# A. Appendix

## A.1. Additional dataset information

Stimuli within both NSD [1] and NSD-Imagery were displayed at $8.4 \times 8.4$ degrees. All fMRI data in the NSD were collected at ultra-high field (7T) using a whole-brain, 1.8-mm, 1.6-s, gradient-echo, echo-planar imaging (EPI) pulse sequence. The fMRI responses are expressed in terms of "betas" ($\beta$; each $\beta$ is a measure of the amplitude of BOLD signal evoked by a single image in a single voxel) obtained from a general linear model (GLM) analysis. Betas indicate BOLD response amplitudes evoked by each stimulus trial relative to the baseline signal level present during the absence of a stimulus (gray screen). All reconstruction methods evaluated in this work were trained using GLMsingle results provided with the NSD data release, and evaluated on the GLMsingle preparations of the NSD-Imagery data, specifically, the 1.8-mm volume preparation of the data and version 3 of the GLM betas (betas_fithrf_GLMdenoise_RR).

The shared1000 test set—which reconstruction methods are typically evaluated on—was sampled from scanning sessions of NSD where training data was also collected. Thus, cross-session non-stationarities are likely to have had similar impacts on the training and evaluation data. The vision trials in NSD-Imagery, however, were collected during a separate scanning session. Thus, it is likely that cross-session non-stationarities had a more detrimental impact on decoder performance when generalizing to NSD-imagery than to the shared1000 data. This could explain why the performance of all decoders on the NSD-Imagery vision trials were generally lower than previously published results for the shared1000 trials. The drop in performance relative to the shared1000 could also have a number of other causes, including a different task being used in the NSD-Imagery vision trials (cue-matching vs continuous recognition), a different number of stimulus repetitions in NSD-Imagery (8 repetitions vs. 3 repetitions), or the possibility that the GLMsingle algorithm used to preprocess the fMRI responses [4] is less effective on small datasets.

Across all methods evaluated in this paper [2, 3, 5, 6, 10, 11], we trained the models using the full 40 sessions of the NSD training data (not including the shared1000), and we perform inference using the same brain region as the original paper authors. For the majority of methods evaluated in the paper, this means we utilize only voxels from the "nsdgeneral" brain region, defined by the NSD authors as the subset of voxels in posterior cortex most responsive to the visual stimuli presented (between 13,000 to 16,000 voxels per participant). The method from Takagi et al. is the only method to deviate from this, instead using the respective ROIs for early and higher (ventral) visual regions included in the streams atlas of the NSD.

Because the NSD-Imagery dataset comprises both vision and imagery trials within a single scanning session, as well as multiple types of discrete stimulus types, we Z-scored the fMRI data within each experimental run separately. A run is defined by a series of consecutive trials comprising the same visual modality and stimulus type (e.g., vision trials comprising simple stimuli), typically lasting 4 minutes. Normalizing trials for vision and imagery runs separately provides some control against non-stationary brain activity (e.g., changes in SNR) across imagery and vision.

## A.2. Additional evaluation metric details

All metrics calculated in Table 1 of the manuscript were calculated across 10 reconstructions sampled from the posterior distribution of each decoding method. 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 other papers, and we do not report the shared1000 2WC metrics in Table 1 of the manuscript. Brain correlation scores are the Pearson correlation between the averaged measured brain response $\beta$ and the predicted brain response $\beta'$ produced by a brain encoding model (GNet [9]) 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, 2, and 3 were calculated and averaged across 10 images sampled from the output distribution of each method using a random seed.

## A.3. Median and worst case reconstructions


Figure 1. Qualitative comparison of the median-case reconstructions on stimuli seen during the vision trials of NSD-Imagery. Samples selected are the median scoring according to the reconstruction metrics in Table 1 of the manuscript.

