MindEye2: Shared-Subject Models Enable fMRI-To-Image With 1 Hour of Data

Paul S. Scotti [1 2 3] **Mihir Tripathy** [† 2] **Cesar Kadir Torrico Villanueva** [† 2] **Reese Kneeland** [† 4] **Tong Chen** [5 2]
Ashutosh Narang [2] **Charan Santhirasegaran** [2] **Jonathan Xu** [6 2] **Thomas Naselaris** [4] **Kenneth A. Norman** [3]
Tanishq Mathew Abraham [1 2]

Figure 1: MindEye2 vs. MindEye1 reconstructions from fMRI brain activity using varying amounts of training data.

Abstract

Reconstructions of visual perception from brain activity have improved tremendously, but the practical utility of such methods has been limited. This is because such models are trained independently per subject where each subject requires dozens of hours of expensive fMRI training data to attain high-quality results. The present work showcases high-quality reconstructions using only 1 hour of fMRI training data. We pretrain our model across 7 subjects and then fine-tune on minimal data from a new subject. Our novel functional alignment procedure linearly maps all brain data to a shared-subject latent space, followed by a shared non-linear mapping to CLIP image space. We then map from CLIP space to pixel space by fine-tuning Stable Diffusion XL to accept CLIP latents as inputs instead of text. This approach improves out-of-subject generalization with limited training data and also attains state-of-the-art image retrieval and reconstruction metrics compared to single-subject approaches. MindEye2 demonstrates how accurate reconstructions of perception are possible from a single visit to the MRI facility. All code is available on GitHub.

1 Introduction

Spurred by the open releases of deep learning models such as CLIP (Radford et al., 2021) and Stable Diffusion (Rombach et al., 2022), along with large-scale functional magnetic resonance imaging (fMRI) datasets such as the Natural Scenes Dataset (Allen et al., 2022) where human participants were scanned viewing tens of thousands of images, there has been an influx of research papers demonstrating the ability to reconstruct visual perception from brain activity with high fidelity (Takagi and Nishimoto, 2022; 2023; Ozcelik et al., 2022; Ozcelik and VanRullen, 2023; Gaziv et al., 2022; Gu et al., 2023; Scotti et al., 2023; Kneeland et al., 2023a;b;c; Ferrante et al., 2023a; Thual et al., 2023; Chen et al., 2023a;b; Sun et al., 2023; Mai and Zhang, 2023; Xia et al., 2023). FMRI indirectly measures neural activity by detecting changes in blood oxygenation. These patterns of fMRI brain activity are translated into embeddings of pretrained deep learning models and used to visualize internal mental representations (Beliy et al., 2019; Shen et al., 2019a;b; Seeliger et al., 2018; Lin et al., 2019).

[†]Core contribution. [1]Stability AI [2]Medical AI Research Center (MedARC) [3]Princeton Neuroscience Institute [4]University of Minnesota [5]The University of Sydney [6]University of Waterloo. Correspondence to: Paul Scotti <scottibrain@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Visualization of internal mental representations, and more generally the ability to map patterns of brain activity to the latent space of rich pretrained deep learning models, has potential to enable novel clinical assessment approaches and brain-computer interface applications. However, despite all the recent research demonstrating high-fidelity reconstructions of perception, the practical adoption of such approaches to these settings has been limited if not entirely absent. A major reason for this is that the high-quality results shown in these papers use single-subject models that are not generalizable across people, and which have only been shown to work well if each subject contributes dozens of hours of expensive fMRI training data. MindEye2 introduces a novel functional alignment procedure that addresses these barriers by pretraining a shared-subject model that can be fine-tuned using limited data from a held-out subject and generalized to held-out data from that subject. This approach yields similar reconstruction quality to a single-subject model trained using $40\times$ the training data. See Figure 1 for selected samples of reconstructions obtained from just 1 hour of data from subject 1 compared to their full 40 hours of training data in the Natural Scenes Dataset.

In addition to a novel approach to shared-subject alignment, MindEye2 builds upon the previous SOTA approach introduced by MindEye1 (Scotti et al., 2023). In terms of similarities, both approaches map flattened spatial patterns of fMRI activity across voxels (3-dimensional cubes of cortical tissue) to the image embedding latent space of a pretrained CLIP (Radford et al., 2021) model with the help of a residual MLP backbone, diffusion prior, and retrieval submodule. The diffusion prior (Ramesh et al., 2022) is used for reconstruction and is trained from scratch to take in the outputs from the MLP backbone and produce aligned embeddings suitable as inputs to any pretrained image generation model that accepts CLIP image embeddings (hereafter referred to as unCLIP models). The retrieval submodule is contrastively trained and produces CLIP-fMRI embeddings that can be used to find the original (or nearest neighbor) image in a pool of images, but is not used to reconstruct a novel image. Both MindEye2 and MindEye1 also map brain activity to the latent space of Stable Diffusion's (Rombach et al., 2022) variational autoencoder (VAE) to obtain blurry reconstructions that lack high-level semantic content but perform well on low-level image metrics (e.g., color, texture, spatial position), which get combined with the semantically rich outputs from the diffusion prior to return reconstructions that perform well across perceptual and semantic features.

MindEye2 innovates upon MindEye1 in the following ways: (1) Rather than the whole pipeline being independently trained per subject, MindEye2 is pretrained on data from other subjects and then fine-tuned on the held-out target subject. (2) We map from fMRI activity to a richer CLIP space provided by OpenCLIP ViT-bigG/14 (Schuhmann et al., 2022; Ilharco et al., 2021), and reconstruct images via a fine-tuned Stable Diffusion XL unCLIP model that supports inputs from this latent space. (3) We merge the previously independent high- and low-level pipelines into a single pipeline through the use of submodules. (4) We additionally predict the text captions of images to be used as conditional guidance during a final image reconstruction refinement step.

The above changes support the following main contributions of this work: (1) Using the full fMRI training data from Natural Scenes Dataset we achieve state-of-the-art performance across image retrieval and reconstruction metrics. (2) Our novel multi-subject alignment procedure enables competitive decoding performance even with only 2.5% of a subject's full dataset (i.e., 1 hour of scanning).

2 MindEye2

MindEye2 involves pretraining and then fine-tuning a single model where brain activity is mapped to the embedding space of pretrained deep learning models. During inference, these embeddings predicted from the brain are fed into frozen image generative models that translate from model space to pixel space. Our strategy to reconstruct seen images from brain activity using minimal training data is to first pretrain the model using data from 7 subjects (30-40 hours of scanning data each) and then to fine-tune the model using data from a held-out 8th subject. The full MindEye2 pipeline is depicted in Figure 2.

Single-subject models were trained/fine-tuned on a single 8xA100 80Gb GPU node for 150 epochs with a batch size of 24. Multi-subject pretraining was done with a batch size of 63 (9 samples per each of 7 subjects). Models were trained with Huggingface Accelerate (Gugger et al., 2022) and DeepSpeed (Rajbhandari et al., 2020) Stage 2 with CPU offloading.

2.1 Shared-Subject Functional Alignment

Every subject has a uniquely shaped brain with different functional organization, meaning that there needs to be an initial alignment step to ensure the model can handle inputs from different brains. Unlike anatomical alignment where every subject's brain is mapped to the same brain template (Talairach and Tournoux, 1990; Mazziotta et al., 2001), we remain in subjects' native brain space and functionally align flattened spatial patterns of fMRI activity to a shared-subject latent space using subject-specific ridge regression. That is, each subject has a separate linear layer with weight decay to map the input fMRI voxels (13,000 to 18,000 voxels depending on the subject) to a 4096-dim latent.

Following this initial linear layer, the rest of the model pipeline is shared across subjects without any subject-

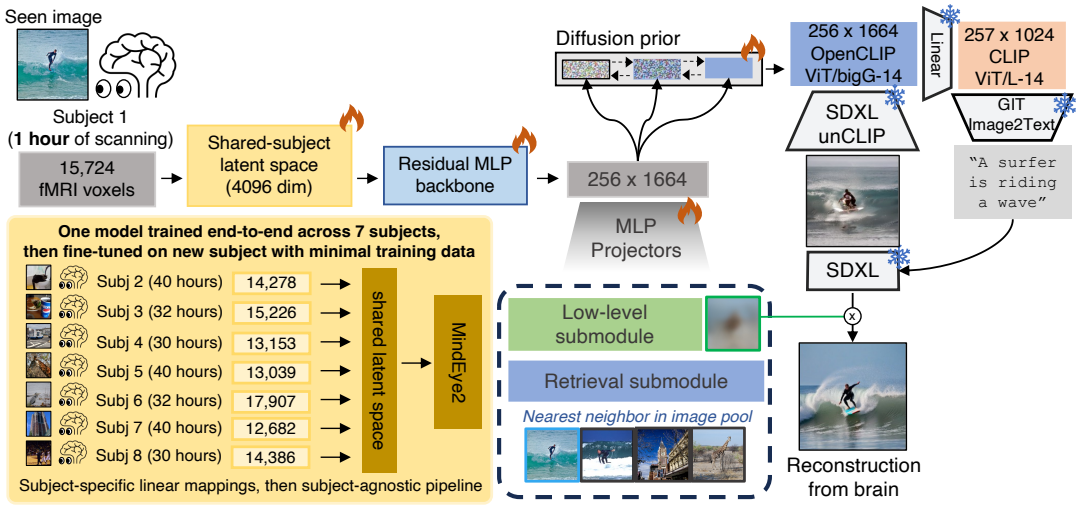

Figure 2: MindEye2 overall schematic. MindEye2 is trained using samples from 7 subjects in the Natural Scenes Dataset and then fine-tuned using a target held-out subject who may have scarce training data. Ridge regression maps fMRI activity to an initial shared-subject latent space. An MLP backbone and diffusion prior output OpenCLIP ViT-bigG/14 embeddings which SDXL unCLIP uses to reconstruct the seen image, which are then refined with base SDXL. The submodules help retain low-level information and support retrieval tasks. Snowflakes=frozen models used during inference, flames=actively trained.

specific mappings. The whole pipeline is trained end-to-end where pretraining involves each batch containing brain inputs from all subjects. That is, alignment to shared-subject space is not trained independently and we do not pretrain models separately for each subject; rather, we pretrain a single model equally sampling across all the subjects except the held-out subject used for fine-tuning.

Two strengths of this novel functional alignment procedure are in its simplicity and flexibility. Using a simple linear mapping for alignment can provide robust, generalizeable performance in low-sample, high-noise settings because simple mappings are less likely to overfit to noise. Also, unlike typical functional alignment approaches that require subjects to experience the same shared set of images (Haxby et al., 2011), our approach has the flexibility to work even when subjects are viewing entirely unique images in the training data. This is critical for the Natural Scenes Dataset, where 90% of the seen images are unique to the subject and the 10% that were seen across subjects are relegated to the test set. Further, this approach holds advantages for data collection of a new subject, where such data collection does not need to be restricted to showing a predefined set of images. This approach is also relevant for application on small target datasets where other pre-training data is available.

