Cross-Subject Mind Decoding from Inaccurate Representations –Supplementary Materials–

Yangyang Xu[1]*, Bangzhen Liu[2,5], Wenqi Shao[3], Yong Du[4]*, Shengfeng He[5], Tingting Zhu[1]
[1]The University of Oxford [2]South China University of Technology [3]Shanghai AI Lab
[4]Ocean University of China [5]Singapore Management University

In this supplementary material, we provide analysis of shared latent space, neuroscience interpretability, the failure cases and limitation of our method, detailed information on the evaluation metrics, the structural components of our framework, additional qualitative comparisons with existing methods, qualitative analyses of various framework variants, analysis of VCM, qualitative evaluations under data-limited scenarios, and the synthesis of fMRI data for specific subjects.

1. Analysis of Shared Latent Space

Our framework learns a shared latent space that aligns fMRI and visual features across subjects, enabling generalization by capturing subject invariant patterns. We present the t-SNE visualization of subject-specific and cross-subject representations in Fig. 1, t-SNE visualizations reveal tighter clustering of cross-subject representations, indicating better alignment across different subjects.

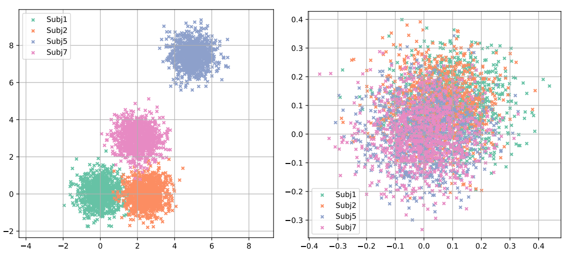
2. Neuroscience Interpretability

We further investigate the neuroscience interpretability of our model by analyzing voxel-level gradients derived from internal representations. As illustrated in Fig. 3, the results indicate that the Low-level Visual Cortex (LVC) predominantly supports edge decoding, while the High-level Visual Cortex (HVC) is more involved in semantic processing. Both regions contribute to color prediction. These findings suggest that our shared representational space captures and preserves the hierarchical structure of visual processing across subjects.

3. Failure cases

We present the failure cases of our method in Fig. 2, our method inherits the limitation of SD, which cannot handle the complex scenes [9, 22]. Additionally, it also fails when encountering unnatural colors.

*Corresponding authors: Yangyang Xu (xuyangyang@hit.edu.cn) and Yong Du (csyongdu@ouc.edu.cn).



(a) Subject-Specific Representations (b) Cross Subject Representations

Figure 1. t-SNE visualizations of representations learned by subject-specific and cross subject decoding.

4. Limitation

One limitation of our work is that the SBMM is a subject-dependent component, which needs to be retrained for every new subject. This is due to the significant inter-subject variability and limited large-scale datasets, resulting in the requirement of subject-specific components, which is a shared limitation of current cross-subject studies [12, 15, 19, 21]. Although not our purpose, our method could be misused for privacy invasion or other unethical purposes. Thus, strict and responsible data privacy protections must be established.

5. Details of Evaluation Metrics

We use 8 evaluation metrics for the quantitative comparison from low and high levels. **PixCorr** measures the pixel-wise correlation of decoded and GT images, **SSIM** measures the structure similarity between two images [20]. **AlexNet(2)** is the two-way comparison of image features extracted from the second layer of AlexNet [7], and **AlexNet(5)** compares the features extracted from the fifth layer. The above four metrics evaluate the low-level similarity of reconstructed images. The high-level metrics including **Inception**, **CLIP**, **EffNet-B**, and **SwAV**. **Inception** is the two-way comparison of the features extracted from the last pooling layer of InceptionV3 [16], CLIP compares the cosine similarity between the features extracted from the CLIP image encoder [13]. **EffNet-B** and **SwAV** are distance metrics based

Table 1. Qualitative comparisons with other methods on three datasets.

Method	NOD			GOD			BOLD5000		
	Acc (%)	PCC	SSIM	Acc (%)	PCC	SSIM	Acc (%)	PCC	SSIM
IC-GAN [11]	-	-	-	29.39	0.449	0.545	-	-	-
MinD-Vis [3]	-	-	-	26.64	0.532	0.527	25.918	0.545	0.524
CMVDM [23]	-	-	-	30.11	0.768	0.632	27.791	0.557	0.535
Ours	**35.12**	**0.734**	**0.745**	**34.311**	**0.794**	0.704	**29.088**	**0.583**	**0.553**

(a) Prediction (b) Stimulus

Figure 2. Failure cases of our method.

