Modeling the Brain's Grammar: ROI-Guided fMRI Pretraining for Transferable and Interpretable Vision Decoding

Supplementary Material

6. A Brief Introduction to MRL

Matryoshka Representation Learning (MRL)[13] learns a single high-dimensional embedding vector that contains nested, coarse-to-fine representations at multiple granularities. Instead of training separate models for different embedding sizes, MRL jointly optimizes a set of nested low-dimensional prefixes of the full embedding. Specifically, given a nested set of dimensions $\mathcal{M} = \{m_1, m_2, \ldots, m_k\}$ with $m_i < m_{i+1}$ (e.g., $\{8, 16, 32, \ldots, 2048\}$), MRL applies an independent classification loss to each prefix $\mathbf{z}_{1:m}$ of the embedding $\mathbf{z} = F(x) \in \mathbb{R}^d$ for every $m \in \mathcal{M}$. The total objective is formulated as:

$$\min_{\theta_F, \{W^{(m)}\}} \frac{1}{N} \sum_{i=1}^{N} \sum_{m \in \mathcal{M}} c_m \cdot \mathcal{L}\big(W^{(m)} \cdot F(x_i; \theta_F)_{1:m},\, y_i\big),$$

where $W^{(m)} \in \mathbb{R}^{L \times m}$ is a linear classifier for dimension m, $\mathcal{L}$ is typically the softmax cross-entropy loss, and $c_m \geq 0$ are optional weighting coefficients (often set to 1). This encourages the first m dimensions to form a standalone, high-quality representation ,as accurate as one trained independently,while sharing parameters across all scales. The result is a flexible, adaptive embedding that can be truncated at inference time to meet varying accuracy or computational requirements, without retraining or additional cost .

In our work, we use an MRL-like design as a regularization method to avoid overfitting and enhance the structure of the learned brain representation.

7. Mind editing

ROITok unlocks more potential applications in neural decoding. Here, we introduce a novel task termed **Mind Editing** that probes how decoded outputs change when ROI tokens are modified. Given an fMRI pattern, we replace selected ROI tokens with those from another image; as shown in Fig. 11, the reconstruction blends visual content from both sources. This demonstrates that our ROI-based decoder flexibly integrates information across brain regions rather than memorizing training examples, underscoring its potential for controllable and interpretable decoding. In the future, we may use ROITok to investigate how mental imagery is organized in the brain.

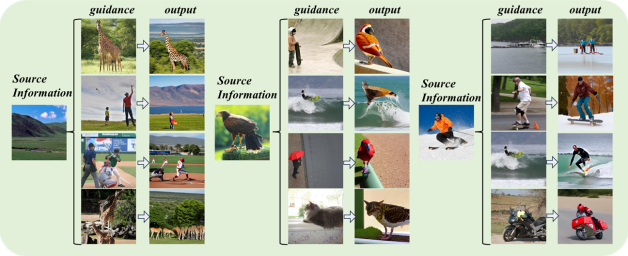

Figure 11. Decoding examples for the Mind Editing task: In these cases, we replace the "General" and "Early" ROI token of source image with that from another guidance image, yielding a reconstruction that blends visual content from both sources.

8. More ablations

8.1. Ablations on ROIs

Reason for ROI selection In the main paper, we employ a set of 12 ROIs, comprising both basic visual regions (such as V1 to hV4) and composite ROIs formed by aggregating multiple basic regions. For instance, the "Early" visual area combines V1, V2, and V3, while the "General" ROI includes all task-related voxels across the visual cortex. This hierarchical ROI design enhances the flexibility of our pretrained model, enabling straightforward adaptation to diverse downstream datasets. Since different datasets often provide varying ROI definitions and exhibit different signal-to-noise ratios, the multi-scale structure allows users to select the most appropriate level of regional granularity for their specific application.