2.2 Backbone, Diffusion Prior, & Submodules

Flattened spatial patterns of brain activity are first linearly mapped to the shared-subject space using an output dimensionality of 4096. Then, these latents are fed through an MLP backbone with 4 residual blocks, followed by a linear mapping that goes from 4096-dim to 256×1664 dimensionality of OpenCLIP ViT-bigG/14 image token embeddings. These backbone embeddings are then simultaneously fed through a diffusion prior (Ramesh et al., 2022) and two MLP projectors (retrieval and low-level submodules). Differences from MindEye1 include linear mapping to a shared-subject space, mapping to OpenCLIP ViT-bigG/14 rather than CLIP ViT-L/14, and adding a low-level MLP submodule.

MindEye2 has three losses that are summed, stemming from the diffusion prior, retrieval submodule, and low-level submodule. The end-to-end loss, with $\alpha_1 = .033$ and $\alpha_2 = .016$, is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{prior}} + \alpha_1 \cdot \mathcal{L}_{\text{BiMixCo|SoftCLIP}} + \alpha_2 \cdot \mathcal{L}_{\text{lowlevel}} \quad (1)$$

2.2.1 Diffusion Prior

Using a diffusion prior to align outputs from a contrastive learning model was inspired by DALL-E 2 (Ramesh et al., 2022), where a "diffusion prior" maps CLIP text embeddings to CLIP image space before using an unCLIP decoder to reconstruct images. Here we trained our own diffusion prior from scratch to map fMRI latents to the OpenCLIP ViT-bigG/14 image space, which was kept frozen as done with locked-image text tuning (LiT) (Zhai et al., 2022). We used the same prior loss as Ramesh et al. (2022), implemented with the same code as MindEye1 which used modified code from the DALLE2-pytorch repository.

2.2.2 RETRIEVAL SUBMODULE

MindEye1 observed a tradeoff if using contrastive loss and MSE loss on the outputs of the diffusion prior directly, such that the model could not effectively learn a single embedding to satisfy both objectives. Instead, applying MSE loss on the diffusion prior and applying contrastive loss on the outputs from an MLP projector attached to the MLP backbone effectively mitigated this tradeoff because the objectives no longer shared identical embeddings. We adopted the same approach here, with the retrieval submodule contrastively trained to maximize cosine similarity for positive pairs while minimizing similarity for negative pairs. We used the same BiMixCo and SoftCLIP losses used in MindEye1 (Scotti et al., 2023), which involved the first third of training iterations using bidirectional MixCo data augmentation (Kim et al., 2020) with hard labels and the last two-thirds of training iterations using soft labels (generated from the dot product of CLIP image embeddings in a batch with themselves) without data augmentation.

2.2.3 LOW-LEVEL SUBMODULE

MindEye1 used an independent low-level pipeline to map voxels to the latent space of Stable Diffusion's variational autoencoder (VAE) such that blurry reconstructions were returned that lacked semantic information but performed well on low-level metrics. Here, we reimplement this pipeline as a submodule, similar to the retrieval submodule, such that it need not be trained independently. The MLP projector feeds to a CNN upsampler that upsamples to the $(64, 64, 4)$ dimensionality of SD VAE latents with L1 loss as well as an additional MLP to the embeddings of a teacher linear segmentation model VICRegL (Bardes et al., 2022) ConvNext-XXL ($\alpha = 0.75$) for an auxilliary SoftCLIP loss (soft labels from VICRegL model).

$$\mathcal{L}_{\text{lowlevel}} = \frac{1}{N}\sum_{i=1}^{N} |\text{VAE}_i - \hat{\text{VAE}}_i| + L_{\text{SoftCLIP}}(\text{VIC}, \hat{\text{VIC}}) \tag{2}$$

2.3 Image Captioning

To predict image captions from brain activity we convert the diffusion prior's predicted ViT-bigG/14 embeddings to CLIP ViT/L-14 space and then feed through a frozen pretrained GenerativeImage2Text (GIT) model (Wang et al., 2022). The use of GIT to caption images from brain activity in the Natural Scenes Dataset was previously shown to be viable by Ferrante et al. (2023b). We independently trained a linear model to convert from OpenCLIP ViT-bigG/14 embeddings to CLIP ViT-L/14 embeddings (see Appendix A.7), which was necessary because there was no existing GIT model that accepted OpenCLIP ViT-bigG/14 embeddings as inputs. Image caption prediction from brain activity lends further flexibility to such decoding approaches and can help refine image reconstructions to match desired semantic content.

2.4 Fine-tuning Stable Diffusion XL for unCLIP

CLIP (Radford et al., 2021) is an example of a multimodal contrastive model that maps images and text captions to a shared embedding space. unCLIP (or image variations) models go from this shared embedding space back to pixel space, and have been used for the creative application of returning variations of a given reference image (Xu et al., 2023; Ye et al., 2023; Pinkney, 2022). As such, previous unCLIP models prioritized replication of high-level semantics over low-level structures. These models can be trained by fine-tuning a base image generation model to accept CLIP image embeddings instead of, or in addition to, text embeddings. Outputs are diffused from pure noise just like the base model, unlike image-to-image models (Meng et al., 2022) that start the diffusion process from a reference image mixed with noise.

Contrary to previous unCLIP models, our goal was to train a model that returns images as close as possible to the reference image across both low-level structure and high-level semantics. This is because our use-case was to exactly return the original image given its CLIP image embedding predicted from the brain.

The base Stable Diffusion XL (SDXL) (Podell et al., 2023) model uses text conditionings from both OpenCLIP ViT-bigG/14 and CLIP ViT-L/14. They condition cross-attention layers on the penultimate text encoder outputs and additionally condition on pooled text embeddings from OpenCLIP ViT-bigG/14 by adding it to the timestep embedding. Here, we fine-tuned the cross-attention layers using the OpenCLIP ViT-bigG/14 image embeddings corresponding to all 256 patch tokens and we dropped the additional conditioning on pooled text embeddings. We opted to only condition on image embeddings because we observed that incorporating any text conditioning worsened the fidelity of the unCLIP reconstructions.

We evaluate the fidelity of our SDXL unCLIP model to reconstruct images from ground truth OpenCLIP ViT-bigG/14 image embeddings in Appendix A.6, showing that reconstructions are nearly identical to the original images. We fine-tuned SDXL on one 8xA100 80GB GPU node using an internal dataset for $110,000$ optimization steps at a resolution of 256×256 pixels and a batch size of 8 with offset-noise (Lin et al., 2024; Guttenberg, 2023) set to 0.04. All other settings were identical to those used with base Stable Diffusion XL. Like Stable Diffusion XL, this unCLIP model can output different aspect ratios, however, we observed best results with 768×768 resolution.

2.5 Model Inference

The pipeline for reconstruction inference is depicted in Figure 2. First, the diffusion prior's predicted OpenCLIP ViT-bigG/14 image latents are fed through our SDXL unCLIP model to output a pixel image. We observed that these reconstructions were often distorted ("unrefined") due to an imperfect mapping to bigG space (see Figure 3). This may be explained by the increased versatility allowed from mapping to the larger dimensionality OpenCLIP bigG latent space. To increase image realism, we feed the unrefined reconstructions from SDXL unCLIP through base SDXL via image-to-image (Meng et al., 2022) with text conditioning guidance from MindEye2's predicted image captions (section 2.3). We skip the first 50% of denoising diffusion timesteps, starting the process from the noised image encoding of the unrefined reconstruction. We simply take the first samples output from these stochastic models without any special 2nd-order selection. Refinement using base SDXL subjectively improves the quality of image outputs without strongly affecting low or high-level image metrics.

Figure 3: SDXL unCLIP reconstructions + predicted image captions (left) are fed to base SDXL for refinement (right).

The final "refined" reconstructions come from combining the outputs from base SDXL with the pixel images output from the low-level submodule via simple weighted averaging (4:1 ratio). This weighted averaging step increases performance on low-level image metrics while minimally affecting reconstructions' subjective appearance, but is overall not a critical component to MindEye2 and can be entirely discarded without a noticeable drop in reconstruction quality. Replacing this weighted averaging approach with conditioning the diffusion model using VAE embeddings as done in MindEye1 resulted in worsened performance, likely due to the OpenCLIP embeddings already doing a good job at retaining low-level image information.

For retrieval inference, only the retrieval submodule's outputs are necessary. Nearest neighbor retrieval can be performed via cosine similarity between the submodule's OpenCLIP ViT-bigG/14 embeddings and all the ViT-bigG/14 embeddings corresponding to the images in the desired image pool.

3 Results

We used the Natural Scenes Dataset (NSD) (Allen et al., 2022), a public fMRI dataset containing the brain responses of human participants viewing rich naturalistic stimuli from COCO (Lin et al., 2014). The dataset spans 8 subjects who were each scanned for 30-40 hours (30-40 separate scanning sessions), where each sesssion consisted of viewing 750 images for 3 seconds each. Images were seen 3 times each across the sessions and were unique to each subject, except for a select 1,000 images which were seen by all the subjects. We follow the standardized approach to train/test splits used by other NSD reconstruction papers (Takagi and Nishimoto, 2022; Ozcelik and VanRullen, 2023; Gu et al., 2023) which is to use the shared images seen by all the subjects as the test set. We follow the standard of evaluating model performance across low- and high-level image metrics averaged across the 4 subjects who completed all 40 scanning sessions. We averaged across same-image repetitions for the test set (1,000 test samples) but not the training set (30,000 training samples). The inputs to the model during training and fine-tuning are always single-trial, individual (non-pooled) brain-image paired samples. For more information on NSD and data preprocessing see Appendix A.2.

Critically, models trained/fine-tuned on a subset of data were selected in chronological order. That is, models fine-tuned from only 1 hour's worth of data come from using the subject's first scanning session of 750 image presentations. These are individual samples not pooled across image repeats. This means our model must be able to generalize to test data collected from scanning sessions entirely held-out during training/fine-tuning and which involve reconstructing images never presented to the model during training or fine-tuning.

3.1 fMRI-to-Image Reconstruction

First, we report performance of MindEye2 when training on the full NSD dataset. We quantitatively compare reconstructions across fMRI-to-image models in Table 1, demonstrating state-of-the-art MindEye2 performance across nearly all metrics. We compare to both the previous MindEye1 results as well as other fMRI-to-image approaches that were open-sourced such that we could replicate their pipelines using the recently updated NSD (which includes an additional 3 scanning sessions for every subject). For exhaustive figures depicting all MindEye2 reconstructions in the test set, see the figs folder in our GitHub repo.