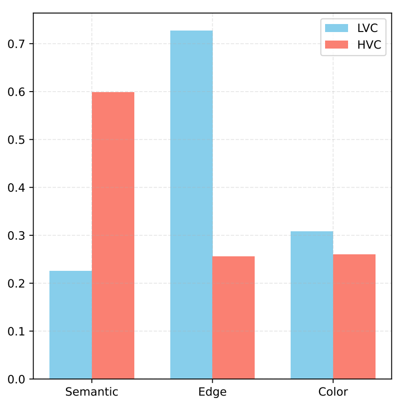


Figure 3. Voxel-level gradient analysis of visual features across brain regions.

on EfficientNet-B1 [18] and SwAV-ResNet50 [1], respectively.

6. Structure Details

Our SRM adopts the Querying Transformer [8] architecture, comprising 12 hidden layers. As illustrated in the middle-right panel of Fig. 2 in the main paper, the predicted semantic representation $\tilde{S}$ is incorporated into the cross-attention layers of the Querying Transformer.

The VCM consists of 13 layers, each designed for different resolutions. Each layer includes three `Conv2D` layers with `SiLU` activation functions between them, and the final output is activated using a `Sigmoid` function.

The pseudo-code for the BAI framework is provided in Alg. 1.

7. Evaluation on More Datasets

We also extend our framework on other mind decoding benchmark, including NOD [4], GOD [5], and BOLD5000 [2] datasets. We evaluate our method on three datasets using the same metrics as CMVDM [23]. As shown in Tab. 1, our method consistently outperforms existing baselines across all datasets and metrics.

8. More Qualitative Comparisons with Competitors

Additional qualitative comparisons with competing methods are presented in Fig. 4. Our method demonstrates higher consistency with the GT stimulus images in terms of semantics, structure, and color.

9. Qualitative Comparison of Variants

We present a qualitative comparison of various model variants, including predicted edges and color representations, in Fig. 5. The variants UM, w/o SBMM, and Ours-SS fail to predict the reasonable edges and color representations, whereas our model successfully predicts the rough edges and color of the image. Directly using the predicted representations does not yield plausible images, as the second-stage mind decoding requires accurate representations. Our proposed modules, SRM and VCM, enhance the quality of the reconstructed image by tolerating rough representations. Finally, our complete framework produces faithfully reconstructed results with plausible appearances.

10. Analysis of VCM

As shown in Fig. 5, the predicted edges and color palettes are dissimilar to the GT edges and colors, why the final reconstruction be faithful with the GT stimulus? Here we answer