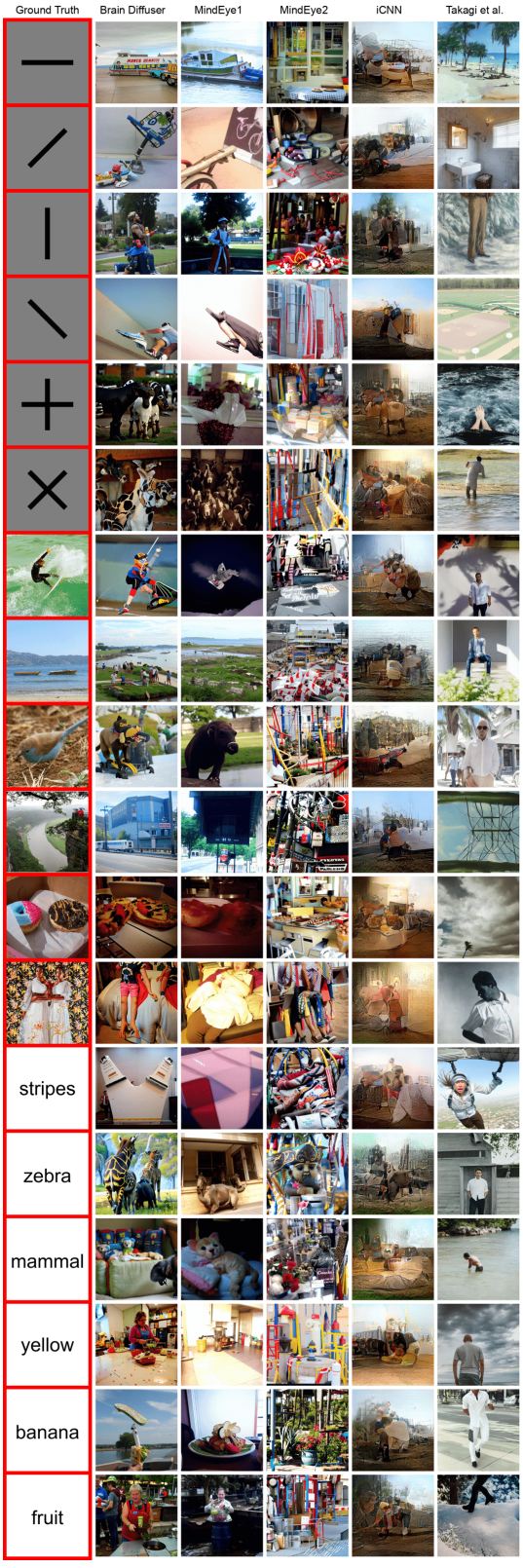


Figure 2. Qualitative comparison of the median-case reconstructions on stimuli imagined during the imagery trials of NSD-Imagery. Samples are selected the same way as Figure 1.


Figure 3. Qualitative comparison of the worst-case reconstructions on stimuli seen during the vision trials of NSD-Imagery. Samples selected are the worst scoring according to the reconstruction metrics in Table 1 of the manuscript.