MindEye2 refined reconstructions using the full NSD dataset performed SOTA across nearly all metrics, confirming that our changes to shared-subject modeling, model architecture, and training procedure benefitted reconstruction

Method	Low-Level				High-Level				Retrieval	
	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓	Image ↑	Brain ↑
MindEye2	**0.322**	**0.431**	**96.1%**	98.6%	95.4%	93.0%	**0.619**	0.344	**98.8%**	**98.3%**
MindEye2 (unrefined)	0.278	0.328	95.2%	**99.0%**	**96.4%**	**94.5%**	0.622	**0.343**	–	–
MindEye1	0.319	0.360	92.8%	96.9%	94.6%	93.3%	0.648	0.377	90.0%	84.1%
Ozcelik and VanRullen (2023)	0.273	0.365	94.4%	96.6%	91.3%	90.9%	0.728	0.421	18.8%	26.3%
Takagi and Nishimoto (2023)	0.246	0.410	78.9%	85.6%	83.8%	82.1%	0.811	0.504	–	–
MindEye2 (low-level)	0.399	0.539	70.5%	65.1%	52.9%	57.2%	0.984	0.673	–	–
MindEye2 (1 hour)	0.195	0.419	84.2%	90.6%	81.2%	79.2%	0.810	0.468	79.0%	57.4%

Table 1: Quantitative comparison of fMRI-to-image models. Results average across subjects 1, 2, 5, and 7 from the Natural Scenes Dataset. Results from all previous work were recalculated using their respective public codebases using the full 40 sessions of NSD data, which was not released until the recent completion of the 2023 Algonauts challenge. Image retrieval refers to the percent of the time the correct image was retrieved out of 300 candidates, given the associated brain sample (chance=0.3%); vice-versa for brain retrieval. PixCorr=pixelwise correlation between ground truth and reconstructions; SSIM=structural similarity index metric (Wang et al., 2004); EfficientNet-B1 ("Eff") (Tan and Le, 2020) and SwAV-ResNet50 ("SwAV") (Caron et al., 2021) refer to average correlation distance; all other metrics refer to two-way identification (chance = 50%). Two-way identification refers to percent correct across comparisons gauging if the original image embedding is more similar to its paired brain embedding or a randomly selected brain embedding (see Appendix A.9). Missing values are from metrics being non-applicable. Bold indicates best performance, underline second-best performance.

and retrieval performance (explored more in section 3.5). Interestingly, we observed that high-level metrics for the unrefined MindEye2 reconstructions outperformed the refined reconstructions across several metrics despite looking visibly distorted. This suggests that the standard evaluation metrics used across fMRI-to-image papers should be further scrutinized as they may not accurately reflect subjective interpretations of reconstruction quality.

We conducted behavioral experiments with online human raters to confirm that people subjectively prefer the refined reconstructions compared to the unrefined reconstructions (refined reconstructions preferred 71.94% of the time, $p < 0.001$). Human preference ratings also confirm SOTA performance compared to previous papers (correct reconstructions identified 97.82% of the time, $p < 0.001$), evaluated via two-alternative forced-choice judgments comparing ground truth images to MindEye2 reconstructions vs. random test set reconstructions. See Appendix A.15 for more details.

We also report performance for MindEye2 fine-tuned with only 1 hour of data in the same Table 1. We qualitatively compare reconstructions side-by-side with models trained on only 1 hour's worth of data in Figure 4, depicting improvements in reconstruction quality for MindEye2. We report more evaluations in the Appendix: see A.3 for MindEye2 results without pretraining, A.4 for evaluations with varying amounts of training data across all models, A.5 for single-subject evaluations, A.11 for MindEye2 evaluations with varying selection of pretraining subjects, and A.14 for visualization of the functional preferences of various brain regions of interest along the visual hierarchy (Serre et al., 2005) for each subject. We also conducted a behavioral experiment with human raters which confirmed that humans subjectively prefer MindEye2 (1-hour) reconstructions to Brain Diffuser (1-hour) reconstructions (Appendix A.15).

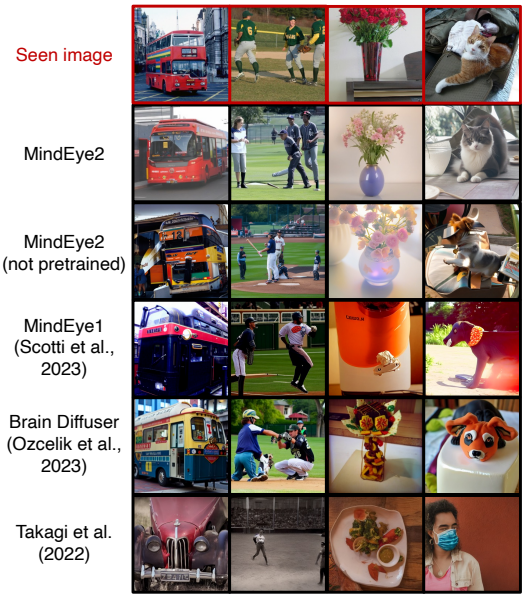

Figure 4: Reconstructions from different model approaches using 1 hour of training data from NSD.

3.1.1 Varying Amounts of Training Data

The overarching goal of the present work is to showcase high-quality reconstructions of seen images from a single visit to an MRI facility. Figure 5 shows reconstruction performance across MindEye2 models trained on varying amounts of data from subject 1. There is a steady improvement across both pretrained and non-pretrained models as more data is used to train the model. "Non-pretrained" refers to single-subject models trained from scratch. The pretrained and non-pretrained results became increasingly more similar as more data was added. The 1-hour setting offers a good balance between scan duration and recon-

struction performance, with notable improvements from pretraining. The non-pretrained models trained with 10 or 30 minutes of data suffered significant instability. These models may have experienced mode collapse where outputs were similarly nonsensical regardless of input. Such reconstructions coincidentally performed well on SSIM, indicating SSIM may not be a fully representative metric.

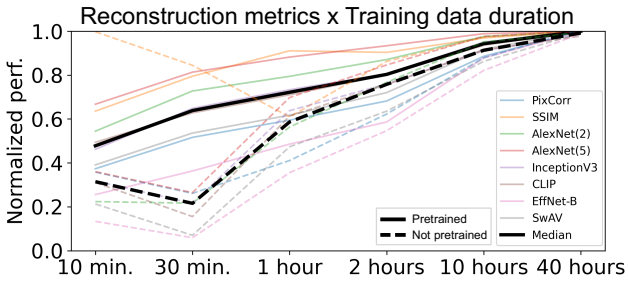

Figure 5: Normalized reconstruction metrics for MindEye2 with (connected) or without (dotted) pretraining on other subjects, using varying amounts of training/fine-tuning data. Normalization was such that 0 on the y-axis corresponds to metrics using random COCO images (not from NSD test set) as reconstructions and 1 corresponds to metrics using 40-session pretrained MindEye2. Black lines indicate median. Test data is the same across all comparisons (see section 3).

3.2 Image Captioning

Predicted image captions are quantitatively compared to previous work in Table 2. UniBrain (Mai and Zhang, 2023) was first to predict captions using NSD, training a diffusion model to predict CLIP ViT-L/14 text latents which get fed through a pretrained Optimus GPT2 model (Radford et al., 2019). Ferrante et al. (2023b) predicted image captions by mapping fMRI inputs to CLIP ViT-L/14 image latents via ridge regression, passing these latents through a pretrained GIT model (Wang et al., 2022).

We adopt the same caption metrics reported in the previous work. ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) capture aspects of text structure and composition. CLIP (Radford et al., 2021) and Sentence-Transformer ("all-MiniLM-L6-v2") (Reimers and Gurevych, 2020) are higher-level metrics that provide insight into textual context, relationships, and semantics. All metrics except ROUGE were calculated using the same code as Ferrante et al. (2023b). MindEye2 captioning performance outperformed previous models across all metrics except one, suggesting high-quality image captions from brain activity.

3.3 Image/Brain Retrieval

Image retrieval metrics help quantify the level of fine-grained image information contained in the fMRI embeddings. There are many images in the test set that contain similar semantic content (e.g., 14 images of zebras), so if the model can identify the exact image corresponding to a given brain sample, that demonstrates such fMRI embeddings contain fine-grained image content. MindEye2 improves upon MindEye1's retrieval evaluations by reaching near-ceiling performance on the retrieval benchmarks used in previous papers (Lin et al., 2022; Scotti et al., 2023) (Table 1). Further, retrieval performance remained competitive when MindEye2 was trained with only 1 hour of data.

	COCO captions		GIT captions	
Metric	MindEye2	UniBrain	MindEye2	Ferrante et al.
METEOR ↑	**0.248**	0.170	**0.344**	0.305
ROUGE-L ↑	**0.326**	0.225	**0.427**	-
ROUGE-1 ↑	**0.353**	0.247	**0.455**	-
Sentence ↑	**47.9%**	-	**52.3%**	44.7%
CLIP-B ↑	**73.7%**	-	**75.4%**	70.5%
CLIP-L ↑	63.8%	**86.1%**	**67.1%**	-

Table 2: FMRI-to-image caption evaluations. Previous works used different ground truth captions for comparison (COCO captions or captions generated from GIT), necessitating separate comparisons. Results were calculated exclusively on NSD subject 1. MindEye2 metrics come from the model trained on all 40 sessions of NSD data whereas previous work used 37 sessions.

Computing the retrieval metrics in Table 1 involved the following steps. The goal for brain retrieval is to identify the correct sample of brain activity that gave rise to the seen image out of a pool of brain samples. The seen image is converted to an OpenCLIP image embedding (or CLIP image embedding, depending on the contrastive space used in the paper) and cosine similarity is computed between its respective fMRI latent (e.g., from the retrieval submodule) as well as 299 other randomly selected fMRI latents in the test set. For each test sample, success is determined if the cosine similarity is greatest between the ground truth OpenCLIP/CLIP image embedding and its respective fMRI embedding (aka top-1 retrieval performance, chance=1/300). We specifically used 300 random samples because this was the approach used in previous work. We averaged retrieval performance across test samples and repeated the entire process 30 times to account for the variability in random sampling of batches. For image retrieval, the same procedure is used except image and brain samples are flipped such that the goal is to find the corresponding seen image in the image pool from the provided brain sample.

3.4 Brain Correlation

To measure whether a reconstruction is faithful to the original brain activity that evoked it, we examine whether it accurately predicts that brain activity when input to a encoding model pretrained to predict brain activity from images (Gaziv et al., 2022). Encoding models provide a more comprehensive analysis of the proximity between images and brain activity (Naselaris et al., 2011), providing a unique

measure of reconstruction quality that is perhaps more informative than the image metrics traditionally used for assessment. This alignment is measured independently of the stimulus image, allowing it to be used to assess reconstruction quality when the ground-truth image is unknown, making it extendable to new data in a variety of domains including covert visual content such as mental images. Given that human judgment is grounded in human brain activity, it could also be the case that brain correlation metrics provide increased alignment with the judgments of human observers. The brain correlation metrics in Table 3 are calculated with the GNet encoding model (St-Yves et al., 2022) using protocol from Kneeland et al. (2023c). "Unrefined" reconstructions performed best, perhaps because refinement sacrifices brain alignment (and reconstruction performance as assessed by some metrics) for the additional boost in perceptual alignment from enforcing a naturalistic prior.

Brain Region	MindEye2	MindEye2 (unrefined)	MindEye2 (1 hour)	Brain Diffuser	Takagi et al.
Visual cortex↑	0.373	**0.384**	0.348	0.381	0.247
V1↑	0.364	**0.385**	0.309	0.362	0.181
V2↑	0.352	**0.366**	0.314	0.340	0.152
V3↑	0.342	**0.353**	0.315	0.332	0.152
V4↑	0.327	**0.339**	0.300	0.323	0.170
Higher vis.↑	0.368	0.373	0.351	**0.375**	0.288

Table 3: Brain correlation scores calculated in different brain regions including visual cortex, early visual cortical regions V1, V2, V3, and V4, and higher visual areas (set complement of visual cortex and early visual cortex).

3.5 Ablations

Here we explain where MindEye2 improvements over MindEye1 come from through ablations. MindEye2 outperforms MindEye1 even without pretraining on other subjects (see Appendix A.3), suggesting improvements in model architecture and training procedure. The following ablation results compare models trained from scratch in reduced capacity (1024-dim shared-subject latent space), skipping base SDXL refinement, using 10 sessions of data solely from subject 1.

