, Mats Leon Richter, Christopher Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gU58d5QeGv.

[46] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. GIT: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=b4tMhpN0JC.

[47] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Image synthesis and editing with stochastic differential equations. *CoRR*, abs/2108.01073, 2021. URL https://arxiv.org/abs/2108.01073.

[48] Matteo Ferrante, Furkan Ozcelik, Tommaso Boccato, Rufin VanRullen, and Nicola Toschi. Brain Captioning: Decoding human brain activity into images and text, May 2023. URL http://arxiv.org/abs/2305.11560. arXiv:2305.11560 [cs].

[49] Emily A Holmes and Andrew Mathews. Mental imagery in emotion and emotional disorders. *Clinical psychology review*, 30(3):349–362, 2010.

[50] Joseph T. Giacino and Kathleen Kalmar. The vegetative and minimally conscious states: A comparison of clinical features and functional outcome. *Journal of Head Trauma Rehabilitation*, 12(4):36–51, 1997. doi: 10.1097/00001199-199708000-00005.

[51] Brian L Edlow, Camille Chatelle, Camille A. Spencer, Catherine J. Chu, Yelena G. Bodien, Kathryn L. O'Connor, Ronald E. Hirschberg, Leigh R. Hochberg, Joseph T. Giacino, Eric S. Rosenthal, and et al. Early detection of consciousness in patients with acute severe traumatic brain injury. *Brain*, 140(9):2399–2414, 2017. doi: 10.1093/brain/awx176.

[52] Alexis F. Turgeon, François Lauzier, Jean-François Simard, Damon C. Scales, Karen E.A. Burns, Lynne Moore, David A. Zygun, Francis Bernard, Maureen O. Meade, Tran Cong Dung, and et al. Mortality associated with withdrawal of life-sustaining therapy for patients with severe traumatic brain injury: A canadian multicentre cohort study. *Canadian Medical Association Journal*, 183(14):1581–1588, 2011. doi: 10.1503/cmaj.101786.

[53] Livia Livinț Popa, Diana Chira, Ștefan Strilciuc, and Dafin F. Mureșanu. Non-invasive systems application in traumatic brain injury rehabilitation. *Brain Sciences*, 13(11), 2023. ISSN 2076-3425. doi: 10.3390/brainsci13111594. URL https://www.mdpi.com/2076-3425/13/11/1594.

[54] Shiyu Luo, Qinwan Rabbani, and Nathan E. Crone. Brain-computer interface: Applications to speech decoding and synthesis to augment communication. *Neurotherapeutics*, 19(1):263–273, Jan 2022. doi: 10.1007/s13311-022-01190-2.

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

[56] Emma C. Gordon and Anil K. Seth. Ethical considerations for the use of brain–computer interfaces for cognitive enhancement. *PLOS Biology*, 22(10):1–15, 10 2024. doi: 10.1371/journal.pbio.3002899. URL https://doi.org/10.1371/journal.pbio.3002899.

[57] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. ISSN 00401706. URL http://www.jstor.org/stable/1267351.

[58] Pawan Sinha and Richard Russell. A perceptually based comparison of image similarity metrics. *Perception*, 40(11):1269–1281, 2011. doi: 10.1068/p7063. URL https://doi.org/10.1068/p7063. PMID: 22416586.

[59] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=G5RwHpBUv0.

[60] Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Mindbridge: A cross-subject brain decoding framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11333–11342, 2024.

[61] Jingyang Huo, Yikai Wang, Xuelin Qian, Yun Wang, Chong Li, Jianfeng Feng, and Yanwei Fu. Neuropictor: Refining fmri-to-image reconstruction via multi-individual pretraining and multi-level modulation, 2024.

[62] Dian Xie, Peiang Zhao, Jiarui Zhang, Kangqi Wei, Xiaobao Ni, and Jiong Xia. Brainram: Cross-modality retrieval-augmented image reconstruction from human brain activity. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 3994–4003, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681296. URL https://doi.org/10.1145/3664647.3681296.

[63] Zhou 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*, 13(4):600–612, April 2004. ISSN 1941-0042. doi: 10.1109/TIP.2003.819861. Conference Name: IEEE Transactions on Image Processing.

[64] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[65] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015. URL http://arxiv.org/abs/1512.00567.

[66] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 2019. URL http://proceedings.mlr.press/v97/tan19a.html.

[67] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *CoRR*, abs/2006.09882, 2020. URL https://arxiv.org/abs/2006.09882.

[68] Ghislain St-Yves, Emily J. Allen, Yihan Wu, Kendrick Kay, and Thomas Naselaris. Brain-optimized deep neural network models of human visual areas learn non-hierarchical representations. *Nature Communications*, 14(1):3329, 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-38674-4. URL https://doi.org/10.1038/s41467-023-38674-4.

[69] Ken Shirakawa, Yoshihiro Nagano, Misato Tanaka, Shuntaro C. Aoki, Kei Majima, Yusuke Muraki, and Yukiyasu Kamitani. Spurious reconstruction from brain activity: The thin line between reconstruction, classification, and hallucination. *Journal of Vision*, 2024. URL https://api.semanticscholar.org/CorpusID:269791182.

[70] David Mayo, Christopher Wang, Asa Harbin, Abdulrahman Alabdulkareem, Albert Eaton Shaw, Boris Katz, and Andrei Barbu. Brainbits: How much of the brain are generative reconstruction methods using? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=KAAUvi4kpb.

A Appendix

A.1 Limitations

The NSD-Imagery dataset utilized as a validation set for this work contains only 18 stimuli, and thus does not allow for large-scale model training or fine-tuning of existing vision decoding models, necessitating a cross-decoding approach. Future models for downstream applications would surely be improved by training or fine-tuning on mental imagery datasets. Such datasets must be a priority for research in this space. Currently, the computational requirements to run these models are an obstacle to realizing medical applications. We trained our ridge regression modules on computing hardware with 512GB of RAM and performed inference for our models using an NVIDIA A100 with 40GB of VRAM. The requirement of such hardware limits the use of our method to researchers with considerable compute resources.

A.2 Ablation Study

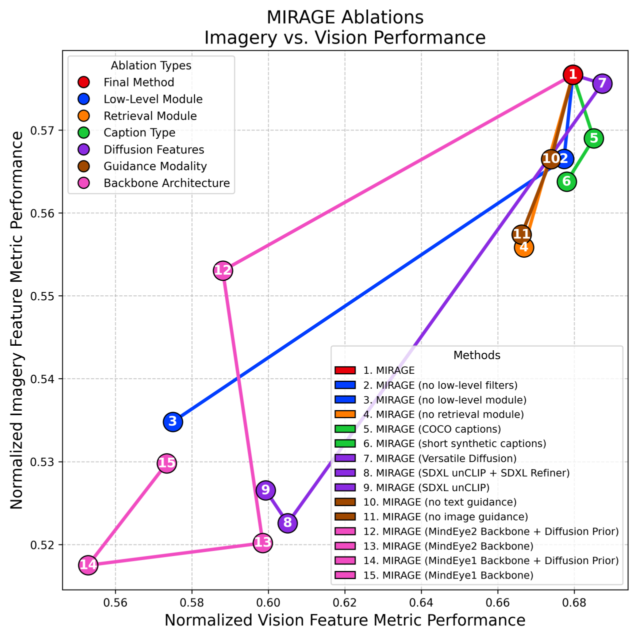

Figure 6: Ablation analyses: model variants (numbered circles) under each ablation type (color) are assessed via the normalized average of all feature metrics (Table 1), with vision on the x-axis and imagery on the y-axis.

We systematically ablated model components to identify which were most important for mental imagery reconstruction. Colored numeric identifiers refer to Figure **??**. **MIRAGE** is identified as (1).

Low-Level Module We found that despite the relatively low spatial resolution of mental imagery, the low-level decoding module provides much of the needed structure to accurately reconstruct the target stimulus (3), and is boosted measurably by our image filtering technique (2).

