

Appendix

A DATASETS

A.1 Natural Scenes Dataset

The Natural Scenes Dataset (NSD) [1] is a large-scale public fMRI dataset conducted at ultra-high-field (7 Tesla). The dataset consists of whole-brain, high-resolution (1.8-mm isotropic, 1.6-s sampling rate) fMRI measurements of 8 healthy adult subjects while they viewed thousands of color natural scenes over the course of 30–40 scan sessions. The visual stimuli comes from 73,000 images in Microsoft Common Objects in Context (MS-COCO) dataset [4]. Each unique image was displayed three times in 30 to 40 scanning sessions for three seconds each time, and each participant participated in a total of 22,000 to 30,000 trials. For four participants who completed all scanning sessions, it leads to 24,980 training samples and 2,770 test samples. The fMRI responses are preprocessed into session-wise z-scored single-trial betas output from *GLMSingle*, which corrected for differences in slice time acquisition, head motion, and spatial distortion.

We download NSD from official website ¹. Follow the setting of previous research [5, 8], for fMRI data, we use *nsdgeneral* mask to select ROI. This region is a general ROI that was manually drawn covering voxels responsive to the NSD experiment in the posterior aspect of cortex. More details about the dataset can be found in Tab 1. Besides, we averaged fMRI responses across same-image repetitions, which leads to 8,859 training samples and 982 test samples. Each visual stimuli presented to the subject have a uniform size of 425×425 . We use CLIP ViT-L/14 to extract image features and text features. The text features are averaged across multiple captions.

A.2 LAION-2B-en

LAION-2B-en is a large image-text paired dataset released by Large-scale Artificial Intelligence Open Network (LAION). It contains approximately 2.32 billion image text pairs with English captions. LAION-2B-en has been widely used for training various large-scale vision-language models, such as CLIP [6], DALL-E [7], etc. We use *img2dataset* and *clip-retrieval* tools [2, 3] to calculate LAION-2B-en embeddings and indices. All images were resized to 256×256 .

B UPPER-BOUND AND CROSS-SUBJECT PERFORMANCE

To further demonstrate the performance of BrainRAM, we show the performance of models train on different subjects. Also, we demonstrated the upper-bound performance, that is, directly using image and text embeddings extracted from CLIP to reconstruct image in Versatile Diffusion dual-guided pipeline. The reconstruction performance and retrieval performance are shown in Table 2 and Table 3, correspondingly. It can be found that BrainRAM has good generalization ability across different subjects.

Subject	Voxels	ROIs	Train	Test
subj01	15724	V1, V2, V3, hV4, VO, PHC, MT, MST, LO, IPS	8859	982
subj02	14278			
subj05	13039			
subj07	12682			

Table 1: Details of the Natural Scenes Dataset.

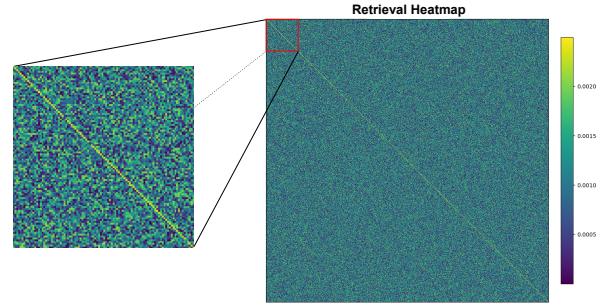


Figure 1: Retrieval heatmap on test set of NSD for subject 1.

C DIFFERENT SELECTION OF NUMBER FOR RETRIEVAL

During the training of Retrieval-Augmentation Module (RAM), the number of k -nearest neighbors has strong influence on the performance of the model. As a result, we use four different $k \in \{1, 2, 4, 8\}$ to test the effect of different k values on model performance. The results are presented in Table 4. It can be found that before $k = 4$, the performance of the model has a visible improvement with the growth of k . However, when $k = 8$, the performance improvement of the model is not significant, and the number parameters are also increasing. Therefore, we choose $k = 4$ in RAM for reconstruction and LAION/testset retrieval.

D VISUALIZATION

We illustrated more reconstruction results from different subjects in Fig 2. These results are consistent with the quantitative indicators in the Table 2, indicating that our BrainRAM has good generalization ability across different subjects. We also visualized retrieval heatmap of subject 1 on NSD test set retrieval in Fig 1. The diagonal with shallower color have the highest similarity. This shows that although there are many similar images in the test set, BrainRAM still accurately captures the fine-grained semantics in different images and distinguishes them.

REFERENCES

- [1] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. 2022. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience* 25, 1 (2022), 116–126.
- [2] Romain Beaumont. 2021. *img2dataset*: Easily turn large sets of image urls to an image dataset. <https://github.com/rom1504/img2dataset>.

¹<http://naturalscenesdataset.org/>

Method	Low-Level				High-Level			
	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
Upper-bound	.215	.387	95.1%	98.9%	99.4%	99.9%	.458	.243
Ours (subj01)	.176	.342	89.9%	95.7%	92.6%	94.1%	.666	.381
Ours (subj02)	.153	.328	89.1%	93.8%	93.1%	93.5%	.692	.393
Ours (subj05)	.162	.332	88.4%	94.3%	92.8%	94.3%	.683	.406
Ours (subj07)	.155	.345	89.3%	95.1%	91.9%	92.9%	.711	.385

Table 2: The upper-bound reconstruction performance and cross-subject reconstruction performance.