Two core differences between MindEye2 and MindEye1 are (1) we used a linear layer, rather than an MLP with dropout, for the initial mapping of voxels to the dimensionality of the residual MLP backbone, and (2) we map to OpenCLIP bigG image latents rather than CLIP L latents. Our ablations show that these changes improve performance across all metrics (Table 4), suggesting that a linear layer with L2 regularization is a more effective means of initially mapping voxels into model space, and that bigG is the richer, more effective CLIP space to map fMRI activity into.

Ablations in Table 5 show evaluations from models trained with various combinations of components. Retrieval metrics were worst when MindEye2 was trained with the diffusion prior and low-level submodules removed, and reconstruction metrics were worst when trained with the retrieval submodule and low-level submodule removed. This indicates that training MindEye2 with multiple objectives leads to mutually beneficial results.

Metric		ME2	ME1	CLIP L
Low-Level	PixCorr ↑	**0.292**	0.225	0.243
	SSIM ↑	**0.386**	0.380	0.371
	Alex(2) ↑	**92.7%**	87.3%	84.8%
	Alex(5) ↑	**97.6%**	94.7%	93.7%
High-Level	Incep ↑	**91.5%**	88.9%	87.7%
	CLIP ↑	**90.5%**	86.2%	89.2%
	Eff ↓	**0.700**	0.758	0.744
	SwAV ↓	**0.393**	0.430	0.427
Retrieval	Fwd ↑	**97.4%**	84.9%	89.6%
	Bwd ↑	**95.1%**	70.6%	82.8%

Table 4: Ablations on how MindEye2 (ME2) improves upon MindEye1. "ME1" results replace the initial linear mapping of fMRI voxels with MindEye1's MLP with dropout. "CLIP L" results map voxels to CLIP L (reconstructions via Versatile Diffusion) instead of OpenCLIP bigG (reconstructions via SDXL unCLIP).

Metric		Prior	Prior+Low	Prior+Ret.	All
Low-Level	PixCorr ↑	0.155	**0.281**	0.233	0.267
	SSIM ↑	0.309	**0.385**	0.319	0.380
	Alex(2) ↑	79.6%	89.4%	**90.6%**	89.7%
	Alex(5) ↑	88.6%	96.2%	**96.8%**	96.4%
High-Level	Incep ↑	85.3%	91.5%	**91.9%**	91.4%
	CLIP ↑	79.5%	88.4%	**89.4%**	87.9%
	Eff ↓	0.805	0.727	**0.717**	0.732
	SwAV ↓	0.490	0.416	**0.410**	0.415
		Ret.	Ret.+Low	Prior.+Ret.	All
Retrieval	Fwd ↑	96.5%	96.9%	96.2%	**98.0%**
	Bwd ↑	92.4%	93.0%	**95.8%**	94.1%

Table 5: Ablations compare reconstruction and retrieval metrics for MindEye2 trained with various combinations of model components. Retr.=Retrieval submodule, Low=Low-level submodule.

4 Related Work

It is common for fMRI analyses to align subjects' brains to a shared space for the purposes of increasing statistical power and/or assessing generality of scientific findings. Such alignment is difficult because structural and functional topography differs substantially across people (Talairach and Tournoux, 1990; Mazziotta et al., 2001). Anatomical alignment, where brains are physically warped into a predefined brain template, is imperfect and can potentially distort functional alignment (Fischl et al., 1999; Sabuncu et al., 2010; Conroy et al., 2009; Thual et al., 2022). Functional alignment can ignore anatomical structure and focus specifically on finding shared patterns of brain activity. There are many approaches to functional alignment but typically they involve subjects experiencing shared stimuli and then using responses to these stimuli to learn an alignment mapping (Chen et al., 2015; Haxby et al., 2011; 2020; Huang et al., 2021; Nastase et al., 2019; Busch et al., 2021). While it is useful to conduct such experiments to identify sources of shared signal across subjects, it is also limiting in that

new subjects would need to be scanned seeing the same images. Other functional alignment approaches avoid such limitations by using self-supervised learning to identify an initial generalizable embedding space with outputs suitable for downstream tasks (Schneider et al., 2023; Chen et al., 2023a;b). Closest to our alignment approach are models that adopt both shared-subject and subject-specific mappings in their model architecture (Défossez et al., 2022; Benchetrit et al., 2023; Yang et al., 2023; Lane and Kiar, 2023).

Ferrante et al. (2023a) previously showed across-subject image reconstruction via ridge regression by training a linear subject-specific decoding model and then separately mapping other subjects to this space via ridge regression. This is similar to our approach in that both involve ridge regression to a shared space, but is distinct in that their approach is capped by the performance of the initial single-subject model from which other subjects are mapped into, is restricted to only linear fine-tuning, and was demonstrated only with a reduced training dataset of images seen by all subjects. MindEye2 is unique in its demonstration that a single neural network model can be pretrained across subjects experiencing unique stimuli and robustly fine-tuned to a new subject with few data points.

5 Conclusion

We introduce MindEye2, a modeling approach that outputs reconstructions of seen images from fMRI activity with a similar quality to previous approaches using only a fraction of the training data. MindEye2 further achieves SOTA across reconstruction and retrieval metrics when supplied with the full training data. Our approach pretrains a model using data from multiple subjects, which is then fine-tuned on scarce data from a held-out subject. Patterns of fMRI activity are mapped to CLIP space and images are reconstructed with the help of our unCLIP model fine-tuned from Stable Diffusion XL. Our work shows the potential to apply deep learning models trained on large-scale neuroimaging datasets to new subjects with minimal data.

5.1 Limitations

fMRI is extremely sensitive to movement and requires subjects to comply with the task: decoding is easily resisted by slightly moving one's head or thinking about unrelated information (Tang et al., 2023). MindEye2 has also only been shown to work on natural scenes such as those in COCO; additional data and/or specialized generative models would likely be required for other image distributions.

5.2 Impact Statement

The present work demonstrates that it is now practical for patients to undergo a single MRI scanning session and produce enough data to perform high-quality reconstructions of their visual perception. Such image reconstructions from brain activity are expected to be systematically distorted due to factors including mental state, neurological conditions, etc. This could potentially enable novel clinical diagnosis and assessment approaches, including applications for improved locked-in (pseudocoma) patient communication (Monti et al., 2010) and brain-computer interfaces if adapted to real-time analysis (Wallace et al., 2022) or non-fMRI neuroimaging modalities. Future work could potentially generalize reconstruction models from perception to mental imagery without training a new model (Stokes et al., 2009; Goebel et al., 2022; Naselaris et al., 2015; Reddy et al., 2010). As technology continues to improve, we note it is important that brain data be carefully protected and companies collecting such data be transparent with their use.

6 Author Contributions

For detailed author contributions see Appendix A.1.

7 Acknowledgements

Special thanks to Dustin Podell, Vikram Voleti, Andreas Blattmann, and Robin Rombach for technical assistance fine-tuning Stable Diffusion XL to support our unCLIP use-case. Thanks to the MedARC Discord community for being the public forum from which this research was developed, particularly thank you to Connor Lane, Alex Nguyen, Atmadeep Bannerjee, Amir Refaee, and Mohammed Baharoon for their helpful discussions. Thanks to Alessandro Gifford and Connor Lane for providing useful feedback on drafts of the manuscript. Thank you to Richard Vencu for help navigating the Stability AI HPC. Thanks to Stability AI for their support for open neuroAI research and providing the computational resources necessary to develop MindEye2. Collection of the Natural Scenes Dataset was supported by NSF IIS-1822683 and NSF IIS-1822929.

References

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, February 2021. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models, April 2022. URL http://arxiv.org/abs/2112.10752. arXiv:2112.10752 [cs].

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.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. preprint, Neuroscience, November 2022. URL http://biorxiv.org/lookup/doi/10.1101/2022.11.18.517004.

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

Furkan Ozcelik, Bhavin Choksi, Milad Mozafari, Leila Reddy, and Rufin VanRullen. Reconstruction of Perceived Images from fMRI Patterns and Semantic Brain Exploration using Instance-Conditioned GANs, February 2022. URL http://arxiv.org/abs/2202.12692. arXiv:2202.12692 [cs, eess, q-bio].

Furkan Ozcelik and Rufin VanRullen. Brain-Diffuser: Natural scene reconstruction from fMRI signals using generative latent diffusion, March 2023. URL http://arxiv.org/abs/2303.05334. arXiv:2303.05334 [cs, q-bio].

Guy Gaziv, Roman Beliy, Niv Granot, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. Self-supervised Natural Image Reconstruction and Large-scale Semantic Classification from Brain Activity. *NeuroImage*, 254:119121, July 2022. ISSN 10538119. doi: 10.1016/j.neuroimage.2022.119121. URL https://linkinghub.elsevier.com/retrieve/pii/S105381192200249X.

Zijin Gu, Keith Jamison, Amy Kuceyeski, and Mert Sabuncu. Decoding natural image stimuli from fMRI data with a surface-based convolutional network, March 2023. URL http://arxiv.org/abs/2212.02409. arXiv:2212.02409 [cs, q-bio].

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, and Tanishq Abraham. Reconstructing the Mind's Eye: fMRI-to-Image with Contrastive Learning and Diffusion Priors. *Advances in Neural Information Processing Systems*, 36:24705–24728, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4ddab70bf41ffe5d423840644d3357f4-Abstract-Conference.html.

Reese Kneeland, Jordyn Ojeda, Ghislain St-Yves, and Thomas Naselaris. Reconstructing seen images from human brain activity via guided stochastic search. Conference on Cognitive Computational Neuroscience, 2023a. doi: 10.32470/CCN.2023.1672-0.

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 2023b. URL http://arxiv.org/abs/2306.00927. arXiv:2306.00927 [cs, q-bio].

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

Matteo Ferrante, Tommaso Boccato, and Nicola Toschi. Through their eyes: multi-subject Brain Decoding with simple alignment techniques, August 2023a. URL http://arxiv.org/abs/2309.00627. arXiv:2309.00627 [cs, q-bio].

Alexis Thual, Yohann Benchetrit, Felix Geilert, Jérémy Rapin, Iurii Makarov, Hubert Banville, and Jean-Rémi King. Aligning brain functions boosts the decoding of visual semantics in novel subjects, December 2023. URL http://arxiv.org/abs/2312.06467. arXiv:2312.06467 [cs, eess, q-bio].

Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic Mindscapes: High-quality Video Reconstruction from Brain Activity, May 2023a. URL http://arxiv.org/abs/2305.11675. arXiv:2305.11675 [cs].

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, March 2023b. URL http://arxiv.org/abs/2211.06956. arXiv:2211.06956 [cs].

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].

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].

Weihao Xia, Raoul de Charette, Cengiz Öztireli, and Jing-Hao Xue. DREAM: Visual Decoding from Reversing Human Visual System, October 2023. URL http://arxiv.org/abs/2310.02265. arXiv:2310.02265 [cs, eess, q-bio].

Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fMRI, July 2019. URL http://arxiv.org/abs/1907.02431. arXiv:1907.02431 [cs, eess, q-bio, stat].