Retrieval Module We see a notable boost from our retrieval procedure (4), demonstrating that high-dimensional image embeddings can be decoded and used to guide a retrieval module to improve performance, even if those same embeddings are not suitable to drive an image generator.

Caption Types We examine the results of using COCO captions (5), as well as a set of synthetic "short" captions (6) that aim to replicate the length of the COCO captions (the median word count is 8 vs 10 for COCO), and evaluate the quality of captions provided by LLaVA v1.5-13B. These short synthetic captions did not provide a boost over COCO, but the longer synthetic captions (average word count $= 34$) utilized in our final method (1) provided an improvement.

Diffusion Features We replaced the 1×768 image embeddings and 77×1280 text embeddings used to drive Stable Cascade with the features used to drive Versatile Diffusion in MindEye1 and Brain Diffuser (7) [38] (257×768 image embedding, 77×768 text embedding), and SDXL unCLIP [1], the embedding used to drive MindEye2 (8,9) (257×1664 image embedding). Our smaller image embedding and larger text embedding provided the most robust guidance for reconstructing mental images (1).

Guidance Modality For both vision and mental imagery, image-only guidance (10) to Stable Cascade afforded better performance than text-only guidance (11). In both cases, but especially for mental imagery, the combination of text and image guidance (1) provided a large performance boost, suggesting that multimodal guidance plays an important role for mental imagery reconstruction.

Backbone Architecture The MLP backbone and diffusion prior architectures of MindEye1 (14,15) and MindEye2 (12,13) perform worse than our ridge regression backbone (1) for both vision and mental imagery, suggesting that these architectures tend to overfit to the training data in the core NSD experiment. Evidently, the potential for increased expressivity afforded by an MLP is a disadvantage when attempting to generalize vision reconstruction performance to new sessions and stimulus types, or to mental imagery.

A.3 Ridge regression backbone implementation details

We chose to use a ridge regression decoding backbone due to its effectiveness in high-dimensional, low-SNR settings [57], and because its singularly effective for aligning noisy brain activity patterns across individual brains [32]. For each set of features, we train a parallel ridge regression model to predict the feature value from our fMRI responses, minimizing the loss function in:

$$\mathcal{L}(\mathbf{w}, \mathbf{b}) = \|\mathbf{X}\mathbf{w} + \mathbf{b} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_2^2 \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{n \times v}$ is the fMRI data matrix of n fMRI trials by v voxels, $\mathbf{y} \in \mathbb{R}^{n \times d}$ is the matrix of target features d for each trial n, $\mathbf{w} \in \mathbb{R}^{v \times d}$ is the weight vector mapping fMRI voxels to the feature dimension d, $\mathbf{b} \in \mathbb{R}^{d}$ is the bias vector added to each trial n, and λ is the ridge parameter controlling the strength of the L_2 penalty.

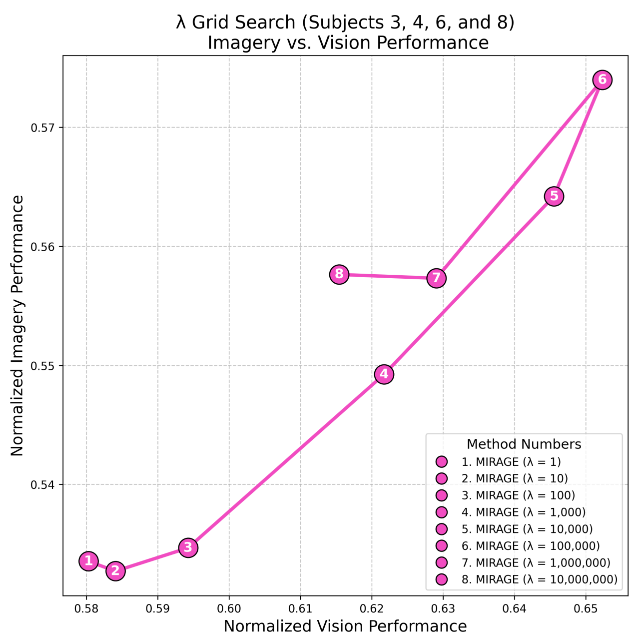

Figure 7: Hyperparameter logarithmic grid search over possible values of λ for use in Equation 1 (Section 3.3. Metrics are the normalized average of all metrics in Table 1, with imagery performance on the Y axis and vision on the X axis.

To select the optimal L_2 weight decay parameter λ for generalizing to fMRI responses of mental images in our regression backbone detailed in Equation 1 of Section 3.3, we performed a logarithmic grid search over possible values of λ on subjects $3, 4, 6$, and 8 of the NSD-Imagery dataset, which we use a hyperparameter tuning set. From this analysis, we select $\lambda = 100,000$.

A.4 VDVAE implementation

For decoding the low-level image in our pipeline, we utilize the Very Deep Variational Autoencoder (VDVAE) model introduced in Child [36]. VDVAEs are generative models that learn to represent an input distribution—such as an image dataset—through a low-dimensional latent space constrained by a predefined prior distribution, typically Gaussian. The VDVAE utilizes a hierarchical structure with multiple layers of conditionally dependent latent variables organized hierarchically, with each layer capturing different levels of detail from coarse to fine as one moves from the top to the bottom of the hierarchy. Appendix Eq. (1) shows that each latent variable depends on the preceding ones and the input x, whereas Appendix Eq. (2) allows for sample generation from the prior distribution without input data.

$$q_\phi(z|x) = q_\phi(z_0|x)q_\phi(z_1|z_0, x)...q_\phi(z_N|z_{<N}, x) \tag{2}$$

$$p_\theta(z) = p_\theta(z_0)p_\theta(z_1|z_0)...p_\theta(z_N|z_{<N}) \tag{3}$$

In our approach, we utilize the VDVAE model [36] trained on the ImageNet dataset at a resolution of 64×64 pixels and consisting of 75 hierarchical layers. We use the latent variables from the first 31 layers, as including additional layers does not yield significant improvements in reconstruction quality. In the testing phase, our predicted latents for the first 31 layers are concatenated with the remaining 44 layers sampled from Eq. (2), and passed through the latent-to-pixel decoder module of the VDVAE to generate reconstructed images at 64×64 pixel resolution.

We observed that when decoding low-level images from fMRI responses to mental images, often the low-level reconstructions have significantly deteriorated features and reduced contrast. To counteract this and boost the influence of our structural layout on subsequent stages of the reconstruction process, we apply a set of image filters to boost the sharpness and contrast of the low-level images. These reconstructions serve as an initial estimate for the diffusion model employed in the subsequent stage of our pipeline.

A.5 Best case reconstructions

Figure 8: Qualitative comparison of reconstruction methods on stimuli seen during the vision trials of NSD-Imagery. Samples selected are the best scoring according to the reconstruction metrics in Table 1 of the manuscript.