The Influence of ROI Sequence Order Since our encoder backbone employs a residual MLP, one might wonder whether the ordering of ROI tokens affects the resulting MRL-like fMRI component representations. In our main experiments, ROI tokens are arranged in the following sequence: V1, V2, V3, hV4, PPA, FFA, OPA, RSC, General, Early, Middle, High. To investigate the sensitivity to this ordering, we reversed the sequence and retrained the model using the same 40-hour training data. The results, presented in Tab. 3, show that the performance remains largely unchanged, indicating that the model is robust to the specific ordering of ROI tokens. This suggests that the residual MLP encoder effectively integrates information across ROIs without relying on a fixed sequential structure.

Table 3. Ablation studies on the order of the ROI tokens. We reverse the ROI order used in the main paper and retain the models with full session NSD data. All results are averaged across subjects. The results show that the ROI order does not have a significant influence on the decoding performance.

Methods	Low- Level Recons				1000-way Retrieval	
	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Image↑	Brain↑
ROITok(40h, order in main paper)	.549	.514	89.6%	83.0%	97.3%	97.2%
ROITok(40h,reverse order)	.550	.514	89.5%	82.7%	97.1%	97.2%

Table 4. Ablation studies on the selection of the used ROI. In the main paper, we trained models with 12 ROIs from different levels. Here, we retrain models with 9 rois, with "Ealy","Middle", and "High" removed. The results show that using more ROIs is slightly better.

Methods	Low- Level Recons				1000-way Retrieval	
	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Image↑	Brain↑
ROITok(40h, 12 rois)	.549	.514	89.6%	83.0%	97.3%	97.2%
ROITok(40h,9 rois)	.539	.508	88.7%	82.3%	97.1%	96.7%

Table 5. Ablation studies on dimension of the MRL-like embedding components, with results averaged across subjects. In the main paper, D' is set to 400. These results shown that MRL-like fMRI embeddings can signifcantly enhence the decoding accuracy. And its performance is pretty stable across different values of D', while smaller value is slightly better for low-level reconstruction.

Methods	Low- Level Recons				1000-way Retrieval	
	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Image↑	Brain↑
ROITok(40h,w/o MRL-like)	.426	.493	85.8%	81.8%	96.3%	94.3%
ROITok(40h,D'=400)	.549	.514	89.6%	83.0%	97.3%	97.2%
ROITok(40h,D'=200)	.551	.514	89.3%	82.7%	97.2%	97.1%
ROITok(40h,D'=50)	.560	.515	89.6%	82.3%	97.1%	97.2%

An Alternative ROI Selection To examine the impact of ROI selection, we also conducted experiments using a reduced set of 9 ROIs: V1, V2, V3, hV4, PPA, FFA, OPA, RSC, and General. The results, reported in Tab. 4, indicate that including the additional composite ROIs (Early, Middle, High) yields a slight improvement in performance. This suggests that the hierarchical grouping of visual regions provides complementary information that marginally enhances model effectiveness.

8.2. Ablations on the dimension D' of fMRI component tokens

In the main paper, we report results using fMRI component tokens with dimension $D' = 400$. To further assess the impact of this design choice, we trained additional models with smaller values of D'. The results, summarized in Tab. 5, demonstrate that the MRL-like embedding structure consistently enhances decoding performance across different component dimensions, and maintains a substantial performance margin over models without this structural constraint. Interestingly, smaller values of D' appear to be more favorable for low-level visual reconstruction tasks, suggesting that a more compact representation better preserves fine-grained visual details.

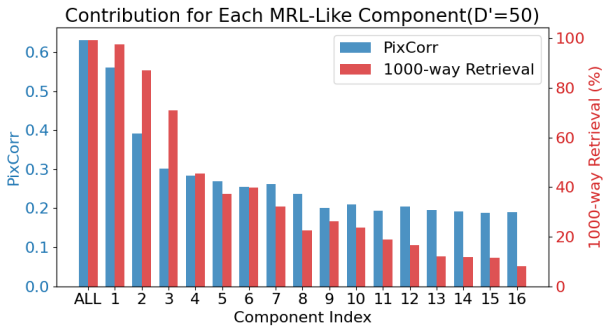

Figure 12. We evaluate each MRL-like component's contribution for retrieval and low-level reconstruction task when $D' = 50$. The results suggest a hierarchical structure of the learned fMRI embeddings.

