MindArt Supplementary Appendix

1. GM-Guided Neural Representation Visualization

To gain an intuitive understanding of the quality of GM-guided neural representations, we present visualization results of different embedding distributions, as shown in Fig. 1, using the *t*-SNE [3] method. In contrast to both the raw fMRI and CLIP visual embedding spaces [1], the representation distribution of the GM-guided fMRI encoder reveals a more pronounced locality-sensitive structure, which is distinguished by the closer positions of similar stimuli and greater distances between dissimilar stimuli within the two-dimensional plane.

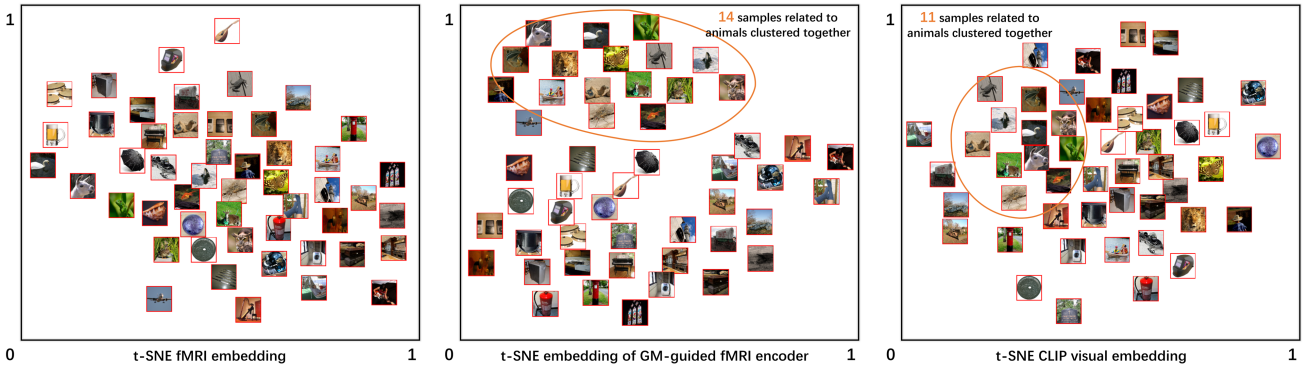

Figure 1. *t*-SNE visualization results for different representations. From left to right: raw fMRI recordings, neural representations learned from our GM-guided fMRI encoder, and pre-trained CLIP visual representations.

2. The Role of Retrieval Prompts

In our design, the retrieval augmentation is based on neural graph node representation $\hat{\mathcal{X}}_i$. However, our MindArt also works when using only a learnable ViT encoder (*i.e.*, *w/o* the GM-guided component). In the circumstances, we just need to utilize $\mathrm{F_i}$ to infer $\mathbf{X}_{z_i}$ via simple ridge regression, and act as node representation $\hat{\mathcal{X}}_i$. We provide an additional ablation study on MindArt-L model, as shown in Tab. 1, to evaluate the roles of the retrieval augmentation. The results indicated the GM-guided representation learning effectively boosts the model's decoding capacity. For another, the retrieval prompt enables model to access external contextual knowledge about decoding targets, thus enhancing text reconstruction capabilities of GPT-2. This means that the performance gains from retrieval prompt are closely related to the GM-guided representations.

Model	GM	Retrieval	SSIM ↑	$\mathrm{CLIP_{Score}}$ ↑	$\mathrm{CLIP_T}$@10 ↑	$\mathrm{CLIP_T}$@50 ↑
MindArt-L	✓	✓	$0.242_{\pm 0.13}$	$\mathbf{0.631}_{\pm 0.13}$	$\mathbf{43.6\%}_{\pm 3.9\%}$	**24.0%**
MindArt-L	✓	×	$\mathbf{0.249}_{\pm 0.14}$	$0.623_{\pm 0.14}$	$42.8\%_{\pm 3.8\%}$	20.0%
MindArt-L	×	✓	$0.198_{\pm 0.13}$	$0.591_{\pm 0.14}$	$32.0\%_{\pm 4.2\%}$	12.0%
MindArt-L	×	×	$0.199_{\pm 0.14}$	$0.589_{\pm 0.15}$	$31.8\%_{\pm 4.0\%}$	14.0%

Table 1. The impact of retrieval prompt on visual reconstruction. The best and worst values are highlighted in **Bold** and red, respectively.

What happens if the generated (retrieval) caption is highly inconsistent with the ground truth. Here, we leverage an example (See Fig. 2) to illustrate the point. From the results, we can observe that the reconstructed image still preserves a similar layout to the visual stimulus due to the constrained variable $\mathbf{X}_{z_i}$ (despite being semantically incorrect).