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

Guohua Shen, Kshitij Dwivedi, Kei Majima, Tomoyasu Horikawa, and Yukiyasu Kamitani. End-to-End Deep Image Reconstruction From Human Brain Activity. *Frontiers in Computational Neuroscience*, 13, 2019b. ISSN 1662-5188. URL https://www.frontiersin.org/articles/10.3389/fncom.2019.00021.

K. Seeliger, U. Güçlü, L. Ambrogioni, Y. Güçlütürk, and M.A.J. van Gerven. Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, 181:775–785, November 2018. ISSN 10538119. doi: 10.1016/j.neuroimage.2018.07.043. URL https://linkinghub.elsevier.com/retrieve/pii/S105381191830658X.

Yunfeng Lin, Jiangbei Li, and Hanjing Wang. DCNN-GAN: Reconstructing Realistic Image from fMRI, January 2019. URL http://arxiv.org/abs/1901.07368. arXiv:1901.07368 [cs, eess].

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents, April 2022. URL http://arxiv.org/abs/2204.06125. arXiv:2204.06125 [cs].

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models, October 2022. URL http://arxiv.org/abs/2210.08402. arXiv:2210.08402 [cs].

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, July 2021. URL https://doi.org/10.5281/zenodo.5143773.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. Accelerate: Training and inference at scale made simple, efficient and adaptable., 2022. URL https://github.com/huggingface/accelerate.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory Optimizations Toward Training Trillion Parameter Models, May 2020. URL http://arxiv.org/abs/1910.02054. arXiv:1910.02054 [cs, stat].

J. Talairach and P. Tournoux. Co-planar stereotaxic atlas of the human brain. 3-Dimensional proportional system: an approach to cerebral imaging. *The Journal of Laryngology & Otology*, 104(1):72–72, January 1990. ISSN 1748-5460, 0022-2151. doi: 10.1017/S0022215100111879. URL https://www.cambridge.org/core/journals/journal-of-laryngology-and-otology/article/abs/co-planar-stereotaxic-atlas-of-the-human-brain-3-dimensional-proportional-system-an-approach-to-cerebral-imaging-1988talairichj-and-tournouxprayportmarkgeorg-thieme-verlag-stuttgart-new-york3-13-711-701-1-price-dm-268-pp-122-illustrations-130/46C98B7A1D9ABB728CB5A5709C09AF89. Publisher: Cambridge University Press.

J Mazziotta, A Toga, A Evans, P Fox, J Lancaster, K Zilles, R Woods, T Paus, G Simpson, B Pike, C Holmes, L Collins, P Thompson, D MacDonald, M Iacoboni, T Schormann, K Amunts, N Palomero-Gallagher, S Geyer, L Parsons, K Narr, N Kabani, G Le Goualher, D Boomsma, T Cannon, R Kawashima, and B Mazoyer. A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). *Philosophical Transactions of the Royal Society of London. Series B*, 356(1412):1293–1322, August 2001. ISSN 0962-8436. doi: 10.1098/rstb.2001.0915. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1088516/.

James V. Haxby, J. Swaroop Guntupalli, Andrew C. Connolly, Yaroslav O. Halchenko, Bryan R. Conroy, M. Ida Gobbini, Michael Hanke, and Peter J. Ramadge. A Common, High-Dimensional Model of the Representational Space in Human Ventral Temporal Cortex. *Neuron*, 72(2):404–416, October 2011. ISSN 08966273. doi: 10.1016/j.neuron.2011.08.026. URL https://linkinghub.elsevier.com/retrieve/pii/S0896627311007811.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. LiT: Zero-Shot Transfer with Locked-image text Tuning, June 2022. URL http://arxiv.org/abs/2111.07991. arXiv:2111.07991 [cs].

Sungnyun Kim, Gihun Lee, Sangmin Bae, and Seyoung Yun. Mixco: Mix-up contrastive learning for visual representation. *ArXiv*, abs/2010.06300, 2020.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicregl: Self-supervised learning of local visual features. *ArXiv*, abs/2210.01571, 2022.

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, December 2022. URL http://arxiv.org/abs/2205.14100. arXiv:2205.14100 [cs].

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

Xingqian Xu, Zhangyang Wang, Eric Zhang, Kai Wang, and Humphrey Shi. Versatile Diffusion: Text, Images and Variations All in One Diffusion Model, March 2023. URL http://arxiv.org/abs/2211.08332. arXiv:2211.08332 [cs].

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models, August 2023. URL http://arxiv.org/abs/2308.06721. arXiv:2308.06721 [cs].

Justin Pinkney. Lambda Diffusers, 2022. URL https://github.com/LambdaLabsML/lambda-diffusers. publicationType: misc; publisher: GitHub; journal: GitHub repository.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations, January 2022. URL http://arxiv.org/abs/2108.01073. arXiv:2108.01073 [cs].

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, July 2023. URL http://arxiv.org/abs/2307.01952. arXiv:2307.01952 [cs].

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common Diffusion Noise Schedules and Sample Steps are Flawed, January 2024. URL http://arxiv.org/abs/2305.08891. arXiv:2305.08891 [cs].

Nicholas Guttenberg. Diffusion with Offset Noise, 2023. URL https://www.crosslabs.org//blog/diffusion-with-offset-noise.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1. doi: 10.1007/978-3-319-10602-1_48.

T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio. A Theory of Object Recognition: Computations and Circuits in the Feedforward Path of the Ventral Stream in Primate Visual Cortex. December 2005. URL https://dspace.mit.edu/handle/1721.1/36407. Accepted: 2007-03-12T16:41:47Z.

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.

Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, September 2020. URL http://arxiv.org/abs/1905.11946. arXiv:1905.11946 [cs, stat].

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments, January 2021. URL http://arxiv.org/abs/2006.09882. arXiv:2006.09882 [cs].

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. 2019.

Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909.

Nils Reimers and Iryna Gurevych. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation, October 2020. URL http://arxiv.org/abs/2004.09813. arXiv:2004.09813 [cs].

Sikun Lin, Thomas Sprague, and Ambuj K. Singh. Mind Reader: Reconstructing complex images from brain activities, September 2022. URL http://arxiv.org/abs/2210.01769. arXiv:2210.01769 [cs, eess, q-bio].

Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2), 2011. doi: 10.1016/j.neuroimage.2010.07.073.

Ghislain St-Yves, Emily J. Allen, Yihan Wu, Kendrick Kay, and Thomas Naselaris. Brain-optimized neural networks learn non-hierarchical models of representation in human visual cortex. *bioRxiv*, 2022. doi: 10.1101/2022.01.21.477293.

B. Fischl, M. I. Sereno, R. B. Tootell, and A. M. Dale. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping*, 8(4):272–284, 1999. ISSN 1065-9471. doi: 10.1002/(sici)1097-0193(1999)8:4<272::aid-hbm10>3.0.co;2-4.

Mert R. Sabuncu, Benjamin D. Singer, Bryan Conroy, Ronald E. Bryan, Peter Jeffrey Ramadge, and James V. Haxby. Function-based intersubject alignment of human cortical anatomy. *Cerebral Cortex*, 20(1):130–140, January 2010. ISSN 1047-3211. doi: 10.1093/cercor/bhp085. URL https://collaborate.princeton.edu/en/publications/function-based-intersubject-alignment-of-human-cortical-anatomy. Publisher: Oxford University Press.

Bryan R. Conroy, Benjamin D. Singer, James V. Haxby, and Peter J. Ramadge. fMRI-Based Inter-Subject Cortical Alignment Using Functional Connectivity. *Advances in neural information processing systems*, 22:378–386, 2009. ISSN 1049-5258. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4572745/.

Alexis Thual, Quang Huy Tran, Tatiana Zemskova, Nicolas Courty, Rémi Flamary, Stanislas Dehaene, and Bertrand Thirion. Aligning individual brains with fused unbalanced Gromov Wasserstein. *Advances in Neural Information Processing Systems*, 35:21792–21804, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/8906cac4ca58dcaf17e97a0486ad57ca-Abstract-Conference.html.

Po-Hsuan (Cameron) Chen, Janice Chen, Yaara Yeshurun, Uri Hasson, James Haxby, and Peter J Ramadge. A Reduced-Dimension fMRI Shared Response Model. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://papers.nips.cc/paper_files/paper/2015/hash/b3967a0e938dc2a6340e258630febd5a-Abstract.html.

James V Haxby, J Swaroop Guntupalli, Samuel A Nastase, and Ma Feilong. Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *eLife*, 9:e56601, June 2020. ISSN 2050-084X. doi: 10.7554/eLife.56601. URL https://elifesciences.org/articles/56601.

Jessie Huang, Erica L. Busch, Tom Wallenstein, Michal Gerasimiuk, Andrew Benz, Guillaume Lajoie, Guy Wolf, Nicholas B. Turk-Browne, and Smita Krishnaswamy. Learning shared neural manifolds from multi-subject FMRI data, December 2021. URL http://arxiv.org/abs/2201.00622. arXiv:2201.00622 [cs, eess, q-bio].

Samuel A Nastase, Valeria Gazzola, Uri Hasson, and Christian Keysers. Measuring shared responses across subjects using intersubject correlation. *Social Cognitive and Affective Neuroscience*, page nsz037, May 2019. ISSN 1749-5016, 1749-5024. doi: 10.1093/scan/nsz037. URL https://academic.oup.com/scan/advance-article/doi/10.1093/scan/nsz037/5489905.

Erica L. Busch, Lukas Slipski, Ma Feilong, J. Swaroop Guntupalli, Matteo Visconti di Oleggio Castello, Jeremy F. Huckins, Samuel A. Nastase, M. Ida Gobbini, Tor D. Wager, and James V. Haxby. Hybrid hyperalignment: A single high-dimensional model of shared information embedded in cortical patterns of response and functional connectivity. *NeuroImage*, 233:117975, June 2021. ISSN 10538119. doi: 10.1016/j.neuroimage.2021.117975. URL https://linkinghub.elsevier.com/retrieve/pii/S1053811921002524.

Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617(7960):360–368, May 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06031-6. URL https://www.nature.com/articles/s41586-023-06031-6. Number: 7960 Publisher: Nature Publishing Group.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech from non-invasive brain recordings, August 2022. URL http://arxiv.org/abs/2208.12266. arXiv:2208.12266 [cs, eess, q-bio].

Yohann Benchetrit, Hubert Banville, and Jean-Remi King. Brain decoding: toward real-time reconstruction of visual perception. October 2023. URL https://openreview.net/forum?id=3y1K6buO8c.

Huzheng Yang, James Gee, and Jianbo Shi. Memory Encoding Model, August 2023. URL http://arxiv.org/abs/2308.01175. arXiv:2308.01175 [cs].

Connor Lane and Gregory Kiar. A Parameter-efficient Multi-subject Model for Predicting fMRI Activity, August 2023. URL https://arxiv.org/abs/2308.02351v1.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, pages 1–9, May 2023. ISSN 1546-1726. doi: 10.1038/s41593-023-01304-9. URL https://www.nature.com/articles/s41593-023-01304-9. Publisher: Nature Publishing Group.

Martin M. Monti, Audrey Vanhaudenhuyse, Martin R. Coleman, Melanie Boly, John D. Pickard, Luaba Tshibanda, Adrian M. Owen, and Steven Laureys. Willful Modulation of Brain Activity in Disorders of Consciousness. *New England Journal of Medicine*, 362(7):579–589, February 2010. ISSN 0028-4793. doi: 10.1056/NEJMoa0905370. URL https://doi.org/10.1056/NEJMoa0905370. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMoa0905370.