We visualize the contribution of each of the first 16 MRL-like components to the retrieval and low-level reconstruction task for $D' = 50$ in Fig. 12. A clear hierarchical pattern emerges in these components. Additional visualizations of low-level reconstructions based on these components are provided in Fig. 13.

9. Individual Performance

In Tab. 6, we report the individual subject decoding performance on the Natural Scenes Dataset (NSD), corresponding to the final reconstruction results with a diffusion prior presented in the main paper.

10. More Experiment Details

Model and Inference Settings The residual MLP backbone comprises 8 MLP layers with a latent dimension of 4800 (equivalent to 12×400, matching the 12 ROI tokens each of dimension 400). The low-level reconstruction module is a VAE decoder consisting of three UpDecoderBlock2D blocks. The total number of trainable parameters during ROITok pretraining is approximately 900 million. The diffusion prior module contains 8 transformer decoder layers and has 81 million trainable parameters.

In the IP-Adapter-based approach, the fMRI component tokens are first projected into 4 semantic tokens before being fed into the pretrained Stable Diffusion U-Net.

At inference time, we adopt the default DDIM scheduler from the IP-Adapter framework for the IP-Adapter-based model, with a denoising process of 50 steps. For the diffusion-prior-based method, we use the UniPCMultistep

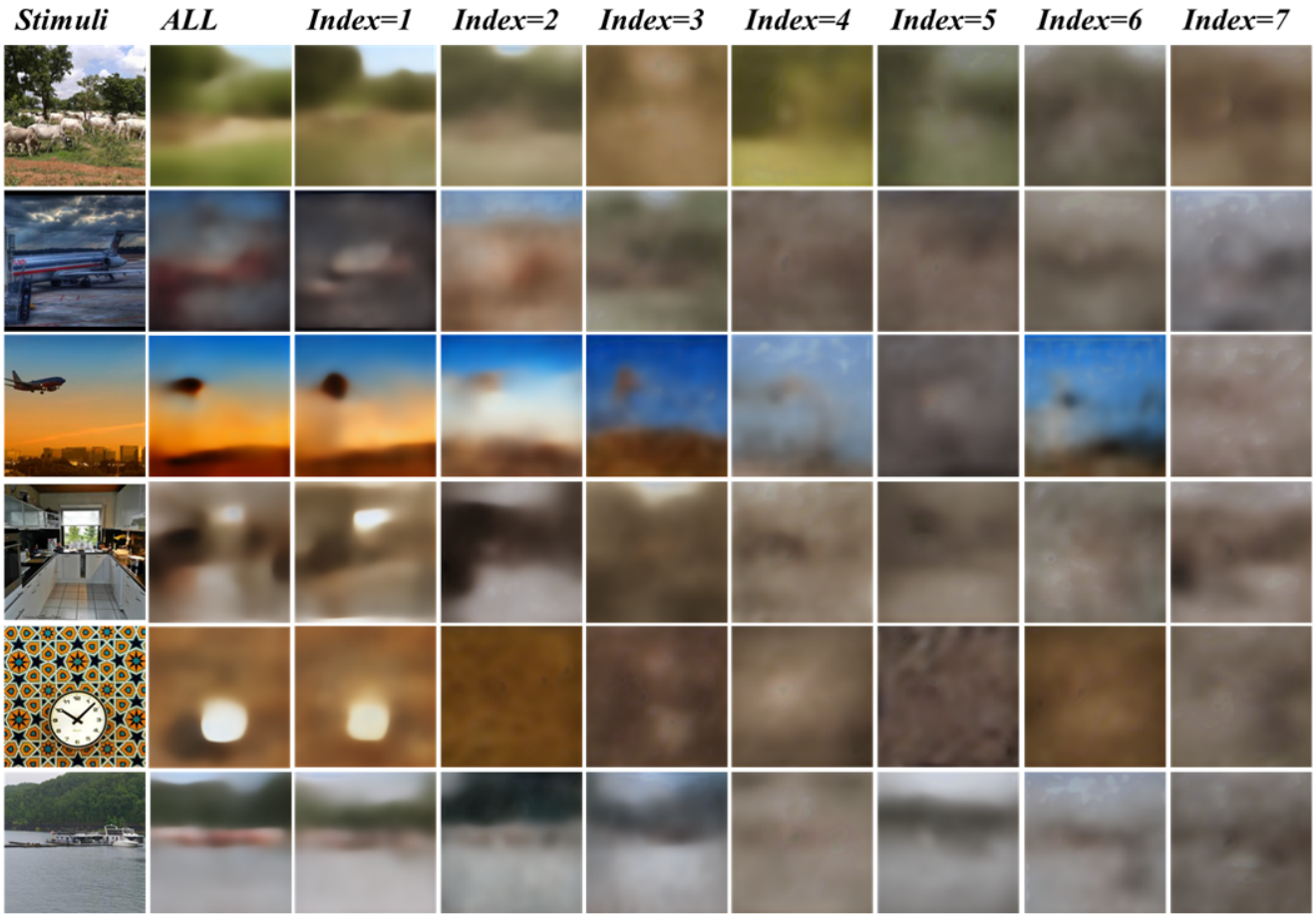

Figure 13. Low-level reconstruction examples for each MRL-like component. We only show the first 7 components, which show the gradual loss of information.