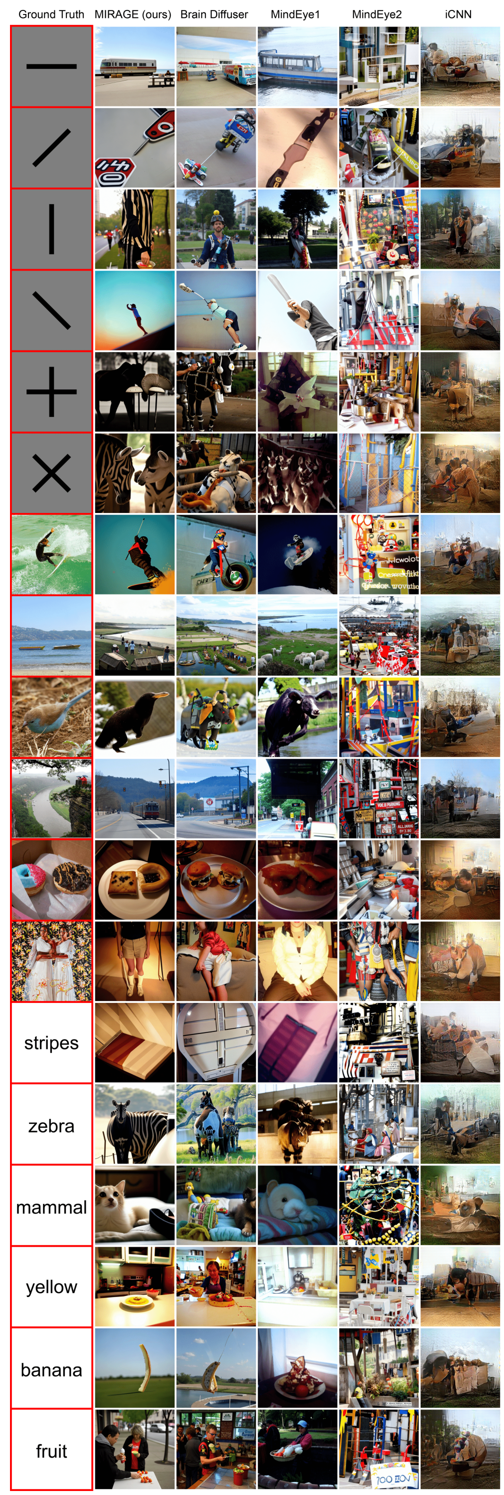

Figure 9: Qualitative comparison of reconstruction methods on imagined stimuli from NSD-Imagery. Samples selected are the best scoring according to the reconstruction metrics in Table 1 of the manuscript.

A.6 Median and worst-case reconstructions

Figure 10: Median-case vision reconstructions from the vision trials of NSD-Imagery. Samples selected as median scoring based on metrics in Table 1 of the manuscript.

Figure 11: Median-case imagery reconstructions from the imagery trials of NSD-Imagery. Samples selected as in Figure 10.

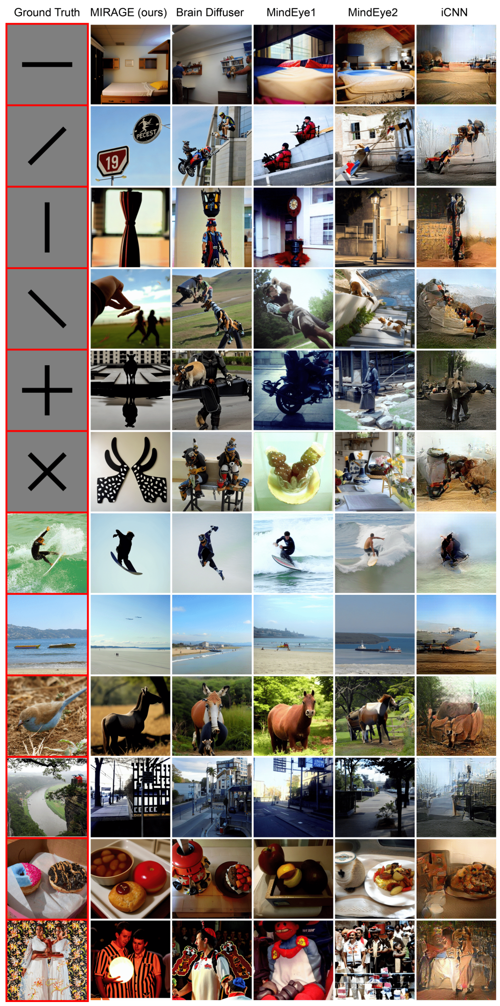

Figure 12: Worst-case vision reconstructions from the vision trials of NSD-Imagery. Samples selected as lowest scoring based on metrics in Table 1 of the manuscript.

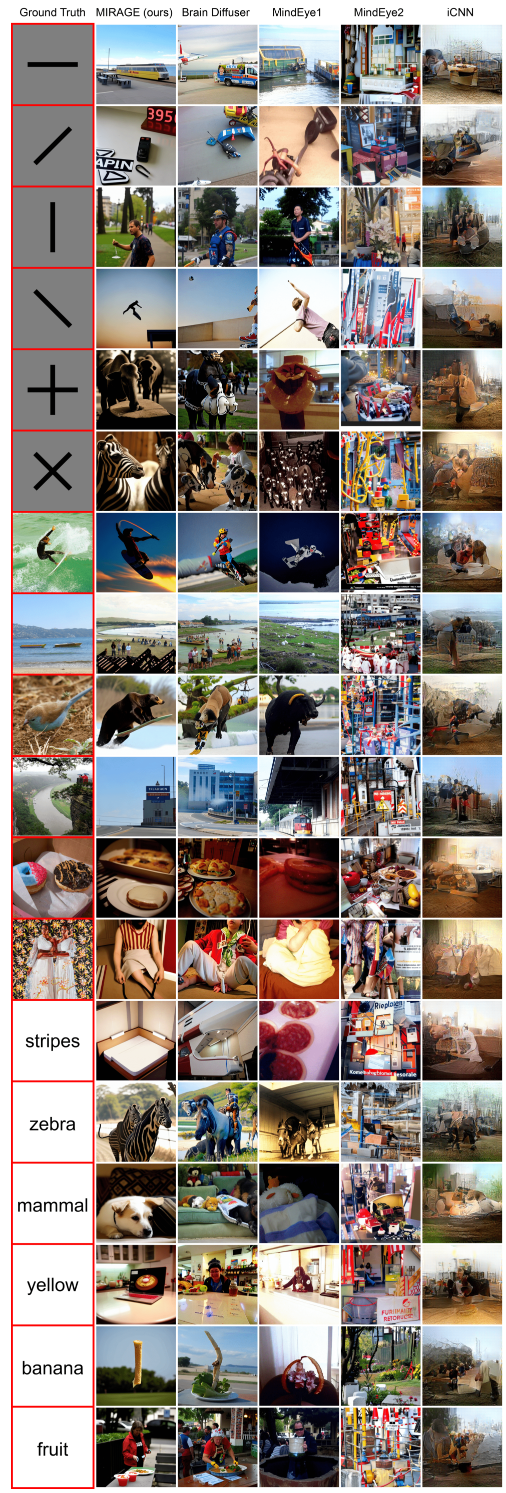

Figure 13: Worst-case imagery reconstructions from the imagery trials of NSD-Imagery. Samples selected as in Figure 12.

A.7 Reconstructions from additional methods on NSD-Imagery

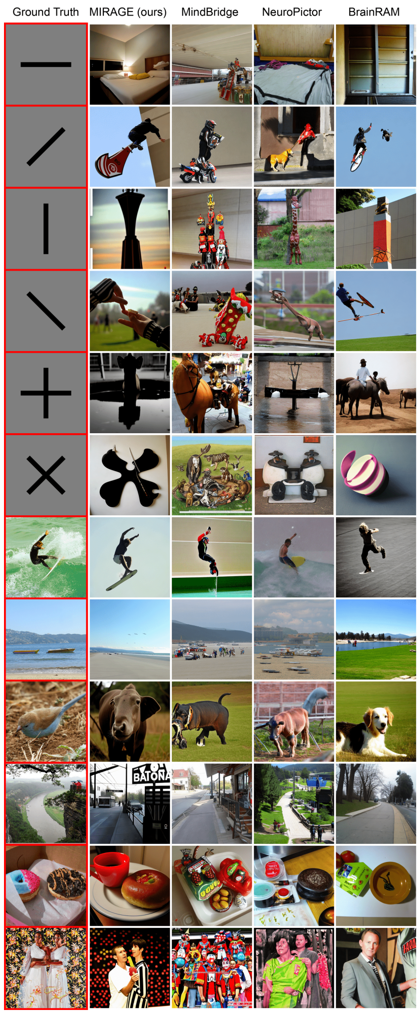

Figure 14: Best-case vision reconstructions (additional methods) from vision trials of NSD-Imagery. Samples selected as highest scoring based on metrics in Table 1 of the manuscript.