Grant Wallace, Stephen Polcyn, Paula P. Brooks, Anne C. Mennen, Ke Zhao, Paul S. Scotti, Sebastian Michelmann, Kai Li, Nicholas B. Turk-Browne, Jonathan D. Cohen, and Kenneth A. Norman. RT-Cloud: A cloud-based software framework to simplify and standardize real-time fMRI. *NeuroImage*, 257:119295, August 2022. ISSN 10538119. doi: 10.1016/j.neuroimage.2022.119295. URL https://linkinghub.elsevier.com/retrieve/pii/S1053811922004141.

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.

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.

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.

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.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742. PMLR, July 2023. URL https://proceedings.mlr.press/v202/li23q.html. ISSN: 2640-3498.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training, September 2023. URL http://arxiv.org/abs/2303.15343. arXiv:2303.15343 [cs].

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, November 2022. ISSN 2050-084X. doi: 10.7554/eLife.77599. URL https://doi.org/10.7554/eLife.77599. Publisher: eLife Sciences Publications, Ltd.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, January 2018. URL http://arxiv.org/abs/1706.08500. arXiv:1706.08500 [cs, stat].

Romain Beaumont. Clip Retrieval: Easily compute clip embeddings and build a clip retrieval system with them, 2022. URL https://github.com/rom1504/clip-retrieval. publicationType: misc; publisher: GitHub; journal: GitHub repository.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The Faiss library. 2024. _eprint: 2401.08281.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, June 2016. doi: 10.1109/CVPR.2016.308. ISSN: 1063-6919.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, September 2020. URL http://arxiv.org/abs/1802.03426. arXiv:1802.03426 [cs, stat].

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning, October 2022. URL http://arxiv.org/abs/2203.02053. arXiv:2203.02053 [cs].

Ken Shirakawa. Umap visualization of CLIP text feature, July 2023. URL https://sore-holly-400.notion.site/230704-Shirakawa-Umap-visualization-of-CLIP-text-feature-281cbcaa681b427f92285d5ef227d665.

Andrew F. Luo, Margaret M. Henderson, Leila Wehbe, and Michael J. Tarr. Brain Diffusion for Visual Exploration: Cortical Discovery using Large Scale Generative Models, November 2023a. URL http://arxiv.org/abs/2306.03089. arXiv:2306.03089 [cs].

Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Logan T. Dowdle, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive and computational neuroscience. preprint, Neuroscience, February 2021. URL http://biorxiv.org/lookup/doi/10.1101/2021.02.22.432340.

Gabriel H. Sarch, Michael J. Tarr, Katerina Fragkiadaki, and Leila Wehbe. Brain Dissection: fMRI-trained Networks Reveal Spatial Selectivity in the Processing of Natural Images, May 2023. URL https://www.biorxiv.org/content/10.1101/2023.05.29.542635v1. Pages: 2023.05.29.542635 Section: New Results.

Andrew Luo, Margaret Marie Henderson, Michael J. Tarr, and Leila Wehbe. BrainSCUBA: Fine-Grained Natural Language Captions of Visual Cortex Selectivity. October 2023b. URL https://openreview.net/forum?id=mQYHXUUTkU&referrer=%5Bthe%20profile%20of%20Leila%20Wehbe%5D(%2Fprofile%3Fid%3D~Leila_Wehbe1).

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting Stable Diffusion Using Cross Attention, December 2022. URL http://arxiv.org/abs/2210.04885. arXiv:2210.04885 [cs].

Alexander G. Huth, Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, April 2016. ISSN 1476-4687. doi: 10.1038/nature17637. URL https://www.nature.com/articles/nature17637. Publisher: Nature Publishing Group.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment, October 2023. URL https://arxiv.org/abs/2310.13018v2.

A Appendix

A.1 Author Contributions

PSS: project lead, drafted the initial manuscript and contributed to all parts of MindEye2 development. MT (core contributor): MindEye2 ablations, SDXL unCLIP vs. Versatile Diffusion comparisons, improved distributed training code, and experimented with approaches not used in the final model including training custom ControlNet and T2I adapters, using retrieval on COCO CLIP captions, and using diffusion priors to align fMRI to text embeddings. CKTV (core contributor): retrained and evaluated MindEye1 models, image captioning evaluations and writing, improved manuscript formatting, ROI-optimized stimuli experiments, and experimented with approaches not used in the final model including trying out different pretrained model embeddings, experimenting with T2I-Adapters and depth conditioning, experimenting with using past/future timepoints as additional conditioning, experimenting with blip2 (Li et al., 2023) for text prediction, and experimenting with behavioral embeddings. RK (core contributor): brain correlations, human preference experiments, recalculated metrics for 40-hour setting Ozcelik and VanRullen (2023) and Takagi and Nishimoto (2023) results, evaluations with varying amounts of training data across all models, assistance with data normalization, significant contributions to manuscript writing. TC: UMAP visualizations, improved the design for Figure 1, and experimented with approaches not used in the final model including using past/future timepoints as additional conditioning and using flattened voxels in MNI space instead of native space. AN: helped with ablations and experimented with replacing soft CLIP loss with soft SigLIP loss (Zhai et al., 2023) (not used in final model). CS: FAISS retrieval with MS-COCO (Appendix A.8) and experimented with approaches not used in the final model including experimenting with using past/future timepoints as additional conditioning, experimenting with blip2 for text prediction, and experimenting with behavioral embeddings. JX: helped with ablations, manuscript revisions and table formatting, experimented with approaches not used in the final model including experimenting with blip2 for text prediction, experimenting with behavioral embeddings, and improving model architecture. TN: assisted with human preference experiments. KN: oversaw the project, manuscript revisions and framing. TMA: oversaw the project, manuscript revisions and framing.

A.2 Additional Dataset Information

fMRI responses correspond to normalized single-trial betas output from GLMSingle (Prince et al., 2022). We use preprocessed flattened fMRI voxels in 1.8-mm native volume space corresponding to 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). MindEye2 was developed using a training and test set of subject 1's data, with other subjects' data untouched until final training of models. The fMRI data from both the training and test set was normalized using a voxel-wise Z-scoring procedure using the mean and standard deviation calculated using only the training set. Despite the shared1000 test trials being distributed across the scanning sessions for each subject, we chose to keep the test set consistent no matter the number of sessions being used for training. We also adjusted the number of training sessions after the normalization step, allowing us to keep the statistical properties of the shared1000 test set consistent between experiments with varying amounts of training data. This may inadvertently give a small normalization advantage to models trained with fewer training sessions, as the models are normalized with additional data not made available for training.

A.3 MindEye2 (not pretrained) vs. MindEye1

Table 6 shows how MindEye2 outperforms MindEye1 even without pretraining on other subjects. Models were trained using the full 40 sessions of training data from subject 1. This suggests that improvements from MindEye1 to MindEye2 are not explained solely from pretraining on other subjects, but that benefits also come from improved model architecture and training procedure.

Method		MindEye2	MindEye1
Low-Level	PixCorr ↑	0.376	**0.388**
	SSIM ↑	**0.440**	0.355
	Alex(2) ↑	**97.5%**	96.1%
	Alex(5) ↑	**99.1%**	98.3%
High-Level	Incep ↑	**95.4%**	95.0%
	CLIP ↑	92.6%	**93.7%**
	Eff ↓	**0.612**	0.635
	SwAV ↓	**0.341**	0.360
Retrieval	Fwd ↑	**100.0%**	95.0%
	Bwd ↑	**99.7%**	89.4%
Brain Corr	NSD General ↑	**0.370**	0.353
	V1 ↑	**0.383**	0.349
	V2 ↑	**0.373**	0.336
	V3 ↑	**0.363**	0.328
	V4 ↑	**0.335**	0.307
	Higher vis. ↑	**0.356**	0.345

Table 6: Performance comparison between MindEye2 (refined) and MindEye1 both trained from scratch across all 40 NSD sessions using only subject 1 data.

A.4 Reconstruction Evaluations Across Varying Amounts of Training Data

Here we present a further analysis of how model performance scales with training data. All of the results presented in Figures 6, 7, and 8 are calculated on only subject 1.

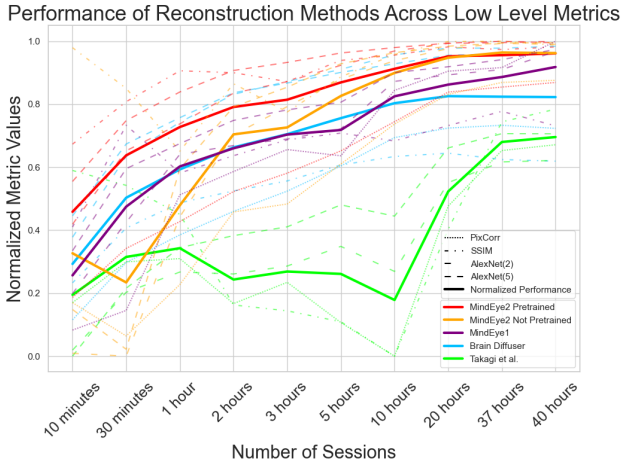

Figure 6: Low-level metric performance (y-axis) plotted against the number of fMRI scanning sessions used in the training data (x-axis) for subject 1. All values are normalized to the same y-axis. The bolded line represents the average performance across all metrics.

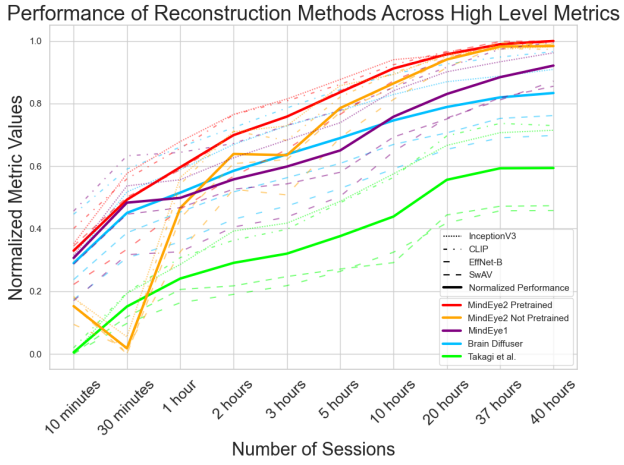

Figure 7: High-level metric performance (y-axis) plotted against the number of fMRI scanning sessions used in the training data (x-axis) for subject 1. All values are normalized to the same y-axis. The bolded line represents the average performance across all metrics. SwAV and EffNet-B scores are inverted in this plot so that higher is better for all metrics.