noise scheduler [45] with 20 denoising timesteps. For each test sample, we generate 16 candidate reconstructions and select the best one according to the similarity score from our retrieval branch.

SoftCLIP loss Let $\{X_i\}_{i=1}^N$ be a batch of predicted embeddings, and $\{Y_i\}_{i=1}^N$ be the target embeddings, where N is the batch size. SoftCLIP[31] loss is defined as:

$$SoftCLIP(X,Y) \tag{1}$$
$$= -\sum_{i=1}^{N}\sum_{j=1}^{N}[\frac{exp(\frac{Y_i \cdot Y_j}{\tau})}{\sum_{m=1}^{N} exp(\frac{Y_i \cdot Y_m}{\tau})} log(\frac{exp(\frac{X_i \cdot Y_i}{\tau})}{\sum_{m=1}^{N} exp(\frac{X_i \cdot Y_m}{\tau})})],$$

where τ is the temperature hyperparameter, and is set to 0.006 in our experiments.

Evaluation Metrics For the reconstruction task, we used both low- and high-level metrics as prior works did. For the retrieval task, apart from 300-way retrieval metric implemented by MindEye2, we add a 1000-way retrieval metric, which is a more strict evaluation that can better assess models' performance gap. If not specially indicated, the retrieval results reported in this paper are for 300-way retrieval task. The representational similarity analysis is conducted with the ROI tokens extracted from the test set of NSD.

11. More Dataset Information

NSD NSD is the largest publicly available fMRI-image dataset comprising recordings from 8 subjects while viewing images sampled from the COCO dataset [17]. The fMRI responses used in this study correspond to normalized beta estimates generated by GLMSingle. We used preprocessed, flattened fMRI voxels in the native 1.8-mm volumetric space. ROITok was initially developed using only data from Subject 1. Data from all other subjects were held out entirely until the final stage of model training and eval-

Table 6. Individual decoding results on NSD.