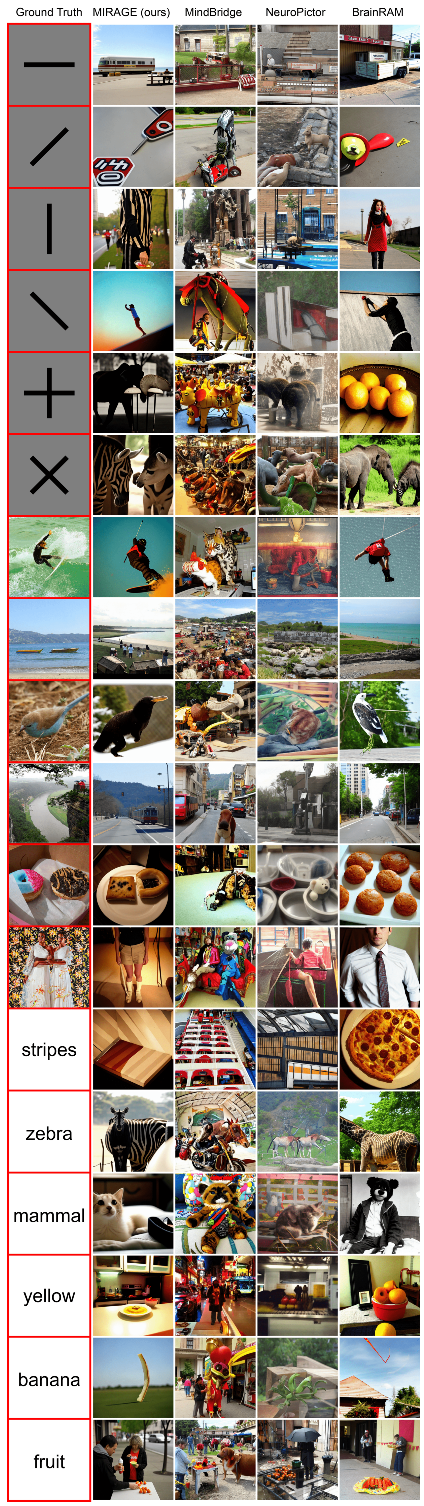

Figure 15: Best-case imagery reconstructions (additional methods) from imagery trials of NSD-Imagery. Samples selected as in Figure 14.

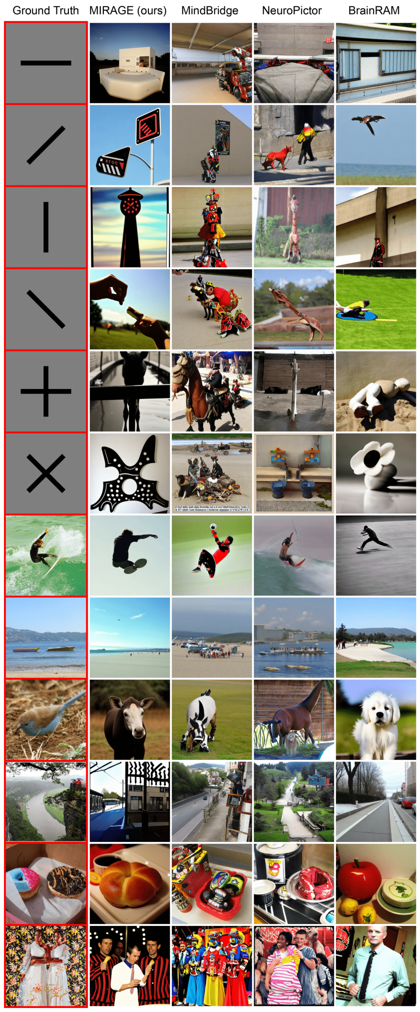

Figure 16: Median-case vision reconstructions (additional methods) from vision trials of NSD-Imagery. Samples selected as median scoring based on metrics in Table 1 of the manuscript.

Figure 17: Median-case imagery reconstructions (additional methods) from imagery trials of NSD-Imagery. Samples selected as in Figure 16.

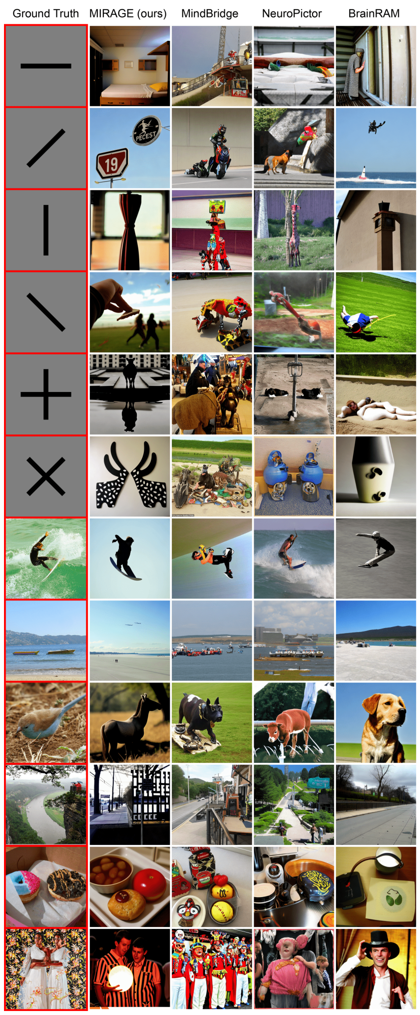

Figure 18: Worst-case vision reconstructions (additional methods) from vision trials of NSD-Imagery. Samples selected as lowest scoring based on metrics in Table 1 of the manuscript.

Figure 19: Worst-case imagery reconstructions (additional methods) from imagery trials of NSD-Imagery. Samples selected as in Figure 18.

A.8 Image feature metrics

We provide performance benchmarks against existing methods evaluated on NSD-Imagery [2] using the metrics in Table 1. For all methods, we output 10 reconstructions per test sample from each method and report averaged metrics across them. Metrics can fluctuate due to the stochastic nature of fMRI-to-image models, and this averaging step increases the reliability of results. Across the majority of metrics, **MIRAGE** shows SOTA performance on mental image reconstructions. We note that although these metrics are often used as a proxy for human judgment, many research efforts have established that these metrics do not closely approximate or align with human assessments of content [58] or quality [59]. We observe that they are particularly volatile with a dataset as small as NSD-Imagery. For this reason, we provide extensive behavioral evaluations of our results by human raters in Section 4.1. Results of our method on the NSD shared1000 test set are also provided in Appendix A.11.