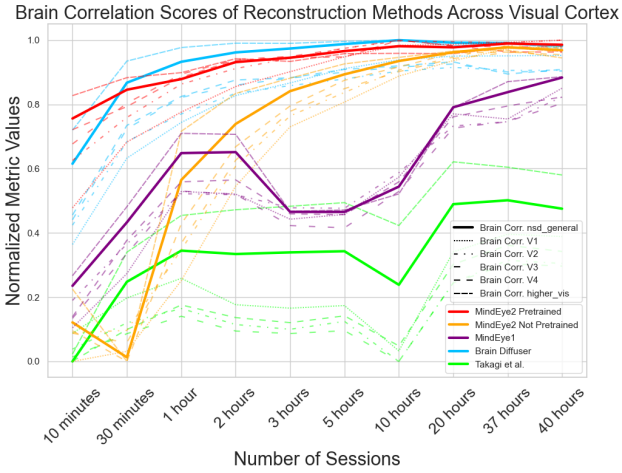

Figure 8: Brain correlation scores (y-axis) in different brain regions including visual cortex (defined by the nsdgeneral mask, bolded), V1, V2, V3, V4 (collectively called early visual cortex) and higher visual areas (the set complement of nsdgeneral and early visual cortex) plotted against the number of fMRI scanning sessions used in the training data (x-axis) for subject 1. All values are normalized to the same y-axis.

A.5 Single-Subject Evaluations

Tables 7 and 8 show more exhaustive evaluation metrics computed for every subject individually using 40-hours and 1-hour of fine-tuning data respectively.

40 Session Subject Results		Subject 1	Subject 2	Subject 5	Subject 7
Low	PixCorr ↑	**0.374**	0.328	0.301	0.283
	SSIM ↑	**0.439**	0.430	0.432	0.423
	Alex(2) ↑	**97.82%**	97.01%	95.32%	94.25%
	Alex(5) ↑	**99.10%**	98.83%	98.54%	97.97%
High	Incep ↑	96.15%	94.90%	**96.52%**	94.09%
	CLIP ↑	93.56%	91.66%	**94.32%**	92.36%
	Eff ↓	0.609	0.631	**0.600**	0.638
	SwAV ↓	0.338	0.347	**0.335**	0.357
Retrieval	Image ↑	**99.96%**	99.88%	98.39%	96.89%
	Brain ↑	**99.87%**	99.84%	96.94%	96.53%
Brain Corr.	Visual cortex ↑	0.374	0.387	**0.413**	0.317
	V1 ↑	0.389	**0.391**	0.354	0.321
	V2 ↑	**0.381**	0.353	0.359	0.314
	V3 ↑	**0.367**	0.362	0.340	0.299
	V4 ↑	0.337	**0.374**	0.321	0.278
	Higher vis. ↑	0.361	0.380	**0.424**	0.309
Captions	METEOR ↑	0.248	0.245	**0.250**	0.240
	ROUGE-L ↑	0.326	0.321	**0.327**	0.319
	ROUGE-1 ↑	0.353	0.349	**0.354**	0.347
	Sentence ↑	47.95%	46.69%	**49.40%**	46.97%
	CLIP-B ↑	73.74%	73.15%	**74.22%**	73.16%
	CLIP-L ↑	63.76%	62.96%	**64.14%**	62.86%

Table 7: Single subject quantitative results for 40 sessions of training data.

1 Session Subject Results		Subject 1	Subject 2	Subject 5	Subject 7
Low	PixCorr ↑	**0.235**	0.200	0.175	0.170
	SSIM ↑	0.428	**0.433**	0.405	0.408
	Alex(2) ↑	88.02%	**85.00%**	83.11%	80.70%
	Alex(5) ↑	**93.33%**	92.13%	91.00%	85.90%
High	Incep ↑	83.56%	81.86%	**84.33%**	74.90%
	CLIP ↑	80.75%	79.39%	**82.53%**	74.29%
	Eff ↓	0.798	0.807	**0.781**	0.854
	SwAV ↓	0.459	0.467	**0.444**	0.504
Retrieval	Image ↑	**93.96%**	90.53%	66.94%	64.44%
	Brain ↑	**77.63%**	67.18%	46.96%	37.77%
Brain Correlation	Visual cortex ↑	0.347	0.350	**0.404**	0.294
	V1 ↑	0.318	0.306	**0.328**	0.283
	V2 ↑	**0.337**	0.296	0.336	0.285
	V3 ↑	**0.341**	0.323	0.323	0.272
	V4 ↑	0.316	**0.336**	0.304	0.243
	Higher vis. ↑	0.345	0.357	**0.415**	0.285
Captions	METEOR ↑	0.200	0.200	**0.207**	0.189
	ROUGE-L ↑	0.278	0.272	**0.280**	0.260
	ROUGE-1 ↑	0.299	0.293	**0.300**	0.279
	Sentence ↑	33.52%	32.36%	**35.12%**	28.00%
	CLIP-B ↑	67.22%	65.98%	**67.63%**	63.15%
	CLIP-L ↑	55.44%	54.00%	**56.19%**	50.60%

Table 8: Single subject quantitative results for 1 session of training data.

A.6 UnCLIP Evaluation

Previous fMRI-to-image papers (Scotti et al., 2023; Ozcelik and VanRullen, 2023; Mai and Zhang, 2023) opted for Versatile Diffusion because it was state-of-the-art in reconstructing images from CLIP image latents with little variation. To compare the image generation capabilities of our unCLIP model with Versatile Diffusion, we computed Fréchet inception distance (FID) (Heusel et al., 2018) scores across 30,000 randomly sampled images from the COCO 2017 validation set. The images were center-cropped and scaled to 480×480 resolution. For Versatile Diffusion, we used Huggingface's VersatileDiffusionDualGuidedPipeline with text_to_image set to 0 to not take any input from text.

Our unCLIP model fine-tuned from Stable Diffusion XL outperforms Versatile Diffusion in terms of returning the original image from CLIP latents (see Appendix 9). This difference is visually obvious as shown in Figure 9. Note that while we observed distortions in our unrefined fMRI-to-image reconstructions using our unCLIP model fine-tuned from SDXL, such distortions were rare when using the ground truth CLIP embeddings.

The ability for this unCLIP model to nearly perfectly return the original image also indicates that OpenCLIP ViT-bigG image embeddings effectively preserve the majority of the information inherent in the original pixel image, retaining both low-level structure and high-level semantic details.

Figure 9: Generating images from their CLIP image embeddings. SDXL unCLIP (middle) outperforms Versatile Diffusion (right) in capturing perceptual details.

Metrics	SDXL unCLIP	VD
FID ↓	**13.69**	22.04
PixCorr ↑	**0.676**	0.266
SSIM ↑	**0.232**	0.055
Alex(2) ↑	**0.998**	0.972
Alex(5) ↑	**0.998**	0.966
Incep ↑	**0.997**	0.994
CLIP ↑	**0.999**	0.997
Eff ↓	**0.240**	0.487
SwAV ↓	**0.029**	0.108

Table 9: SDXL unCLIP reconstructions from ground truth OpenCLIP image latents consistently outperform Versatile Diffusion reconstructions from ground truth CLIP image latents.

A.7 OpenCLIP BigG to CLIP L Conversion

To map from OpenCLIP ViT-bigG/14 image latents to CLIP ViT-L/14 image latents during MindEye2 inference we independently trained a linear model using ground truth images from the COCO 2017 train and validation dataset. This

conversion was necessary to use the pretrained GIT image captioning model. The PyTorch code used to train this model is depicted in Algorithm 1.

Algorithm 1 PyTorch code to convert OpenCLIP bigG to CLIP L.

```
class BigG_to_L(torch.nn.Module):
    def __init__(self):
        super(BigG_to_L, self).__init__()
        self.linear1 = nn.Linear(clip_seq_dim,
                                 clip_text_seq_dim)
        self.linear2 = nn.Linear(clip_emb_dim,
                                 clip_text_emb_dim)
    def forward(self, x):
        x = self.linear1(x)
        x = self.linear2(x.permute(0,2,1))
        return x
```

A.8 COCO Retrieval

MindEye1 scaled up image retrieval using a pool of billions of image candidates contained in the LAION-5B dataset (Schuhmann et al., 2022). This was possible because all LAION images were already converted to CLIP L embeddings and made available for nearest neighbor lookup via the CLIP Retrieval client (Beaumont, 2022). We were not able to use this approach for MindEye2 because it would require converting all images to the 256×1664 dimensionality bigG latent space which was not feasible. That said, cursory investigation with the comparatively smaller MS-COCO dataset suggests that retrieval from a pool of images not containing the original image may not work as well with OpenCLIP bigG embeddings compared to the CLIP L embeddings used in MindEye1. To test retrieval, we used FAISS (Douze et al., 2024) for k-nearest neighbor search through an index of flattened OpenCLIP bigG embeddings of 73,000 MS-COCO images. We found that for incorrect retrievals, the 3 nearest neighbors usually were dissimilar to the original image both semantically and in low-level appearance. This could be due to the latents corresponding to the 256 image patch tokens of OpenCLIP bigG representing a more complex combination of different levels of information. This could cause the OpenCLIP bigG embeddings to not be as effective for nearest neighbor retrieval in terms of subjective intepretation, as the last layer of CLIP ViT-L/14 is highly semantic but lacks in low-level image content. Although we demonstrated improved retrieval performance for MindEye2 compared to MindEye1 using random subsets of 300 images for MindEye2 compared to MindEye1 (Table 4), we suggest that mapping to the last layer of CLIP ViT-L/14 image space would work better if the intended application is to find semantically related nearest neighbors in a large image pool.

A.9 Reconstruction Evaluations: Additional Information

Two-way comparisons were performed for AlexNet (Krizhevsky et al., 2012) (second and fifth layers), InceptionV3 (Szegedy et al., 2016) (last pooling layer), and CLIP (final layer of ViT-L/14). We followed the same image preprocessing and the same two-way identification steps as Ozcelik and VanRullen (2023) and Scotti et al. (2023). For two-way identification, for each model, we computed the Pearson correlation between embeddings for the ground truth image and the reconstructed image, as well as the correlation between the ground truth image and a different reconstruction elsewhere in the test set. If the correlation for the former was higher than the latter, this was marked as correct. For each test sample, performance was averaged across all possible pairwise comparisons using the other 999 reconstructions to ensure no bias from random sample selection. This yielded 1,000 averaged percent correct outputs, which we averaged across to obtain the metrics reported in Table 1.

A.10 UMAP Dimensionality Reduction

As discussed in Scotti et al. (2023), UMAP dimensionality reduction (McInnes et al., 2020) plots of disjointed CLIP fMRI embeddings next to aligned CLIP fMRI embeddings visualize how the diffusion prior effectively addresses the disjointed embedding spaces problem. Theoretically, multimodal contrastive learning will always produce disjointed embeddings because of the "modality gap" phenomenon whereby encoding modalities into a shared space restricts the effective embedding space to a narrow cone in geometric space (Liang et al., 2022).

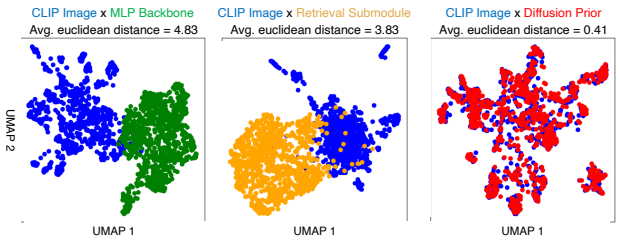

Figure 10: UMAP plots depict CLIP image latents (blue), backbone latents (green), retrieval submodule latents (orange), and diffusion prior latents (red). UMAPs were estimated across the 1,000 test samples for subject 1, using the full 40-session model. CLIP image latents correspond to the 256×1664 dimensionality of OpenCLIP ViT-bigG/14 image token embeddings. Euclidean distance between the given MindEye2 embedding space and CLIP image space is lowest for the diffusion prior, suggesting that the diffusion prior helps to align the two embedding spaces.