Figure 2. What happens if the generated caption is highly inconsistent with the visual stimulus.

3. More Reconstruction Samples

At last we offer the complete reconstruction results of our MindArt (*w/o* extra style) for all test samples from subject 3 in the DIR dataset [2], as illustrated in Fig. 3. Although our MindArt can achieve satisfactory reconstruction results in many cases, there are still some failure examples, which tend to generate semantically inaccurate targets that are similar in appearance. We think it was due to low quality in fMRI signals or the subject was not fully focused when viewing the image. Under this circumstance, it is difficult to capture the fine-grained semantic representations hidden in the fMRI pattern, which can be a limitation of this work. To overcome that, incorporating more semantic prior knowledge to guide neural representation learning is a topic worth exploring.

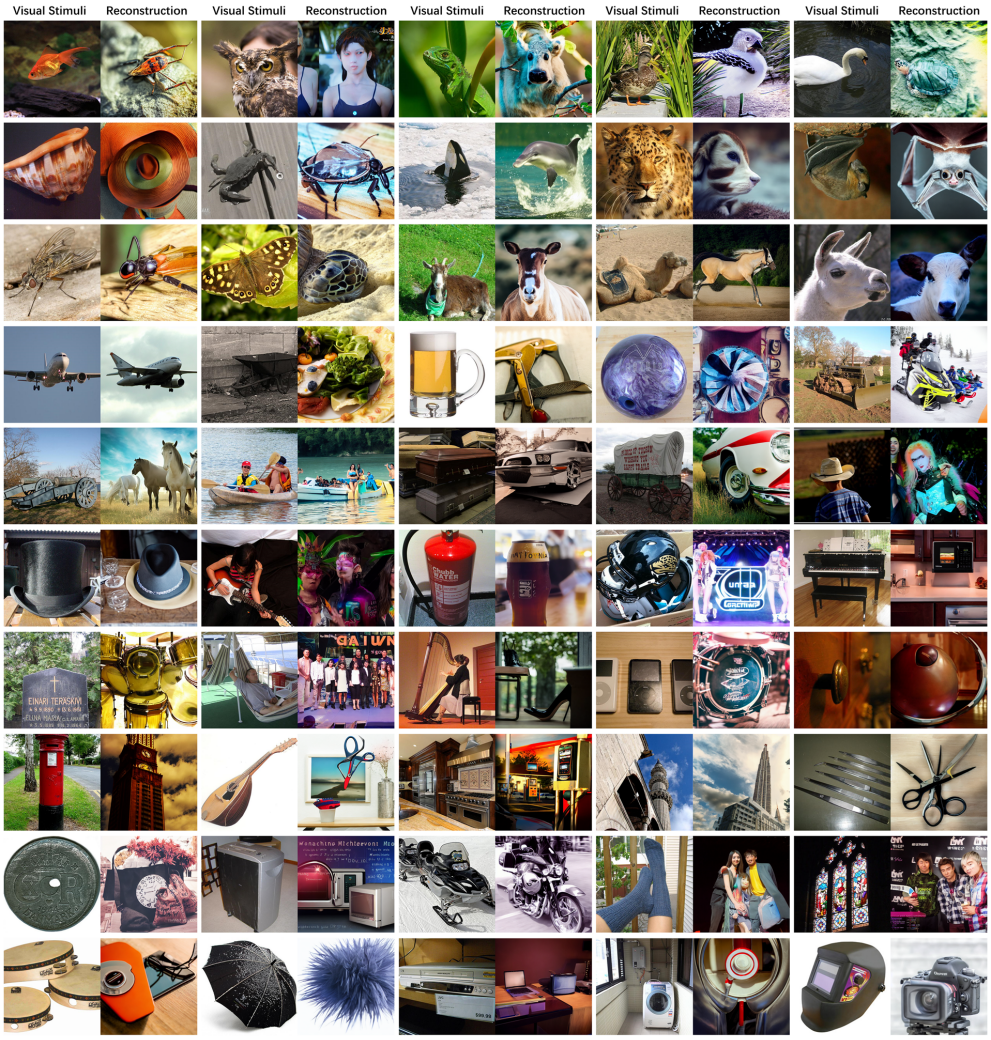

Figure 3. Full reconstruction samples for subject 3 in the DIR dataset. For each group, the left represents ground truth visual stimuli.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[2] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLOS Computational Biology*, 15(1):1–23, 2019. 2

[3] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008. 1