Method	Low-Level				High-Level				Brain Correlation			Captions			
	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓	Early Vis. ↑	Higher Vis. ↑	Visual Cortex ↑	ROUGE-L ↑	METEOR ↑	Sentence ↑	CLIP-L ↑
Mental Imagery Reconstructions															
MIRAGE (ours)	0.104	0.398	**63.92%**	**62.46%**	52.25%	**57.46%**	**0.914**	0.575	**0.204**	0.142	**0.168**	**0.154**	**0.095**	**0.176**	0.469
MindEye1 [28]	0.086	0.349	59.56%	61.00%	52.03%	54.72%	0.948	0.564	0.180	0.135	0.155	–	–	–	–
Brain Diffuser [27]	0.064	0.401	52.14%	58.35%	**52.73%**	54.07%	0.935	0.585	0.133	0.127	0.141	–	–	–	–
iCNN [20]	**0.108**	0.340	50.57%	55.25%	49.39%	41.72%	0.994	**0.560**	0.113	0.062	0.081	–	–	–	–
MindEye2 [1]	0.036	**0.414**	47.60%	55.38%	46.02%	50.78%	0.966	0.591	0.069	0.055	0.061	0.143	0.080	0.162	**0.484**
MindBridge [60]	0.030	0.200	44.13%	52.46%	45.55%	49.70%	0.980	0.627	0.079	0.079	0.075	–	–	–	–
NeuroPictor [61]	0.022	0.305	43.18%	44.85%	44.15%	46.40%	0.994	0.612	0.084	0.054	0.140	–	–	–	–
BrainRAM [62]	0.056	0.372	51.78%	56.29%	52.52%	53.73%	0.927	0.577	0.139	**0.145**	0.112	–	–	–	–
Vision Reconstructions															
MIRAGE (ours)	0.221	0.442	**79.03%**	76.57%	**69.75%**	**66.69%**	**0.879**	0.546	0.363	**0.262**	0.316	0.157	0.105	0.216	0.469
MindEye1 [28]	0.218	0.412	73.56%	80.81%	62.44%	65.34%	0.881	**0.510**	0.374	0.253	0.311	–	–	–	–
Brain Diffuser [27]	0.107	0.455	60.34%	72.84%	60.95%	58.31%	0.908	0.555	0.247	0.229	0.255	–	–	–	–
iCNN [20]	**0.224**	0.385	71.67%	**81.35%**	61.16%	49.03%	0.926	0.524	**0.442**	0.246	0.338	–	–	–	–
MindEye2 [1]	0.161	**0.480**	70.10%	77.52%	62.69%	65.93%	0.886	0.512	0.352	0.237	0.290	**0.171**	**0.118**	**0.249**	**0.515**
MindBridge [60]	0.117	0.352	58.47%	70.76%	58.83%	64.49%	0.915	0.565	0.245	0.227	0.232	–	–	–	–
NeuroPictor [61]	0.055	0.364	62.27%	66.42%	49.92%	53.49%	0.949	0.571	0.272	0.192	**0.363**	–	–	–	–
BrainRAM [62]	0.097	0.409	63.11%	67.99%	58.50%	59.79%	0.894	0.530	0.215	0.226	0.175	–	–	–	–

Table 1: Quantitative comparison between reconstruction methods for both imagery and vision trials on simple and complex stimuli (conceptual stimuli have no ground truth images). PixCorr is the pixel-level correlation score. SSIM is the structural similarity index metric [63]. AlexNet(2) and AlexNet(5) are the 2-way comparisons (2WC) of layers 2 and 5 of AlexNet [64]. CLIP is the 2WC of the output layer of the CLIP ViT-L/14 Vision model [22]. Inception is the 2WC of the last pooling layer of InceptionV3 [65]. EffNet-B and SwAV are distance metrics gathered from EfficientNet-B13 [66] and SwAV-ResNet50 [67] models. Each brain correlation score is calculated using voxels from within the respective regions of the visual cortex. For EffNet-B and SwAV distances, lower is better. For all other metrics, higher is better. Bold indicates best performance, and underlines second-best performance. Additional details on the metrics used, including explanations of 2-way comparisons and brain correlation scores, are in Appendix A.9. A breakdown of model performance across the different types of stimuli is in Appendix A.12. Details for our implementation of iCNN are provided in Appendix A.18.

A.9 Additional evaluation metric details

For the metrics in Table 1, a two-way comparison evaluates whether the feature embedding of the stimulus image is more similar to the feature embedding of the target reconstruction, or the feature embedding of a randomly selected "distractor" reconstruction. Two-way identification refers to percent correct across a set of two-way comparisons performed on a pool of distractor images. The two-way identification metrics we report, which are calculated using reconstructions of the 11 other NSD-Imagery stimuli as distractors, are notably different from the two-way identification metrics presented in individual reconstruction papers that perform evaluations using reconstructions of the shared1000 as the pool of distractors. The pool of distractor images for NSD-Imagery is much smaller, and contains multiple distinct types of stimuli that may significantly alter the resulting identification accuracy metrics. Because of this difference, the two-way identification accuracy numbers are not directly comparable to two-way identification results evaluated on the shared1000 in our work or in other papers. Brain correlation scores are the Pearson correlation between the averaged measured brain response β and the predicted brain response β' produced by a brain encoding model (GNet [68]) averaged across voxels within a respective ROI in visual cortex, including the whole visual cortex, early visual cortical regions (V1, V2, V3, and V4), and higher visual areas (set complement of visual cortex and early visual cortex). All metrics in Tables 1, Appendix Table 4, and Appendix Table5 were calculated and averaged across 10 images sampled from the output distribution of each method using a random seed. The caption metrics were computed against the ground truth image captions provided with the NSD-Imagery dataset. Metrics in Table 1 of the manuscript are the original values reported in each of the respective papers, except for the iCNN method, which has never been benchmarked on the NSD shared1000 test set and so results reported are from our reproduction of the method utilizing the author's open source code. For the results from our method (MIRAGE) in the table, we compute values across 5 output repetitions sampled from the posterior of our method, and average those values together for the table.

A.10 Statistical significance of metrics

Method	Low-Level				High-Level				Brain Correlation		
	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓	Early Vis. ↑	Higher Vis. ↑	Visual Cortex ↑
Mental Imagery Reconstructions											
MIRAGE	±0.0061	±0.0091	±0.89%	±1.06%	±1.27%	±1.24%	±0.0044	±0.0040	±0.0075	±0.0067	±0.0059
MindEye1	±0.0082	±0.0082	±1.07%	±0.72%	±1.43%	±1.35%	±0.0064	±0.0047	±0.0073	±0.0075	±0.0069
Brain Diffuser	±0.0068	±0.0087	±1.40%	±1.07%	±1.47%	±1.46%	±0.0053	±0.0041	±0.0086	±0.0079	±0.0070
iCNN	±0.0081	±0.0055	±0.93%	±0.54%	±1.42%	±1.21%	±0.0041	±0.0023	±0.0055	±0.0074	±0.0056
MindEye2	±0.0085	±0.0079	±1.07%	±0.87%	±1.41%	±1.27%	±0.0067	±0.0051	±0.0073	±0.0081	±0.0070
MindBridge	±0.0063	±0.0080	±1.38%	±1.11%	±1.45%	±1.24%	±0.0053	±0.0045	±0.0079	±0.0079	±0.0071
NeuroPictor	±0.0062	±0.0072	±1.23%	±1.22%	±1.53%	±1.37%	±0.0046	±0.0037	±0.0087	±0.0080	±0.0072
BrainRAM	±0.0074	±0.0105	±1.30%	±1.21%	±1.49%	±1.26%	±0.0058	±0.0056	±0.0088	±0.0074	±0.0074
Vision Reconstructions											
MIRAGE	±0.0072	±0.0097	±1.23%	±1.21%	±1.56%	±1.41%	±0.0040	±0.0045	±0.0052	±0.0050	±0.0044
MindEye1	±0.0086	±0.0086	±1.35%	±1.31%	±1.53%	±1.53%	±0.0035	±0.0036	±0.0053	±0.0046	±0.0042
Brain Diffuser	±0.0052	±0.0082	±1.38%	±1.35%	±1.49%	±1.50%	±0.0040	±0.0038	±0.0055	±0.0051	±0.0045
iCNN	±0.0077	±0.0052	±1.24%	±1.26%	±1.40%	±1.40%	±0.0021	±0.0025	±0.0058	±0.0052	±0.0044
MindEye2	±0.0049	±0.0084	±1.45%	±1.43%	±1.60%	±1.51%	±0.0034	±0.0037	±0.0054	±0.0053	±0.0048
MindBridge	±0.0046	±0.0041	±1.43%	±1.48%	±1.54%	±1.48%	±0.0029	±0.0036	±0.0064	±0.0048	±0.0047
NeuroPictor	±0.0046	±0.0060	±1.40%	±1.54%	±1.47%	±1.44%	±0.0021	±0.0028	±0.0053	±0.0053	±0.0048
BrainRAM	±0.0063	±0.0087	±1.48%	±1.36%	±1.49%	±1.43%	±0.0044	±0.0047	±0.0053	±0.0048	±0.0044

Table 2: Standard error measurements for evaluation metrics of fMRI-to-Image reconstruction models evaluated on both the vision and mental imagery trials of NSD-Imagery. Values correspond to the standard error spread of values in Table 1 in the manuscript.

