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

Paul S. Scotti[1,2], Mihir Tripathy[†,2], Cesar Torrico[†,2], Reese Kneeland[†,3], Tong Chen[4,2], Ashutosh Narang[2], Charan Santhirasegaran[2], Jonathan Xu[5,2], Thomas Naselaris[3], Kenneth A. Norman[6], Tanishq Mathew Abraham[1,2]
[1]Stability AI, [2]Medical AI Research Center (MedARC), [3]University of Minnesota, [4]The University of Sydney, [5]University of Waterloo, [6]Princeton Neuroscience Institute

*Our **MedARC Neuroimaging & AI Lab** is now working on real-time reconstructions and foundation neuroimaging models. Join our lab as a volunteer contributor: https://medarc.ai/fmri*

Reconstructions of seen images from human brain activity using ONE hour of fMRI training data (previous work used FORTY hours)

Seen image | 10 min. | 30 min. | 1 hour | 2 hours | 10 hours | 40 hours | Seen image | 10 min. | 30 min. | 1 hour | 2 hours | 10 hours | 40 hours

Background

Functional magnetic resonance imaging (fMRI) measures neural activation as changes in blood oxygenation. Decoding seen images from fMRI enables better understanding of brain function and potential for mind-reading applications in brain-computer interfaces. fMRI is expensive and time-consuming so generalization with sparse training data is essential for practical adoption. We used the *Natural Scenes Dataset* (NSD) [1], a public fMRI dataset containing brain responses of human participants looking at naturalistic photographs (MS-COCO).

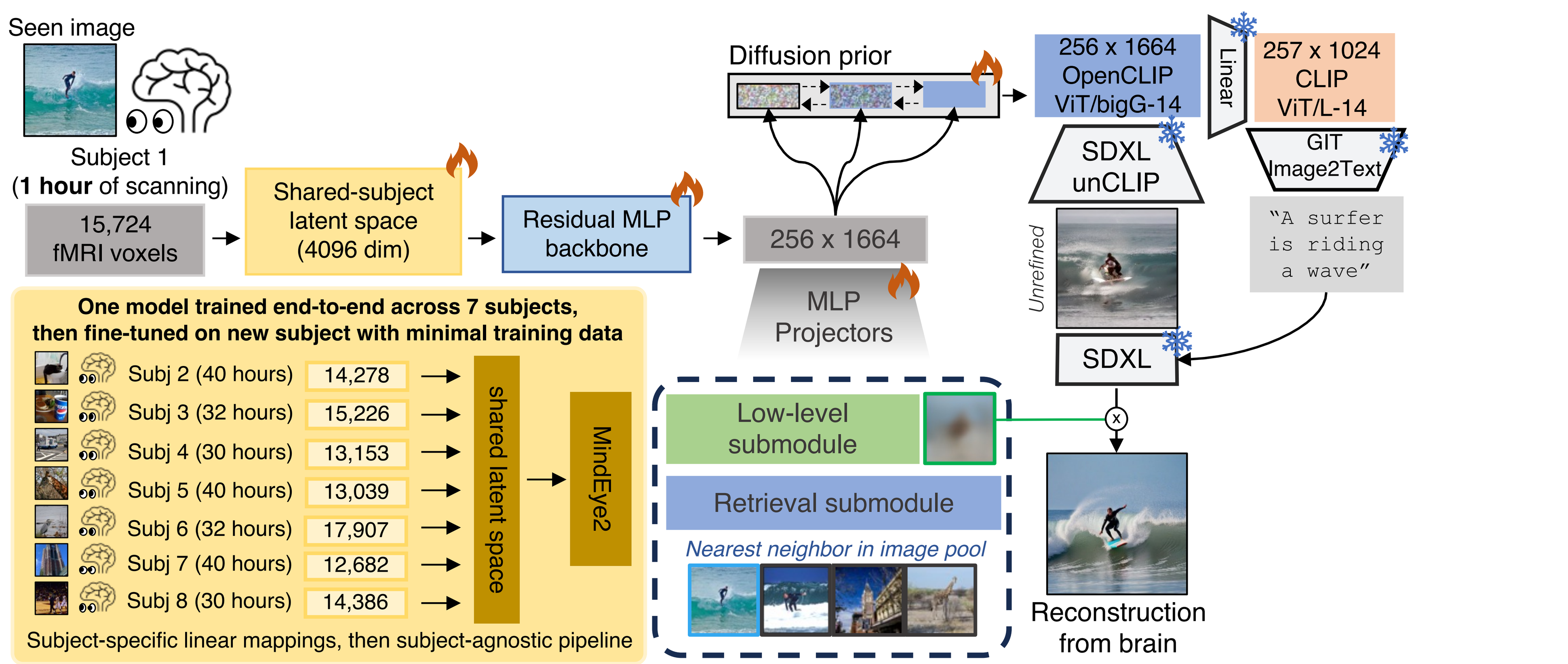
MindEye2 achieves state-of-the-art performance across *retrieval* and *reconstruction*.