Methods	Low- Level				High-Level				Retrieval	
	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓	Image↑	Brain↑
40h data of NSD										
ROITok(sub1)	.538	.363	99.3%	99.5%	96.5%	96.0%	.578	.321	99.9%	99.9%
ROITok(sub2)	.514	.358	98.9%	99.2%	95.9%	95.2%	.597	.330	99.9%	99.9%
ROITok(sub5)	.417	.343	96.9%	99.0%	96.3%	95.6%	.606	.335	98.4%	98.3%
ROITok(sub7)	.409	.341	96.9%	98.4%	94.2%	94.2%	.638	.358	96.6%	96.6%
ROITok(sub1, Low-level)	.628	.527	93.7%	87.1%	62.3%	63.2%	.983	.646	-	-
ROITok(sub2, Low-level)	.596	.522	92.3%	85.6%	61.3%	62.7%	.985	.648	-	-
ROITok(sub5, Low-level)	.488	.504	86.6%	80.1%	60.0%	60.9%	.995	.646	-	-
ROITok(sub7, Low-level)	.483	.503	85.7%	79.1%	58.9%	60.7%	.996	.648	-	-
1h data of NSD										
ROITok(sub1)	.362	.330	95.0%	96.7%	89.9%	89.1%	.711	.404	95.5%	95.1%
ROITok(sub2)	.328	.325	93.9%	96.4%	89.2%	88.3%	.729	.415	94.7%	94.0%
ROITok(sub5)	.281	.309	90.3%	95.2%	90.3%	89.2%	.720	.414	83.3%	80.3%
ROITok(sub7)	.243	.311	88.7%	91.6%	82.8%	83.8%	.790	.464	73.9%	72.5%
ROITok(sub1, Low-level)	.429	.498	80.7%	74.8%	58.1%	58.8%	1.003	.647	-	-
ROITok(sub2, Low-level)	.390	.494	78.4%	73.0%	57.4%	58.1%	1.005	.649	-	-
ROITok(sub5, Low-level)	.342	.485	73.8%	69.7%	57.1%	58.1%	1.007	.649	-	-
ROITok(sub7, Low-level)	.299	.481	70.4%	66.7%	56.4%	56.2%	1.008	.650	-	-

uation. Both training and test fMRI data were normalized using voxel-wise Z-scoring, with the mean and standard deviation for each voxel computed exclusively from the training set. Following MindEye2,The normalization applied to the test set remains identical across experiments that vary in the amount of training data.

GOD To evaluate cross-dataset transferability, we adopt the Generic Object Decoding (GOD) dataset [8], which comprises 1,200 training and 50 test samples per subject under a zero-shot setting (test categories are disjoint from training categories).

12. More Visualization

Decoding Noisy fMRI Fig. 14 presents reconstruction results under varying levels of additive noise. These results demonstrate that fMRI signals retain decodable visual information across a wide range of noise conditions.

High-Resolution Reconstruction for Each MRL-like Component We present high-resolution decoding results for individual MRL-like components in Fig. 15.

Additional Examples for Mind Editing Additional examples for the Mind Editing task are provided in Fig. 16. We replace the "General" and "Early" token of the source fMRI pattern with that from the guidance fMRI pattern. These results highlight ROITok's capacity to effectively integrate distributed neural information across multiple ROIs, enabling coherent and contextually aligned image generation.

More Reconstructions Examples for NSD and GOD More low-level and final reconstruction examples for NSD and GOD are provided in Fig. 17 and Fig. 18, respectively.

12.1. Limitations

First, our method is based on fMRI, which has low temporal resolution and is not suitable for real-time brain decoding. Second, the current model is trained and evaluated primarily on the Natural Scenes Dataset , whose visual stimuli are drawn from the COCO dataset; this limits the model's generalization to other visual domains such as abstract art, medical images, or non-naturalistic scenes. Future work will explore incorporating more diverse image sources, through supervised or unsupervised pretraining, to improve cross-domain robustness and visual reconstruction quality.

13. Potential Negative Societal Impacts

Our work uses only open-source fMRI data (e.g., NSD) and involves no new data collection, thus avoiding direct ethical issues around consent or privacy. And our current model requires subject-specific training and does not directly generalize across individuals, which limits immediate misuse

Figure 14. Examples of decoding noisy fMRI with IP-adapter and Diffusion Prior. For step 0 to step 999, the fMRI patterns are gradually corrupted by additive gaussian noises.

risk.