A.11 NSD test set feature metric evaluations

NSD Shared1000 Test Set	Low-Level				High-Level				Brain Correlation		
Method	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓	Early Vis. ↑	Higher Vis. ↑	Visual Cortex ↑
MIRAGE (ours)	0.285	0.361	94.30%	95.73%	91.18%	90.92%	0.732	0.473	0.337	0.371	0.372
MindEye1 [28]	0.319	0.360	92.49%	96.44%	<u>93.55%</u>	<u>92.14%</u>	<u>0.648</u>	<u>0.377</u>	0.350	<u>0.374</u>	0.378
Brain Diffuser [27]	0.273	<u>0.365</u>	<u>94.39%</u>	96.64%	91.28%	90.90%	0.728	0.421	<u>0.353</u>	**0.375**	<u>0.381</u>
iCNN [20]	<u>0.321</u>	0.336	94.33%	<u>97.09%</u>	90.46%	74.47%	0.797	0.528	**0.410**	0.371	**0.396**
MindEye2 [1]	**0.322**	**0.431**	**96.10%**	**98.61%**	**95.42%**	**92.98%**	**0.619**	**0.344**	<u>0.360</u>	0.368	0.373

Table 3: Quantitative comparison between reconstruction methods on the NSD Shared1000 Test Set. Metrics are the same as Table 1 of the manuscript.

A.12 Comparison of image feature metrics across stimuli types

Method	Low-Level				High-Level				Brain Correlation		
	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓	Early Vis. ↑	Higher Vis. ↑	Visual Cortex ↑
Mental Imagery Reconstructions (Simple Stimuli)											
MIRAGE (ours)	0.027	0.511	**53.11%**	**67.27%**	42.39%	60.30%	**0.939**	0.563	**0.224**	**0.118**	**0.164**
MindEye1 [28]	0.033	0.456	43.71%	61.67%	37.46%	58.37%	0.974	0.563	0.200	0.107	0.148
Brain Diffuser [27]	0.013	**0.524**	30.68%	50.68%	34.43%	44.51%	0.983	0.603	0.152	0.091	0.128
iCNN [20]	**0.063**	0.427	27.42%	47.65%	**45.11%**	**67.99%**	1.006	**0.546**	0.138	0.045	0.081
MindEye2 [1]	0.011	0.448	23.37%	45.34%	31.14%	49.02%	0.987	0.590	0.074	0.035	0.051
Vision Reconstructions (Simple Stimuli)											
MIRAGE (ours)	**0.159**	0.569	**74.24%**	**82.77%**	**56.78%**	63.71%	**0.913**	0.537	0.395	**0.174**	0.279
MindEye1 [28]	0.129	0.506	62.01%	76.36%	43.33%	60.64%	0.961	0.549	0.370	0.140	0.243
Brain Diffuser [27]	0.075	**0.586**	40.19%	66.67%	38.30%	42.20%	0.988	0.601	0.209	0.106	0.169
iCNN [20]	0.132	0.454	57.01%	74.89%	37.69%	**69.02%**	0.992	**0.534**	**0.447**	0.133	**0.278**
MindEye2 [1]	0.040	0.487	50.87%	68.98%	43.52%	52.46%	0.980	0.577	0.334	0.108	0.204

Table 4: Quantitative comparison between reconstruction methods for both imagery and vision trials on simple stimuli. Metrics are the same as Table 1

Method	Low-Level				High-Level				Brain Correlation		
	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓	Early Vis. ↑	Higher Vis. ↑	Visual Cortex ↑
Mental Imagery Reconstructions (Complex Stimuli)											
MIRAGE (ours)	**0.181**	0.285	74.74%	57.65%	62.121%	54.62%	0.888	0.587	**0.183**	**0.165**	**0.172**
MindEye1 [28]	0.138	0.243	**75.42%**	60.34%	66.591%	51.06%	0.921	**0.566**	0.159	0.164	0.161
Brain Diffuser [27]	0.114	**0.278**	73.60%	**66.02%**	**71.02%**	**63.64%**	**0.888**	0.567	0.114	0.163	0.154
iCNN [20]	0.153	0.253	73.71%	62.84%	53.674%	15.46%	0.982	0.575	0.089	0.079	0.081
MindEye2 [1]	0.032	0.231	70.42%	65.11%	61.97%	51.93%	0.943	0.601	0.062	0.074	0.068
Vision Reconstructions (Complex Stimuli)											
MIRAGE (ours)	0.282	0.315	83.83%	70.38%	82.727%	69.659%	0.845	0.555	0.331	0.350	0.353
MindEye1 [28]	0.308	0.318	85.11%	85.27%	81.55%	70.038%	0.800	0.471	0.378	0.365	0.379
Brain Diffuser [27]	0.139	**0.323**	80.49%	79.02%	83.60%	74.43%	0.829	0.509	0.284	0.353	0.341
iCNN [20]	**0.316**	0.316	**86.33%**	**87.80%**	**84.62%**	29.05%	0.860	0.514	**0.437**	0.358	**0.397**
MindEye2 [1]	0.223	0.333	84.28%	85.83%	80.08%	**77.46%**	**0.794**	**0.454**	0.378	**0.360**	0.376

Table 5: Quantitative comparison between reconstruction methods for both imagery and vision trials on complex stimuli. Metrics are the same as Table 1.

A.13 Impact of trial repetition averaging on performance

One of the experimental details that varies between NSD [24] and NSD-Imagery [2] is the number of times each stimulus was presented in the experiment, also called the number of trial repetitions. NSD contained 3 trial repetitions of each stimulus in both the training and test sets, while NSD-Imagery contains 8 trial repetitions for the vision task and 16 trial repetitions for the imagery task. In Figure 20, we plot the effect of these additional trial repetitions on the performance of MIRAGE relative to the other methods we compare against.

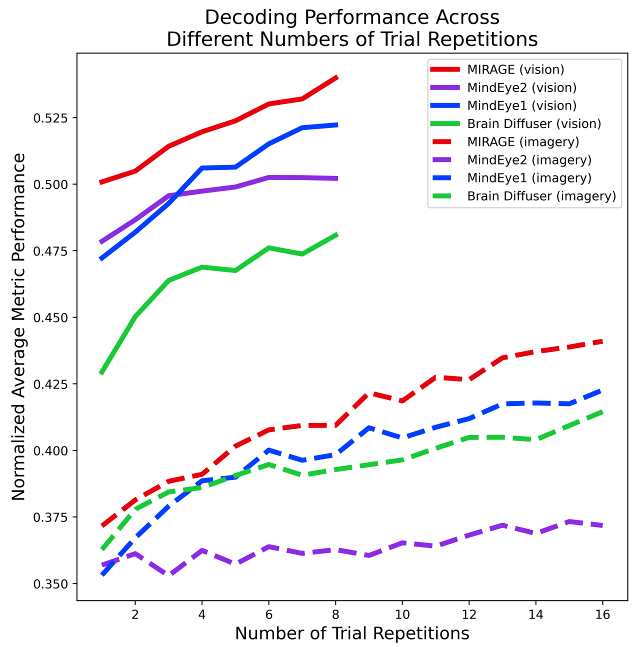

Figure 20: Performance of MIRAGE and other methods when averaging across brain activity responses to multiple trial repetitions of the same stimulus. Y-axis is the normalized average of all metrics in Table 1, X-axis is the number of averaged trial repetitions.