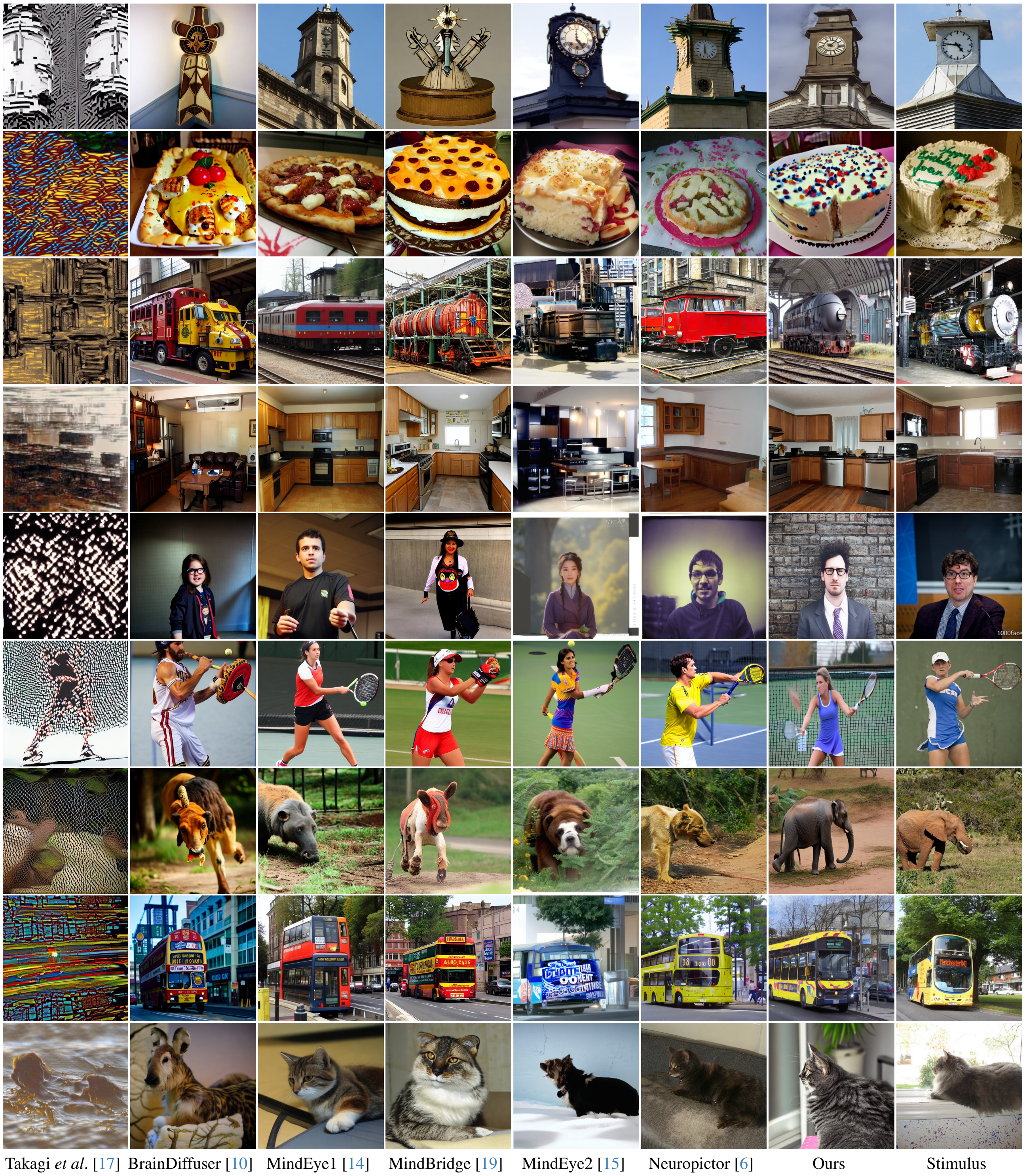


Figure 4. More qualitative comparison with competitors on mind decoding.

this question by visualizing the output of VCM. The visualization of VCM's output weights α_e and α_c are shown in Fig. 6 (the brighter indicates a higher value). The predicted α_e and α_c control fusion weights to relax the influence of predicted edge and color conditions to output, though predicted representations are inaccurate, the dissimilar representations can also lead to faithful mind decoding.

11. Qualitative Comparison under Different Data Limitation Scenarios

We present a qualitative comparison under different data limitation scenarios in Fig. 7. As the number of training samples increases, the reconstruction quality also improves. Compared to training from sketches with limited data, our adapted method reconstructs the image more faithfully.

12. Synthesis fMRI for Specific Subject

Given an unseen stimulus image, our framework mimics the visual system by synthesizing the corresponding fMRI for a specific subject: $\{S, E, C\} \Rightarrow \hat{V}_x$. We then decode the synthesized fMRI voxels into representations: $\hat{V}_x \Rightarrow \{\hat{\hat{S}}, \hat{\hat{E}}, \hat{\hat{C}}\}$. The reconstructed images are shown in Fig. 8, where the synthesized fMRI faithfully reconstructs the stimulus image.

13. Comparison with MindBridge on Cross Subject Mind Decoding

We present a cross-subject comparison with MindBridge and MindEye2 in Fig. 9. Our decoded images exhibit greater consistency with the stimulus image across different subjects. For instance, both MindBridge and MindEye2 fail to decode the *"Broccoli"* in the second sample, whereas our method successfully reconstructs it.