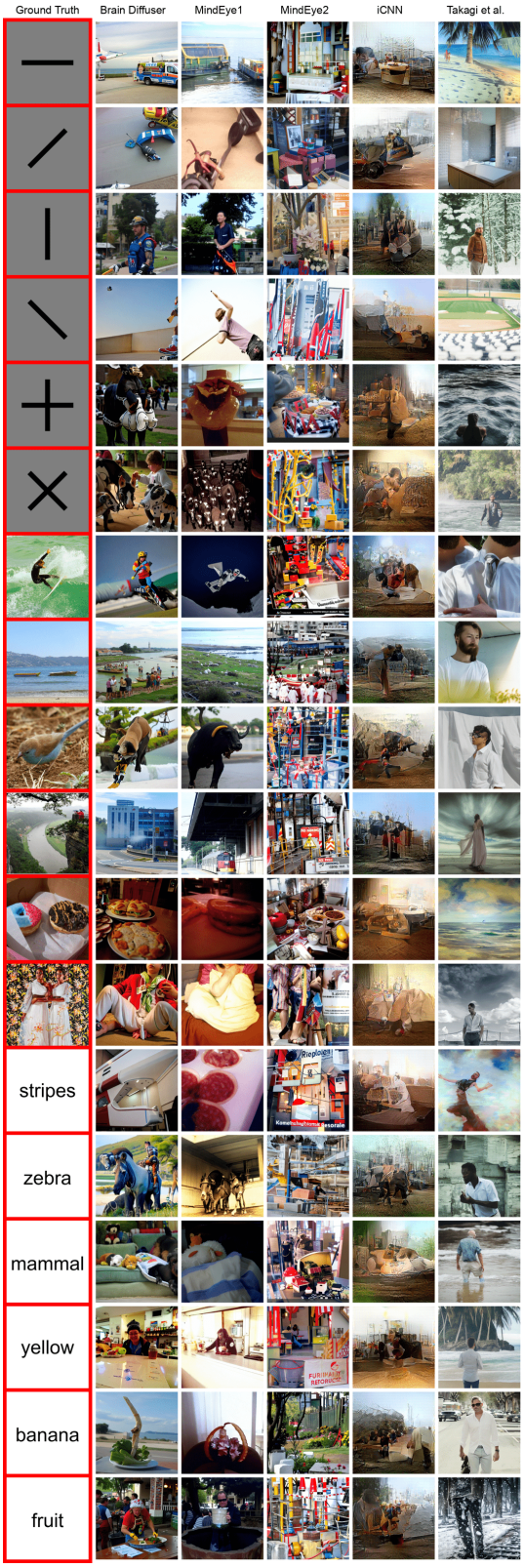


Figure 4. Qualitative comparison of the worst-case reconstructions on stimuli imagined during the imagery trials of NSD-Imagery. Samples are selected the same way as Figure 3.

## A.4. 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)** | | | | | | | | | | | |
| **MindEye1 [5]** | 0.033 | 0.456 | **43.71%** | **61.67%** | 37.46% | 58.37% | **0.974** | 0.563 | **0.200** | **0.107** | **0.148** |
| **Brain Diffuser [3]** | 0.013 | 0.524 | 30.68% | 50.68% | 34.43% | 44.51% | 0.983 | 0.603 | 0.152 | 0.091 | 0.128 |
| **iCNN [7]** | **0.063** | 0.427 | 27.42% | 47.65% | **45.11%** | **67.99%** | 1.006 | **0.546** | 0.138 | 0.045 | 0.081 |
| **MindEye2 [6]** | 0.011 | 0.448 | 23.37% | 45.34% | 31.14% | 49.02% | 0.987 | 0.590 | 0.074 | 0.035 | 0.051 |
| **Takagi et al. [10, 11]** | 0.027 | **0.595** | 29.70% | 50.30% | 37.12% | 54.70% | 0.980 | 0.591 | -0.002 | -0.001 | 0.002 |
| **Vision Reconstructions (Simple Stimuli)** | | | | | | | | | | | |
| **MindEye1 [5]** | 0.129 | 0.506 | **62.01%** | **76.36%** | 43.33% | 60.64% | **0.961** | 0.549 | 0.370 | **0.140** | 0.243 |
| **Brain Diffuser [3]** | 0.075 | **0.586** | 40.19% | 66.67% | 38.30% | 42.20% | 0.988 | 0.601 | 0.209 | 0.106 | 0.169 |
| **iCNN [7]** | **0.132** | 0.454 | 57.01% | 74.89% | 37.69% | **69.02%** | 0.992 | **0.534** | **0.447** | 0.133 | **0.278** |
| **MindEye2 [6]** | 0.040 | 0.487 | 50.87% | 68.98% | **43.52%** | 52.46% | 0.980 | 0.577 | 0.334 | 0.108 | 0.204 |
| **Takagi et al. [10, 11]** | 0.015 | 0.542 | 22.16% | 50.68% | 32.73% | 55.19% | 0.968 | 0.588 | 0.012 | -0.007 | 0.001 |