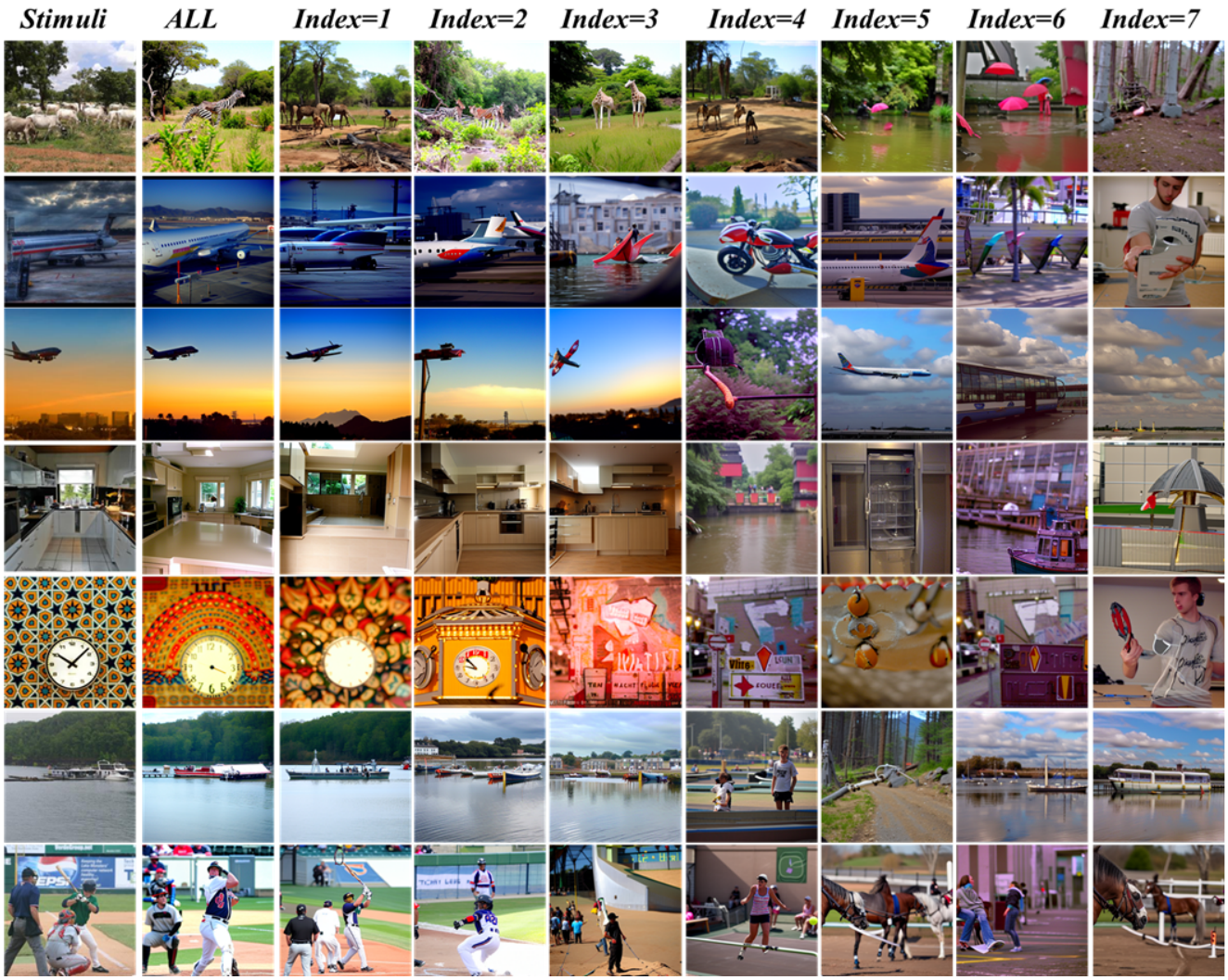

Figure 15. High-resolution reconstruction examples generated by diffusion-prior-based model for each MRL-like component.