References

[1] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS*, 33:9912–9924, 2020. 2

[2] Nadine Chang, John A Pyles, Austin Marcus, Abhinav Gupta, Michael J Tarr, and Elissa M Aminoff. Bold5000, a public fmri dataset while viewing 5000 visual images. *Scientific data*, 6(1):49, 2019. 2

[3] 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. In *CVPR*, pages 22710–22720, 2023. 2

[4] Zhengxin Gong, Ming Zhou, Yuxuan Dai, Yushan Wen, Youyi Liu, and Zonglei Zhen. A large-scale fmri dataset for the visual processing of naturalistic scenes. *Scientific Data*, 10 (1):559, 2023. 2

[5] Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):15037, 2017. 2

[6] Jingyang Huo, Yikai Wang, Yun Wang, Xuelin Qian, Chong Li, Yanwei Fu, and Jianfeng Feng. Neuropictor: Refining fmri-to-image reconstruction via multi-individual pretraining and multi-level modulation. In *ECCV*, pages 56–73, 2024. 3

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 25, 2012. 1

[8] 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 *ICML*, pages 19730–19742, 2023. 2

[9] Minghao Liu, Le Zhang, Yingjie Tian, Xiaochao Qu, Luoqi Liu, and Ting Liu. Draw like an artist: Complex scene generation with diffusion model via composition, painting, and retouching. *arXiv preprint arXiv:2408.13858*, 2024. 1

[10] Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023. 3

[11] 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. In *IJCNN*, pages 1–8, 2022. 2

[12] Ruijie Quan, Wenguan Wang, Zhibo Tian, Fan Ma, and Yi Yang. Psychometry: An omnifit model for image reconstruction from human brain activity. In *CVPR*, pages 233–243, 2024. 1

[13] 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 *ICLR*, pages 8748–8763, 2021. 1

[14] Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. In *NeurIPS*, 2023. 3

[15] Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. In *ICML*, 2024. 1, 3

[16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 1

[17] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *CVPR*, pages 14453–14463, 2023. 3

[18] Mingxing Tan. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114, 2019. 2

[19] Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Mindbridge: A cross-subject brain decoding framework. In *CVPR*, pages 11333–11342, 2024. 1, 3

[20] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 1

[21] Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. Umbrae: Unified multimodal brain decoding. In *ECCV*, pages 242–259, 2024. 1

[22] Binbin Yang, Yi Luo, Ziliang Chen, Guangrun Wang, Xiaodan Liang, and Liang Lin. Law-diffusion: Complex scene generation by diffusion with layouts. In *ICCV*, pages 22669–22679, 2023. 1

[23] Bohan Zeng, Shanglin Li, Xuhui Liu, Sicheng Gao, Xiaolong Jiang, Xu Tang, Yao Hu, Jianzhuang Liu, and Baochang Zhang. Controllable mind visual diffusion model. In *AAAI*, pages 6935–6943, 2024. 2

Algorithm 1: Structure details of BAI.

```
class BAI:
   # Shared Encoder
   shared_encoder = Sequential([
      Linear(8192, 1024),
      LayerNorm((1024,))
   ])

   # SBMM in Encoders
   encoder_alpha_subj = ModuleDict({
      subj_id: Sequential([
         Linear(1024, 1024)
      ]) for subj_id in [1, 2, 5, 7]
   })

   encoder_beta_subj = ModuleDict({
      subj_id: Sequential([
         Linear(1024, 1024)
      ]) for subj_id in [1, 2, 5, 7]
   })

   # SBMM in Decoders
   decoder_alpha_subj = ModuleDict({
      subj_id: Sequential([
         Linear(1024, 1024)
      ]) for subj_id in [1, 2, 5, 7]
   })

   decoder_beta_subj = ModuleDict({
      subj_id: Sequential([
         Linear(1024, 1024)
      ]) for subj_id in [1, 2, 5, 7]
   })

   # Shared Decoder
   shared_decoder = Sequential([
      Linear(1024, 1024),
      LayerNorm((1024,)),
      Linear(1024, 8192)
   ])

   # Edge Prediction from Voxel Features
   vox2edge = Sequential([
      FC2Img(ConvTransposeAndResNet()), # Custom
          architecture combining ConvTranspose and
          ResNet blocks
      Sigmoid()
   ])
   # Color Prediction from Voxel Features
   vox2color = Sequential([
      FC2Img(ConvTransposeAndResNet()),
      Tanh()
   ])
   # Text Prediction from Voxel Features
   vox2text = Sequential([
      Linear(1024, 1024),
      ResMLP([MLPBlock(1024) for _ in range(2)]),
      Linear(1024, vocab_size) # E.g., 59136
   ])


   # Voxel Prediction from Edge
   edge2vox = Img2FC(ConvAndPoolingBlocks())

   # Voxel Prediction from Color
   color2vox = Img2FC(ConvAndPoolingBlocks())

   # Voxel Prediction from Semantic
   text2vox = Sequential([
      Linear(vocab_size, 1024),
      LayerNorm((1024,)),
      Linear(1024, 1024)
   ])


   # Translator MLP from Voxels to Representation
   translator2Rep = Sequential([
      ResMLP([MLPBlock(1024) for _ in range(4)])
   ])

   # Translator MLP from Representation to Voxels
   translator2Vox = Sequential([
      ResMLP([MLPBlock(1024) for _ in range(4)])
   ])
```

Figure 5. Qualitative comparison with various variants.