A.14 Impact of training data scale on performance

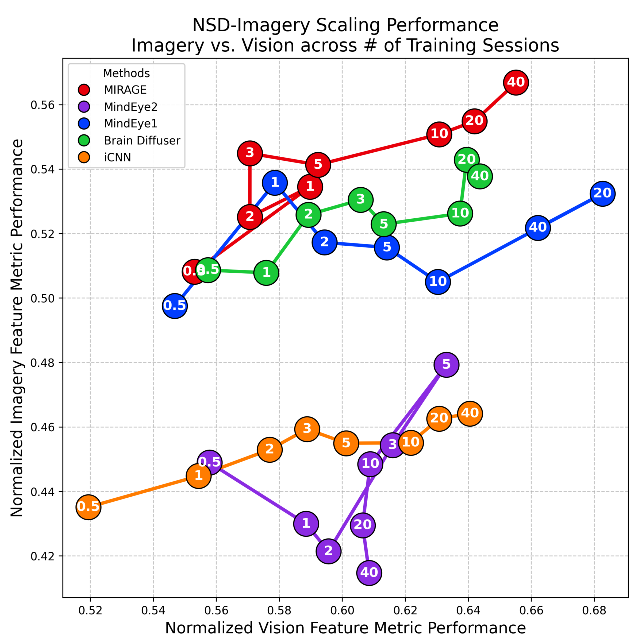

Figure 21: Performance of MIRAGE and other methods on NSD-Imagery for Subject 1 when trained on different numbers of fMRI sessions present in NSD. Each session includes approximately one hour of fMRI data. Metrics are the normalized average of all metrics in Table 1, with imagery performance on the Y axis and vision on the X axis. Methods are indicated by color, with the number of training sessions indicated by the numbers in each dot.

An additional challenge in deploying these fMRI-to-image decoding methods lies in making them more generalizable to new subjects. MIRAGE, along with all of the other methods examined in this paper, were trained with 40 hours of subject-specific fMRI data comprising $10,000$ unique stimuli. Collecting this much training data for new subjects in practical settings is currently impractical or impossible for certain clinical patients. Recent work in MindEye2 [1] has tackled this problem head-on by using a multi-subject pretraining step, however as evaluated in Figure 21, this technique generalizes poorly to mental imagery data. By contrast, MIRAGE outperforms all other methods for mental image reconstruction using only 3 hours of fMRI training data, and continues to scale robustly up to 40 sessions. We additionally note that the methods that used ridge regression decoding backbones (MIRAGE, Brain Diffuser, iCNN) all produce much more consistent scaling improvements on mental images than the models that utilize deep neural network backbones (MindEye1, MindEye2).

A.15 Impact of diffusion strength on performance

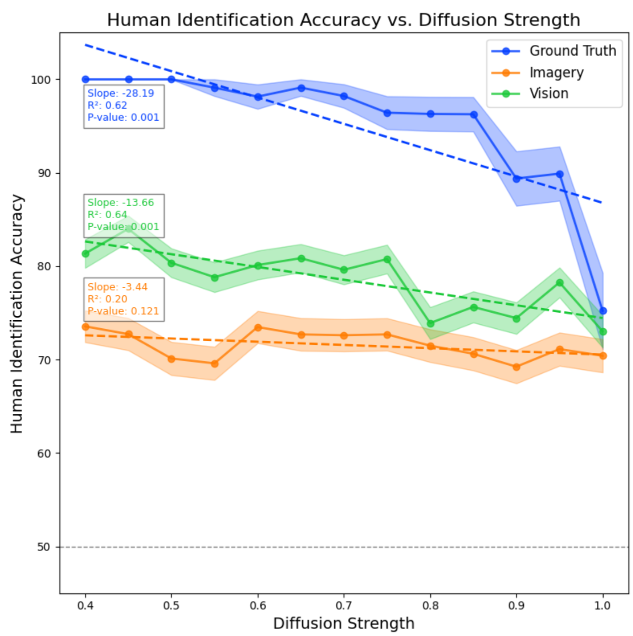

Figure 22: Human identification accuracy of MIRAGE (with no CLIP-Image guidance) as a function of diffusion model strength for imagery trials (orange line), vision trials (green line), and a control experiment that used the features directly from the ground truth image and caption (blue line). A dashed line is placed at the 50% chance threshold. Results are from a behavioral experiment that is identical to Experiment 1 (Table 5A), but varied across strength parameters.

Recent work [69, 70] has raised the question of how much of the detail in fMRI-to-image reconstructions originates in the brain and how much is simply hallucinated by the strong natural priors enforced by a diffusion model. This critical perspective makes a clear prediction: As the strength of the natural prior increases, the results should improve. One way of examining this relationship is by modulating the strength parameter in the img2img mode of the diffusion model, by which an initial image (in our case the low-level reconstruction provided by the VDVAE model discussed in Section 3.2) is partially noised and then denoised with CLIP guidance. This denoising process is where the natural priors are enforced, and the amount of denoising (and thus the amount of the final image that is guided by the natural prior) is modulated by the strength parameter. We repeated Experiment 1 presented in Table 5A of the paper across a wide range of strength parameters to investigate the potential influence of diffusion strength on the results. We test strength parameters between 0.4—the lowest strength value that yields meaningful variation from the input image—and 1.0, which destroys the entire input image before denoising. In this experiment, we use only CLIP-text semantic guidance for the diffusion process to increase the contrast between the original input image and the purely semantic guidance during the denoising process, although we acknowledge that this slightly reduces the performance of **MIRAGE** during the experiment relative to the results in Table 5A. The results demonstrate that increasing the strength parameter (and therefore increasing the influence of the prior in determining the ultimate reconstruction) induces no significant change in the rate at which humans can correctly identify mental image reconstructions as corresponding to the stimulus image, and seen image reconstructions experience a significant *decrease* in identifiability as diffusion strength increases. Although the diffusion model prior clearly plays a role in improving the quality and aesthetics of the reconstructions, our results show that there is still plenty of decoded signal present in the reconstructions to facilitate identification even with minimal guidance from the diffusion model.

To further emphasize our improved signal recovery and decreased reliance on the priors of the image generator, MIRAGE employs the lowest diffusion strength parameter (0.7) among comparable "dual stream" (high-level/low-level) reconstruction methods, including as Brain Diffuser (0.75) and MindEye1 (0.85), meaning MIRAGE explicitly relies less on the diffusion model's prior than other approaches. Additionally, MIRAGE achieves state-of-the-art (SOTA) performance on both simple and conceptual stimuli types, which are outside the natural prior of the diffusion model. These findings underscore MIRAGE's ability to extract and utilize more imagery signal from the brain independent of the natural prior of the diffusion model, setting it apart from alternative methods.

A.16 Head-to-head evaluation on the conceptual stimuli

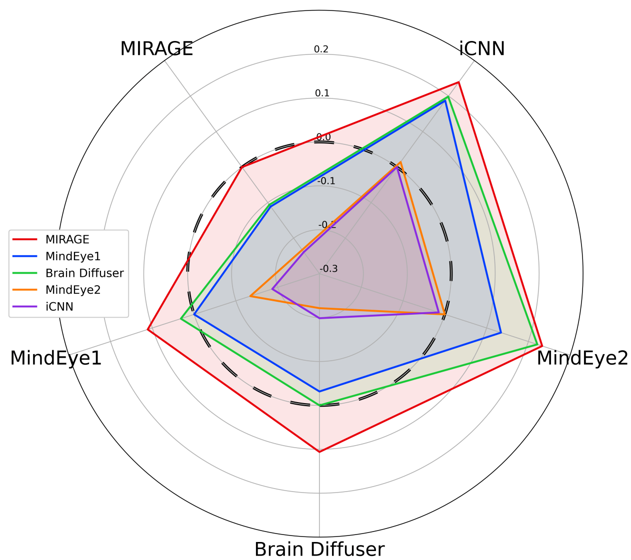

Figure 23: Head-to-head human similarity score results for the conceptual stimuli. The Y-axis represents the similarity score advantage (difference between target method's score and the alternative, on the radial X-axis); a larger colored polygon area indicates a stronger advantage, and the dashed circle at unity denotes equal performance.