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

| 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)** | | | | | | | | | | | |
| **MindEye1 [5]** | 0.138 | 0.243 | **75.42%** | 60.34% | 66.51% | 51.06% | 0.921 | **0.566** | **0.159** | **0.164** | **0.161** |
| **Brain Diffuser [3]** | 0.114 | 0.278 | 73.60% | **66.02%** | **71.02%** | **63.64%** | **0.888** | 0.567 | 0.114 | 0.163 | 0.154 |
| **iCNN [7]** | **0.153** | 0.253 | 73.71% | 62.84% | 53.67% | 15.46% | 0.982 | 0.575 | 0.089 | 0.079 | 0.081 |
| **MindEye2 [6]** | 0.032 | 0.231 | 70.42% | 65.11% | 61.97% | 51.93% | 0.943 | 0.601 | 0.062 | 0.074 | 0.068 |
| **Takagi et al. [10, 11]** | -0.039 | **0.315** | 54.05% | 30.08% | 49.39% | 25.46% | 0.972 | 0.622 | -0.001 | 0.009 | 0.003 |
| **Vision Reconstructions (Complex Stimuli)** | | | | | | | | | | | |
| **MindEye1 [5]** | 0.308 | 0.318 | 85.11% | 85.27% | 81.55% | 70.04% | 0.800 | 0.471 | 0.378 | **0.365** | 0.379 |
| **Brain Diffuser [3]** | 0.139 | **0.323** | 80.49% | 79.02% | 83.60% | 74.43% | 0.829 | 0.509 | 0.284 | 0.353 | 0.341 |
| **iCNN [7]** | **0.316** | 0.316 | **86.33%** | **87.80%** | **84.62%** | 29.05% | 0.860 | 0.514 | **0.437** | 0.358 | **0.397** |
| **MindEye2 [6]** | 0.223 | 0.333 | 84.28% | 85.83% | 80.08% | **77.46%** | **0.794** | **0.454** | 0.378 | 0.360 | 0.376 |
| **Takagi et al. [10, 11]** | -0.041 | 0.281 | 60.95% | 27.84% | 45.80% | 30.83% | 0.971 | 0.632 | -0.024 | 0.039 | 0.016 |

Table 2. Quantitative comparison between reconstruction methods for both imagery and vision trials on complex stimuli. Metrics are the same as Table 1 of the manuscript.

## A.5. Behavioral experiment

### A.5.1. 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.

### A.5.2. 2AFC identification task

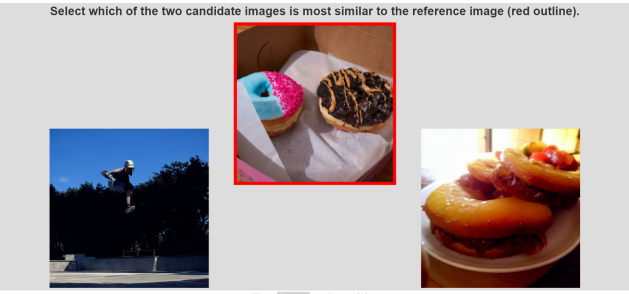


Figure 5. An example of the 2 alternative forced choice task used in the first behavioral experiment performed by human raters.

Our first experiment 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 5. 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 4.5, 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.5.3. Continuous similarity rating task


Figure 6. An example of similarity score task used in experiment 2 of the behavioral experiment performed by human raters.

The second experiment we conducted was facilitated by the "Drag-Rate" task on the Meadows platform. An example of the task can be seen in Figure 6. 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 subjects, and across 10 reconstructions sampled from the output distribution of each reconstruction method. 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 task was designed with this in mind, configured to directly compare the difference in quality between reconstructions of vision and imagery for each method. Our analysis of these results in Section 4.5 provides a detailed analysis of how reconstruction performance scales across vision and imagery.

## A.6. iCCN implementation

Originally introduced in Shen et al. [7], and first trained on NSD in Shirakawa et al. [8], 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.7. Impact of trial repetition averaging on performance