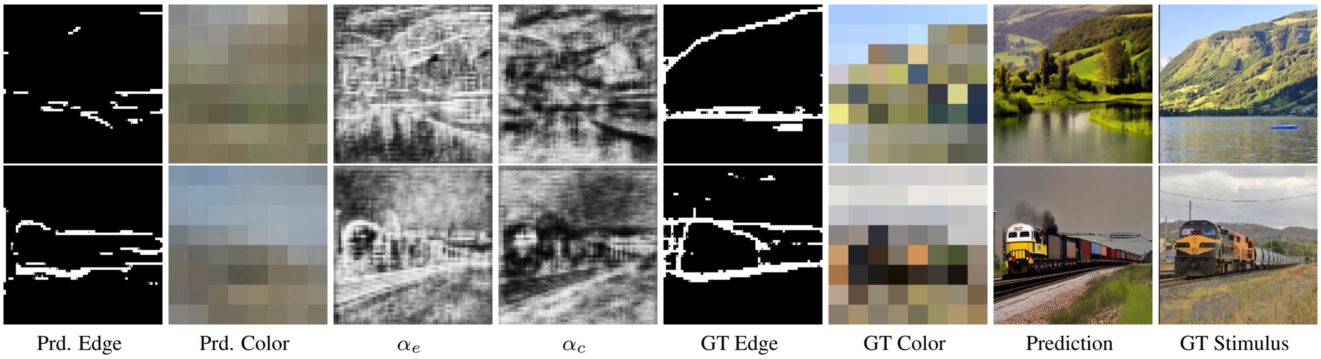


Figure 6. Visualization of VCM's output weights α_e and α_c, they control the fusion weights to relax the influence of predicted edge and color conditions to output.