Mental image reconstructions of the conceptual stimuli are particularly difficult to evaluate, as they do not have associated ground truth images or a meaningful match to vision trials. We pitted reconstruction methods head-to-head by presenting human raters with a ground truth conceptual stimulus and two reconstructions of that stimulus sampled at random from the collection of methods we evaluate against in this work. Figure 23 plots the average differences between similarity scores in these head-to-head comparisons, or "similarity score advantage" for all combinations of methods. Reconstructions from our method are the most strongly preferred in all head-to-head comparisons.

A.17 Behavioral experiment

A.17.1 Experiment protocols

We conducted a set of behavioral experiments on 500 human raters online. For our experiment, we identified no risks to the human participants, and our institution's IRB approved our experiment. We probed 3 experiments intermixed into two discrete sections within the same behavioral tasks, with each experiment consisting of trials sampled evenly from the different stimulus types and the 4 NSD subjects who completed all 40 scanning sessions (subjects 1, 2, 5, 7). The experimental trials within each task were shuffled and 36 trials were presented to each subject. Our subjects were recruited through the Prolific platform, with our experimental tasks hosted on Meadows. Each human rater was paid $\$1.50$ for the completion of the experiment, and the median completion time was 6 minutes and 17 seconds, resulting in an average payment rate of $\$14.32$/hour. Each human rater was presented with 6 attention check trials during the experiment. An attention check is a trial in which the ground truth image is presented as a candidate image during the trial. Because the ground truth image will always be the image that is most similar to itself, these trials were used to identify whether subjects were paying attention to the task and the instructions. We identified 5 human raters who failed at least 2 attention checks and removed those raters from our data before conducting our analysis. Code to reproduce our experiment can be found in our anonymized GitHub repository. All human subjects provided informed consent. All procedures were approved the Institutional Review Board at the University of Minnesota.

A.17.2 2AFC identification task

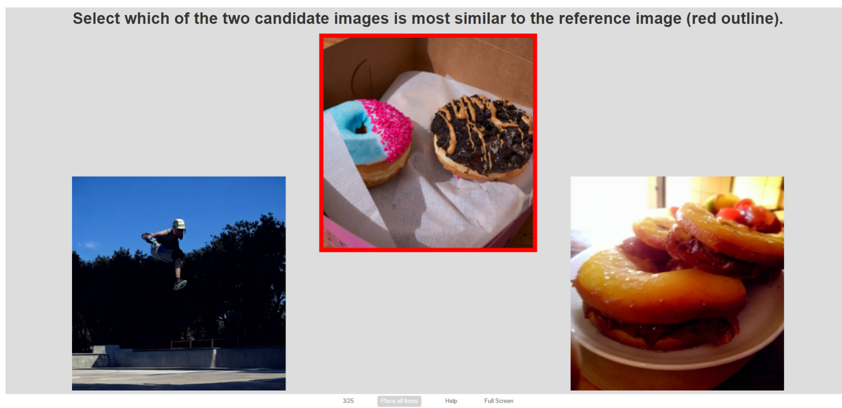

Figure 24: An example of the 2 alternative forced choice task used in the first behavioral experiment performed by human raters.

Our first experiment, which made up the entirety of the first task, was a 2 alternative forced choice task (2AFC) facilitated by the "Match-To-Sample" task on the Meadows platform. An example of the first experiment can be seen in Figure 24. In this experiment, human raters were asked to select which of two candidate images was more similar to a reference image. The reference image provided is the ground truth image the NSD-Imagery subject either saw or imagined, and the 2 candidate images were the target reconstruction of the reference image, or a randomly selected reconstruction from an fMRI scan corresponding to a different stimulus of the same stimulus type. The two candidate images were always sampled from the same reconstruction method and NSD-Imagery subject. This experiment was repeated for all reconstruction methods, visual modalities, NSD subjects, and across 10 reconstructions sampled from the output distribution of each reconstruction method. With the results presented in Section 4.2, we establish a baseline for human-rated image identification accuracy of mental image reconstructions, as no other paper has conducted behavioral evaluations of mental image reconstructions.

A.17.3 Continuous similarity rating task

The second and third experiments we conducted were shuffled together inside the second task of the experiment, which was facilitated by the "Drag-Rate" task on the Meadows platform. An example of the task used in experiments 2 and 3 can be seen in Figure 25. In this task, human raters were presented with a reference image, two candidate images, and a continuous two-dimensional plot that they could drag the candidate images onto, where the Y-axis represented "similarity to the reference image" and the X-axis represented the rater's confidence. The reference image provided was always the ground truth image the NSD-Imagery subject either saw or imagined. For experiment 2, the 2 candidate images were reconstructions of the reference image from the imagery and vision trials of the NSD-Imagery trials. Experiment 2 was repeated for the simple and complex stimuli (as conceptual stimuli do not have meaningful vision reconstructions), all reconstruction methods, NSD

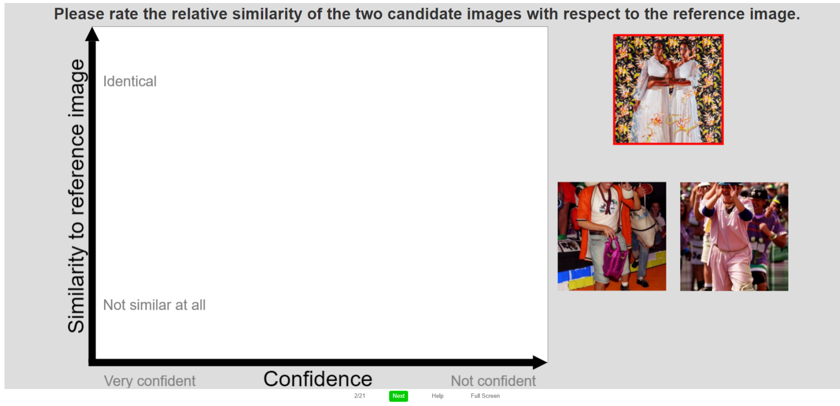

Figure 25: An example of similarity score task used in experiments 2 and 3 of the behavioral experiment performed by human raters.

subjects, and across 10 reconstructions sampled from the output distribution of each reconstruction method. For experiment 3, we designed a head-to-head experiment for conceptual reconstructions, where the candidate images were reconstructions of the imagery trials for the conceptual stimuli produced by different reconstruction methods. Experiment 3 contained trials for all NSD subjects, 10 reconstructions sampled from the output distribution of each reconstruction method, and 3 unique combinations of reconstruction methods for each sample. One-dimensional similarity ratings—like the ones used in this section of the experiment—can be extremely sensitive to the context of the alternative samples being compared against, and so are primarily useful for comparing the relative similarity of the candidate stimuli presented during each individual trial. The two comparison tasks evaluated within this task of the experiment were designed with this in mind, each configured to more directly compare the difference in quality between reconstructions of vision and imagery, as well as to compare the differences in quality between reconstruction methods on the conceptual stimuli. Our analysis of these results in Section 4.2 provides a detailed analysis of how reconstruction performance scales across vision and imagery, and of how each method performs on the conceptual stimuli.

A.18 iCCN implementation

Originally introduced in Shen et al. [20], and first trained on NSD in Shirakawa et al. [69], we adapt the author's open source implementation to try and faithfully replicate their results, making the following changes to the implementation:

1. **Normalization of images:** We disabled normalization of images when computing VGG19 features. During our initial trials, normalization led to unexpected color distortions in the reconstructed images. Removing normalization allowed the reconstructions to maintain their original color integrity, which is particularly crucial for visual comparisons in tasks requiring precise color representation.
2. **Feature decoding with Ridge Regression:** Instead of the `fastl2lir` library, we employed the Ridge Regression implementation from the `sklearn` library. This change enhanced compatibility with the rest of our workflow and provided better support for managing memory-intensive computations. For VGG19 layers with a large feature space, feature decoding was performed in chunks. This approach enabled the simultaneous calculation of features and fitting of the Ridge Regression model without requiring intermediate results to be saved to disk, thereby optimizing both time and memory usage.

A.19 Anonymized code release

We provide an anonymized GitHub repository with the code to reproduce our method.