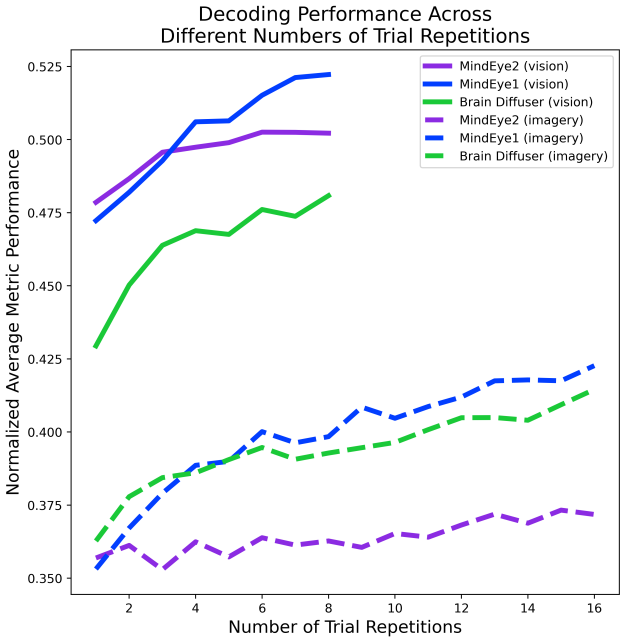


Figure 7. Performance of various methods when averaging across brain activity responses to multiple trial repetitions of the same stimulus. Y axis is the normalized average of all metrics in Table 1 of the manuscript, X axis is the number of averaged trial repetitions.

One of the experimental details that varies between NSD [1] and NSD-Imagery 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 7, we plot the effect of these additional trial repetitions on the performance of a subset of the reconstruction methods evaluated in this work.

## A.8. Impact of training data scale on performance

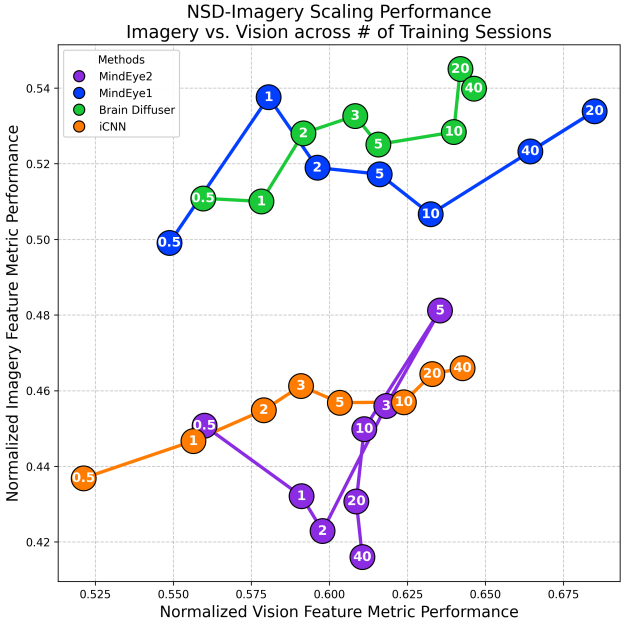


Figure 8. Performance of various methods on NSD-Imagery for Subject 1 when trained on different numbers of fMRI sessions present in NSD. Each session includes approximately one hour of fMRI data. Metrics are the normalized average of all metrics in Table 1 of the manuscript, with imagery performance 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. All of the 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 clinical settings is currently impractical or impossible for certain patients. Recent work in MindEye2 [6] has made strides in scaling decoding procedures using a multi-subject pretraining step, however as demonstrated in Figure 8, this approach generalizes poorly to mental imagery data. We additionally note that the methods that used ridge regression decoding backbones (Brain Diffuser, iCNN) produce much more consistent scaling improvements on mental images than the models that utilize deep neural network backbones (MindEye1, MindEye2).

## References


[1] 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, 2022. 1, 6

[2] Reese Kneeland, Jordyn Ojeda, Ghislain St-Yves, and Thomas Naselaris. Brain-optimized inference improves reconstructions of fMRI brain activity, 2023. arXiv:2312.07705 [cs, q-bio]. 1

[3] Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, 13, 2023. 1, 4

[4] Jacob S Prince, Ian Charest, Jan W Kurzawski, John A Pyles, Michael J Tarr, and Kendrick N Kay. Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *eLife*, 11:e77599, 2022. Publisher: eLife Sciences Publications, Ltd. 1

[5] 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. 1, 4

[6] Paul Steven Scotti, Mihir Tripathy, Cesar Torrico, 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 *ICLR 2024 Workshop on Representational Alignment*, 2024. 1, 4, 6

[7] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLOS Computational Biology*, 15(1):e1006633, 2019. 4, 6

[8] 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. 6

[9] 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. 1

[10] 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. 1, 4

[11] Yu Takagi and Shinji Nishimoto. Improving visual image reconstruction from human brain activity using latent diffusion models via multiple decoded inputs, 2023. 1, 4