Retrieval: identify the original (or most similar) image out of a pool of candidates (i.e., nearest neighbor)
Reconstruction: recreate the original seen image (i.e., output from latent diffusion model)

Methods

Compared to past work, MindEye2 innovates by:

1. Training model *across* subjects
2. Mapping to stronger CLIP space (OpenCLIP bigG)
3. Fine-tuning a SOTA Stable Diffusion XL [3] unCLIP model
4. Predict image captions from brain for added guidance



Each of 10,000 unique images was viewed 3x for 3 sec. Corresponding fMRI voxels (1.8mm cubes of cortex) were collected for each image presentation. We pretrain our model across 7 subjects and fine-tune on minimal data from a new subject. We linearly map all brain data to a shared-subject latent space, followed by a shared non-linear mapping to OpenCLIP [2] 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.

References. [1] Allen et al. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuro.* [2] Ilharco et al. (2021). OpenCLIP. [3] Podell et al. (2023). Sdxl: Improving latent diffusion models for high-resolution image synthesis. *ICLR*. [4] Meng et al. (2022). SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. *ICLR*. [5] Reddy et al. (2010). Reading the mind's eye: Decoding category information during mental imagery. *NeuroImage*. [6] Wallace et al. (2022). RTCloud: A cloud-based software framework to simplify and standardize real-time fMRI. *NeuroImage*. [7] Scotti et al. (2023). Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. *NeurIPS*.

Qualitative comparison to past work

Seen image | Reconstructions using 1 hour of training data

MindEye2 | MindEye2 (not pretrained across subjects) | MindEye1 (Scotti et al., 2023) [7] | Brain Diffuser (Ozcelik et al., 2023) | Takagi et al. (2022)

unCLIP comparison

unCLIP models can convert CLIP image embeddings back to pixel space.

We fine-tuned SDXL to support CLIP image embedding input instead of text, raising ceiling reconstruction performance.

Reconstructions from ground truth CLIP image embeddings

Original image | Versatile Diffusion (CLIP ViT-L/14) | SDXL unCLIP (OpenCLIP ViT-bigG/14)

Refinement with image caption prediction

Unrefined SDXL unCLIP recon + predicted caption | **Refined** reconstruction

"Unrefined" reconstructions = pixel images output directly from SDXL unCLIP

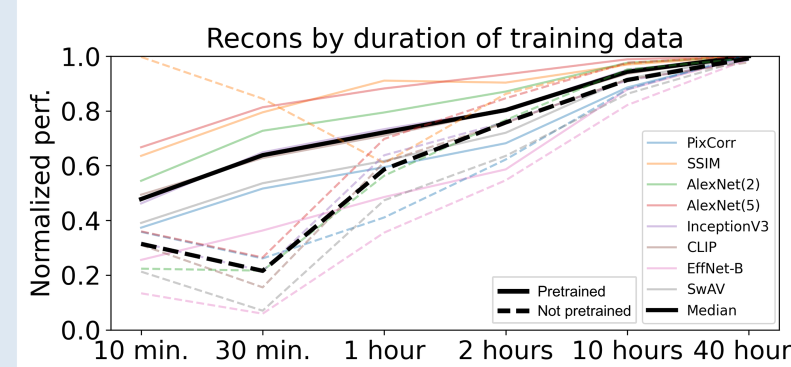
We observed unrefined reconstructions were SOTA but subjectively distorted. To improve image realism, we use image-to-image [4] with base SDXL, feeding unrefined recons alongside a MindEye2 predicted image caption.

Quantitative comparison to past work

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%

Results are from full 40-hours training data, averaged across the same 4 participants. PixCorr=pixelwise correlation between ground truth and reconstructions; SSIM=structural similarity index metric; EfficientNet-B1 and SwAV-ResNet50 refer to average correlation distance; all other metrics refer to two-way identification (chance = 50%). 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. **Bold**=best performance, underline= 2nd best.

Varying amt. of training data



The 1-hour setting offers a good balance between scan duration and reconstruction performance, with notable improvements from pretraining.

Ablations

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%

Ablations show importance of both shared-subject modeling and leveraging improved CLIP image space.

ME1 = MindEye1 MLP instead of shared-subject linear mapping
CLIP L = Mapping to CLIP-L instead of OpenCLIP bigG

Conclusions: Benefits & Risks/Limitations

- Potential for new clinical diagnostic methods: reconstructions are expected to be systematically distorted due to mental state.
- Potential to generalize to mental imagery: similar patterns of brain activity are observed across perception and mental imagery [5].
- Real-time brain-computer interfaces [6] e.g., communication with patients in a pseudocoma.
- 1-hour generalization enables practical adoption.
- MindEye2 is limited to natural scene image distributions.
- Data easily becomes too noisy with slight movement or inattention to the task.
- Privacy: IRB approval and participant consent for data sharing was obtained. Medical data should be carefully protected and transparently used.