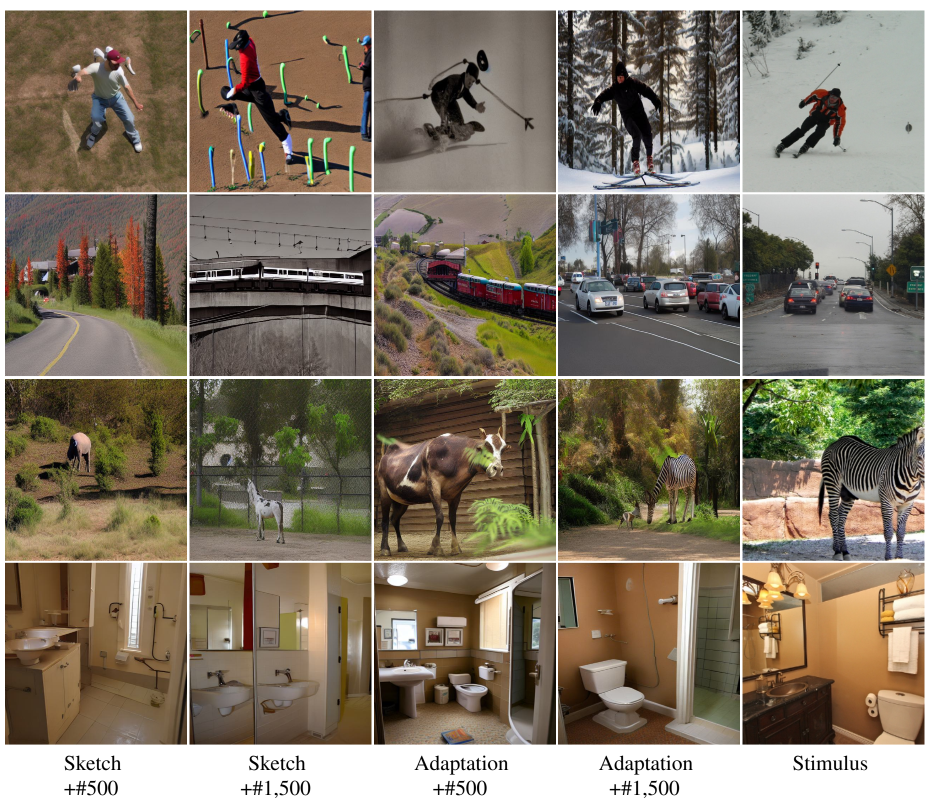
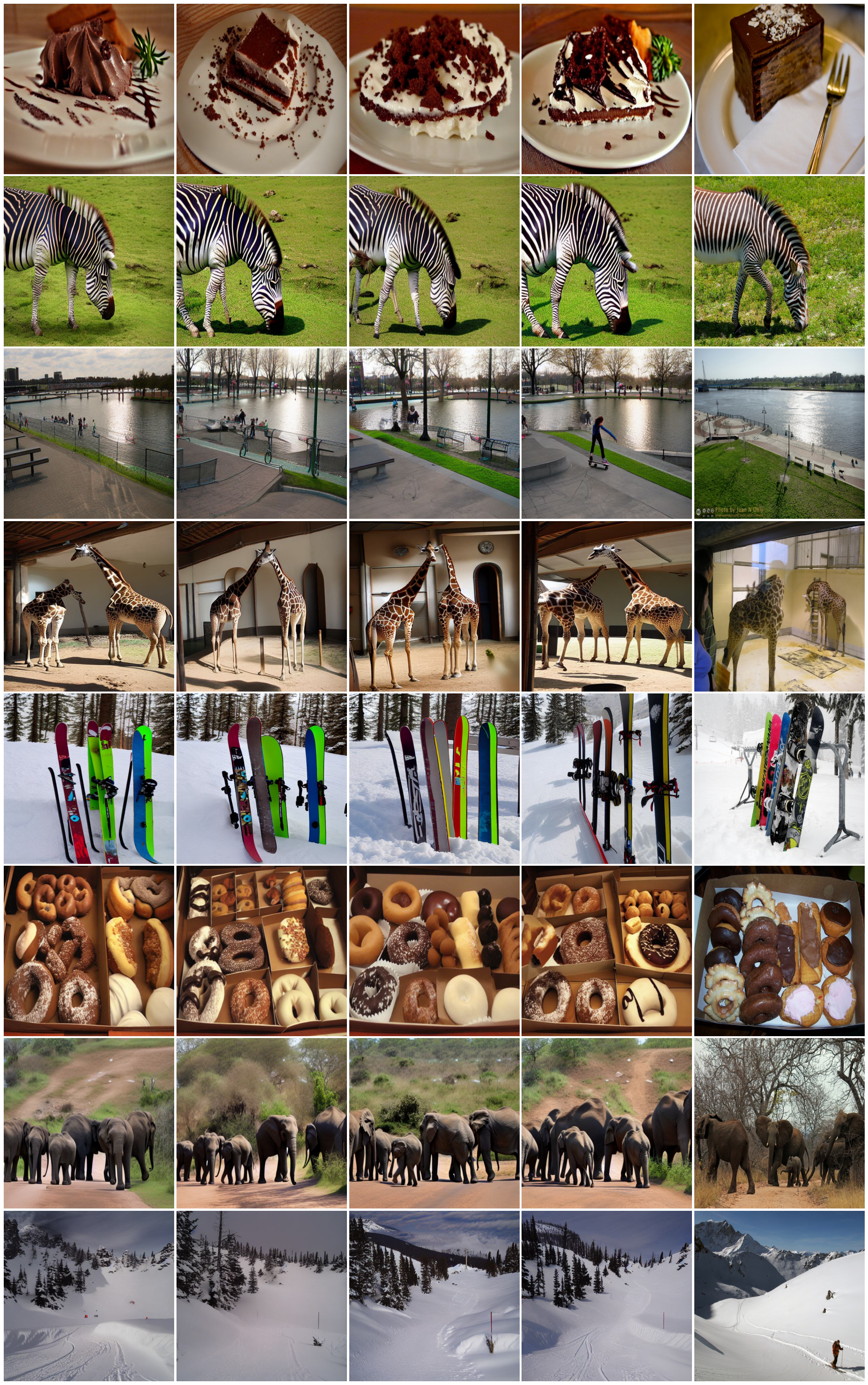


Figure 7. Qualitative comparison under different data limitation scenarios.