A.11 Pretraining with Less Subjects

To determine the relative impact of using additional subjects for pretraining, we separately fine-tuned a MindEye2 model for subject 1 (using 1 hour of their training data) that was pretrained only on subjects 2, 5, and 7 (these are the subjects who completed all 40 scanning sessions), as well as only on subject 5 (the subject whose single-subject model performed the best). Results in Table 10 show similar performance for these models compared to pretraining on the full set of available subjects, suggesting that the number of pretraining subjects does not seem to play a major role in subsequent fine-tuning performance.

Metric		Sub 2-8	Sub 2,5,7	Sub 5
Low-Level	PixCorr ↑	**0.235**	0.234	0.232
	SSIM ↑	**0.428**	0.421	0.421
	Alex(2) ↑	88.0%	**89.1%**	88.9%
	Alex(5) ↑	93.3%	**94.1%**	93.6%
High-Level	Incep ↑	83.6%	**83.8%**	**83.8%**
	CLIP ↑	80.8%	**83.5%**	82.7%
	Eff ↓	0.798	0.790	**0.787**
	SwAV ↓	0.459	0.448	**0.447**
Retrieval	Fwd ↑	**94.0%**	92.7%	90.6%
	Bwd ↑	77.6%	**80.8%**	80.3%
Brain Corr	Visual cortex ↑	0.347	**0.353**	0.352
	V1 ↑	**0.318**	0.316	**0.318**
	V2 ↑	**0.337**	0.331	0.336
	V3 ↑	0.341	0.339	**0.342**
	V4 ↑	0.316	**0.319**	0.318
	Higher vis. ↑	0.345	**0.355**	0.354

Table 10: Evaluation metrics for MindEye2 models both fine-tuned on 1 hour of training data from subject 1 but pretrained on different numbers of subjects.

A.12 Enhanced Reconstructions using Empty Text Prompts

To investigate whether our generated text captions from brain activity were improving subsequent image generation, we conducted an additional ablation. We remade reconstructions for subject 1 using their fine-tuned MindEye2 model but used empty text prompts instead of the image captions predicted from brain activity. We observed modest improvements in high-level metrics due to the use of predicted image captions, as shown in Table 11.

Metrics	Brain captions	Empty prompt
PixCorr ↑	**0.235**	**0.235**
SSIM ↑	**0.428**	0.423
Alex(2) ↑	**88.02%**	87.46%
Alex(5) ↑	**93.33%**	93.22%
Incep ↑	**83.56%**	82.67%
CLIP ↑	80.75%	**81.21%**
Eff ↓	**0.798**	0.810
SwAV ↓	**0.459**	0.461

Table 11: Using image captions predicted from brain activity for refinement seems to modestly improve subsequent image reconstructions from fMRI activity compared to using empty text prompts.

A.13 Subject-specific semantically labeled UMAPs

Appendix Figure 11 shows subject-specific UMAP plots with labeled semantic clusters to visualize what categories of images were better or worse reconstructed for each subject. If a plot does not show dense clustering of images within the same semantic category, this indicates that reconstructions from that subject for this category were likely not of good quality. Broad semantic category labels were manually labeled and taken from a Notion report by Shirakawa (2023), which details his explorations of the Natural Scenes Dataset images.

A.14 ROI-Optimized Stimuli

Here we try to visualize the functional organization of the brain by feeding synthetic brain activity through pretrained MindEye2. Inspired by the ROI-optimal analyses of Ozcelik and VanRullen (2023) and the synthetic brain maximization approach of BrainDIVE (Luo et al., 2023a), we utilized four ROIs derived from population receptive field (pRF) experiments and four ROIs derived from functional localization (fLoc) experiments. These pRF and fLoc experiments were provided by the NSD dataset. The ROIs are as follows (region names following the terminology adopted in Allen et al. (2021)): V1 is the concatenation of V1 ventral (V1v) and V1 dorsal (V1d), and similarly for V2 and V3; V4 is the human V4 (hV4); the Face-ROI consists of the union of OFA, FFA-1, FFA-2, mTL-faces, and aTL-faces; the Word-ROI consists of OWFA, VWFA-1, VWFA-2, mfs-words, and mTL-words; the Place-ROI consists of OPA, PPA, and RSC; and the Body-ROI consists of EBA, FBA-1, FBA-2, and mTL-bodies.

To observe the functional specialization associated with each of the ROIs, we used MindEye2 to reconstruct images based on synthetic fMRI patterns where flattened voxels were either set to 0 if outside the ROI or 1 if inside the ROI. Results are shown in Figure 12.

Subjectively interpreting these reconstructions, it seems that Face-ROI reconstructions depicted human faces, aligned with our expectations for the functional specialization of this region. Word-ROI reconstructions depicted distorted characters written on signs (with the exception of subject 7). Place-ROI reconstructions depicted enclosed environments, mostly rooms. Body-ROI reconstructions depicted strange mixtures of human body parts and animals. V1 reconstructions were dark with a few points of high contrast. V2 reconstructions showed somewhat softer colors. V3 and V4 reconstructions were more abstract with amorphous shapes

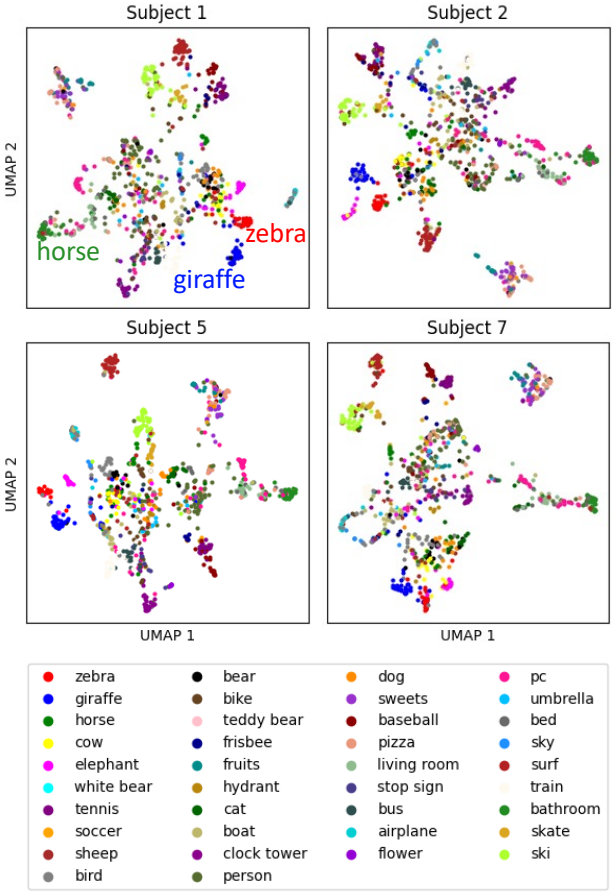

Figure 11: Subject-specific UMAPs on the diffusion prior outputs from the 40-hour MindEye2 models help to visualize individual differences with regard to reconstuction quality for different categories of natural images.

Figure 12: Unrefined reconstructions and decoded captions from synthetic fMRI activity. Voxels in the desired target brain region were set to a one while all other voxels were set to zero, and this synthetic brain data was fed through each subject's MindEye2 model.

and more vivid colors.

Such results demonstrate the potential to directly visualize preferential stimuli for any desired region of interest; further functional specialization exploration could be performed using more sophisticated methods (c.f., Sarch et al. (2023); Luo et al. (2023a;b); Tang et al. (2022); Huth et al. (2016); Sucholutsky et al. (2023)).

A.15 Human Preference Experiments

We conducted two-alternative forced-choice experiments on 58 human raters online. We probed three comparisons intermixed into the same behavioral experiment, with each comparison consisting of 1200 trials sampled evenly from the 1000 NSD test samples across the 4 subjects who completed all 40 scanning sessions (subjects 1, 2, 5, 7). The total 3600 experimental trials were shuffled and 87 trials were presented to each subject. Our subjects were recruited through the Prolific platform, with our experimental tasks hosted on Meadows. All other experiment details follow the protocol used in Kneeland et al. (2023c).

A.15.1 MindEye2 vs. MindEye2 (random)

The first comparison was a two-way identification task in which participants were asked to select which of two images was more similar to a ground truth image. The two images provided for comparison were both reconstructions from MindEye2 (40-hour), one being a randomly selected reconstruction from the test set and the other being the correct, corresponding reconstruction. Raters correctly identified the corresponding reconstruction 97.82% of the time ($p < 0.001$). This establishes a new SOTA for human-rated image identification accuracy, as the only other papers to perform such an experiment were the original method proposed by (Takagi and Nishimoto, 2022), whose method achieved 84.29%, and the MindEye1 + BOI (brain-optimized inference) method proposed by (Kneeland et al., 2023c), whose enhancement to the MindEye1 method achieved 95.62%. The method in (Takagi and Nishimoto, 2022) is different from the "+Decoded Text" method we compare against in Table 1, which was released in a later technical report (Takagi and Nishimoto, 2023), and which does not report human subjective evaluations.

A.15.2 MindEye2 (refined) vs. MindEye2 (unrefined)

The second comparison was the same task as the first but this time comparing refined MindEye2 (40-hour) reconstructions against unrefined MindEye2 reconstructions (both correctly corresponding to the appropriate fMRI activity). This comparison was designed to empirically confirm the subjective improvements in naturalistic quality provided by MindEye2's refinement step. This is particularly important to confirm because the quantitative evaluation metrics displayed in Table 4 sometimes preferred the unrefined reconstructions. Refined reconstructions were rated as more similar to the ground truth images 71.94% of the time ($p < 0.001$), demonstrating that the final refinement step improves reconstruction quality and accuracy when assessed by humans.

A.15.3 MindEye2 (1-hour) vs. Brain Diffuser (1-hour)

The final comparison was likewise the same task but this time comparing reconstructions from MindEye2 against reconstructions from the Brain Diffuser method (Ozcelik and VanRullen, 2023), where both methods were trained using only the first hour of scanning data from the 4 NSD subjects. This experiment demonstrated that the MindEye2 reconstructions were preferred 53.01% of the time ($p = 0.044$), demonstrating a statistically significant improvement in scaling performance compared to the previous state-of-the-art model for reconstructions using only 1 hour of training data. This confirms the results in the main text that MindEye2 achieves SOTA in the low-sample 1-hour setting. We visualized cases where Brain Diffuser (1-hour) was preferred over MindEye2 (1-hour) in Appendix Figure 13. We observed that Brain Diffuser reconstructions were often preferred in situations where both MindEye2 and Brain Diffuser reconstructions were low quality, but MindEye2 reconstructions were "confidently" wrong (in the sense that MindEye2 reconstructions enforce a naturalistic prior from the refinement step) whereas Brain Diffuser reconstructions were producing distorted outputs that contained subtle elements corresponding to the target image. This may indicate human raters prefer distorted outputs with recognizable features, and disfavored the model that enforces a naturalistic prior, and may lose these features.

Figure 13: Examples of trials where subjects preferred Brain Diffuser (1-hour) reconstructions over MindEye2 (1-hour) reconstructions. It seems possible human raters tended to select the more distorted reconstruction of the two when both reconstructions were of bad quality, but the distortions trend towards correct features.