Method	High-Level				Image↑	Brain↑
	Incep↑	CLIP↑	Eff↓	SwAV↓		
Upper-bound	95.6%	97.3%	.493	.258	-	-
Ours (subj01)	88.1%	91.0%	.726	.444	93.4%	90.3%
Ours (subj02)	89.6%	90.3%	.752	.460	91.8%	89.4%
Ours (subj05)	87.3%	89.8%	.819	.483	86.4%	88.2%
Ours (subj07)	90.2%	89.6%	.795	.481	83.1%	85.7%

Table 3: The upper-bound retrieval performance and cross-subject retrieval performance. The first four columns refer to high-level metrics computed on retrieved images from LAION-2B-en. The last two columns represent image retrieval and fMRI retrieval performance on the test set.

Number of retrieved samples	High-Level				Params↓
	Incep↑	CLIP↑	Eff↓	SwAV↓	
$k = 1$	88.7%	90.3%	.693	.432	4.85M
$k = 2$	92.3%	92.5%	.703	.393	4.89M
$k = 4$	92.6%	94.1%	.666	.381	4.98M
$k = 8$	93.1%	94.4%	.672	.389	5.15M

Table 4: Effect of different k .

- [3] Romain Beaumont. 2022. Clip Retrieval: Easily compute clip embeddings and build a clip retrieval system with them. <https://github.com/rom1504/clip-retrieval>.
- [4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common

objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer, 740–755.

- [5] Furkan Ozelik and Rufin VanRullen. 2023. Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports* 13, 1 (2023), 15666.
- [6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [7] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.
- [8] Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. 2023. Reconstructing the mind’s eye: fMRI-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems* 36 (2023).

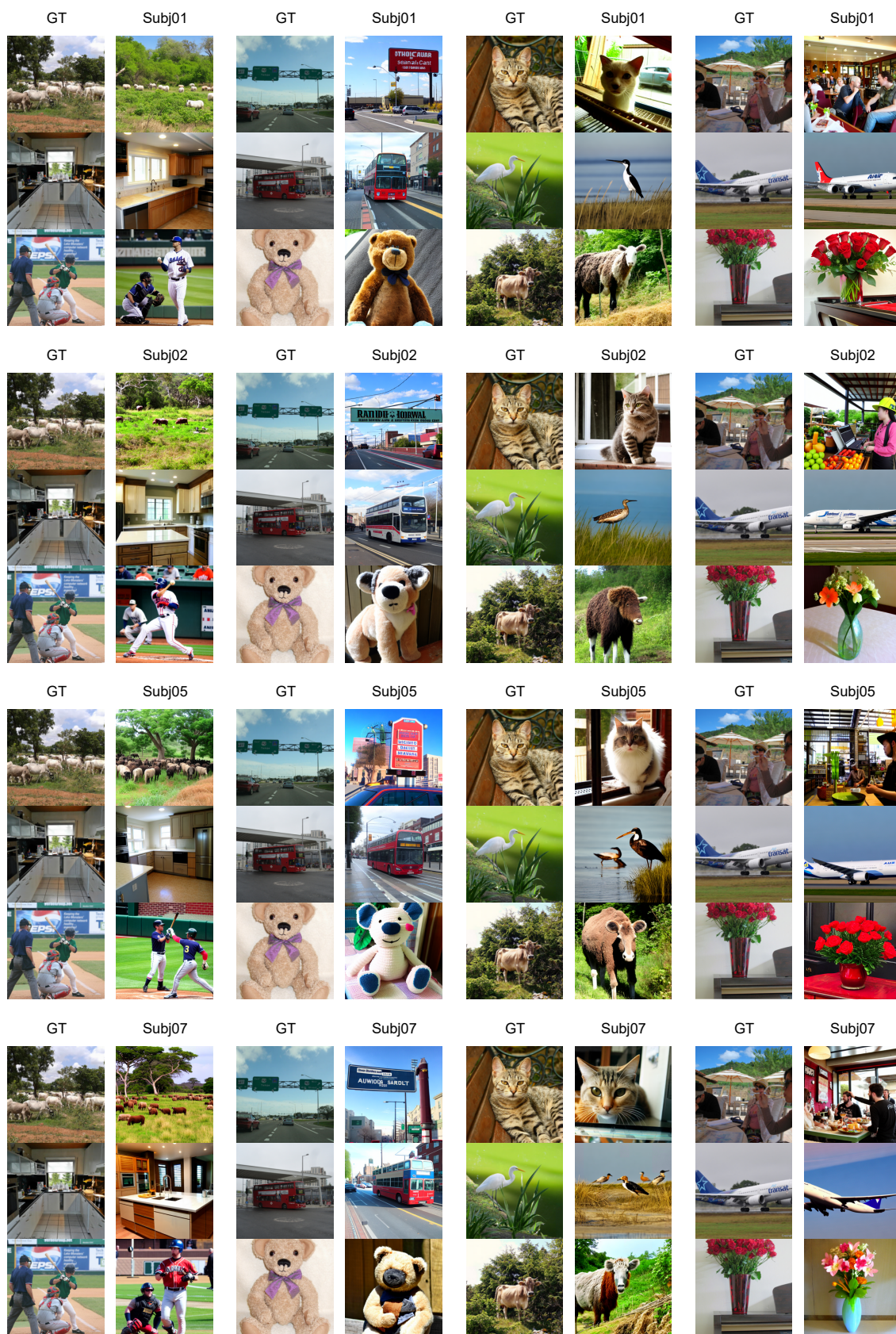


Figure 2: Visual comparison of reconstruct results from different subjects.