Subj.1 Subj.2 Subj.5 Subj.7 Stimulus

Figure 8. Our framework also supports the synthesis of fMRI for specific subject based on an unseen image, and the synthesized fMRI voxels can reconstruct the stimulus image faithfully.

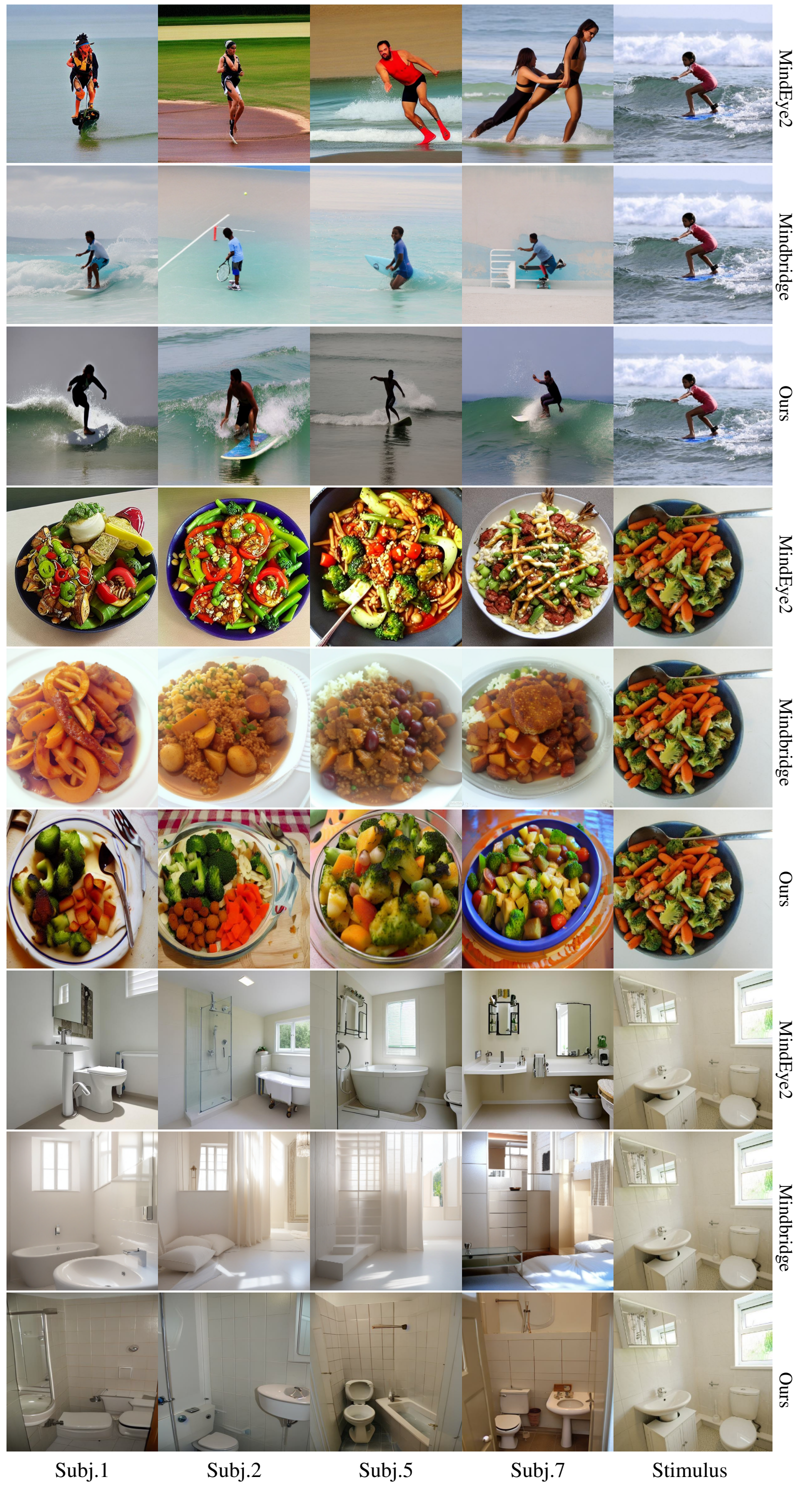


Figure 9. Comparison on Cross-subject Mind decoding.