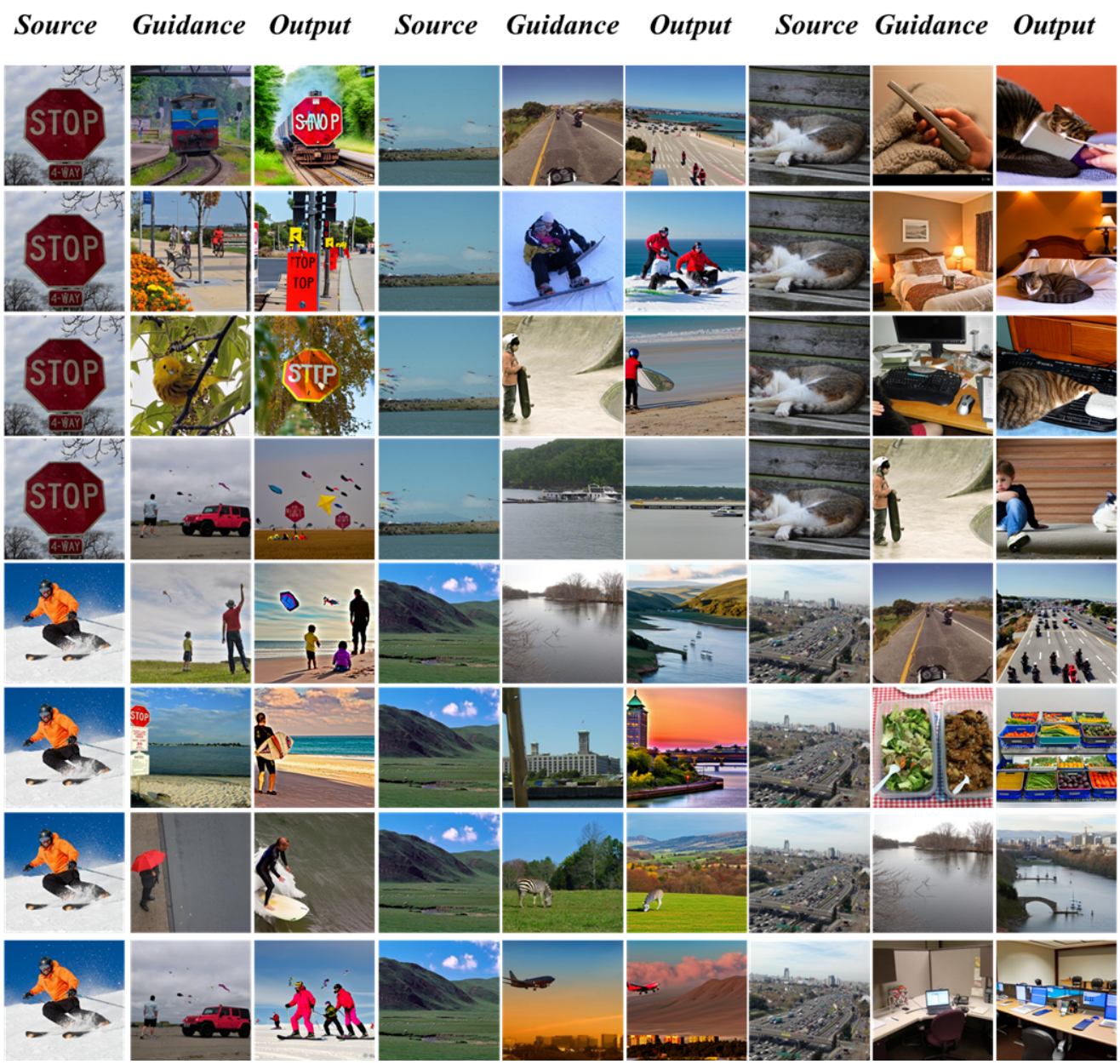

Figure 16. More examples of Mind Editing task. We replace the ”General” and ”Early” token of the source fMRI pattern with that from the guidance fMRI pattern. These results demonstrate ROITok’s ability of integrating distributed information from different ROIs.

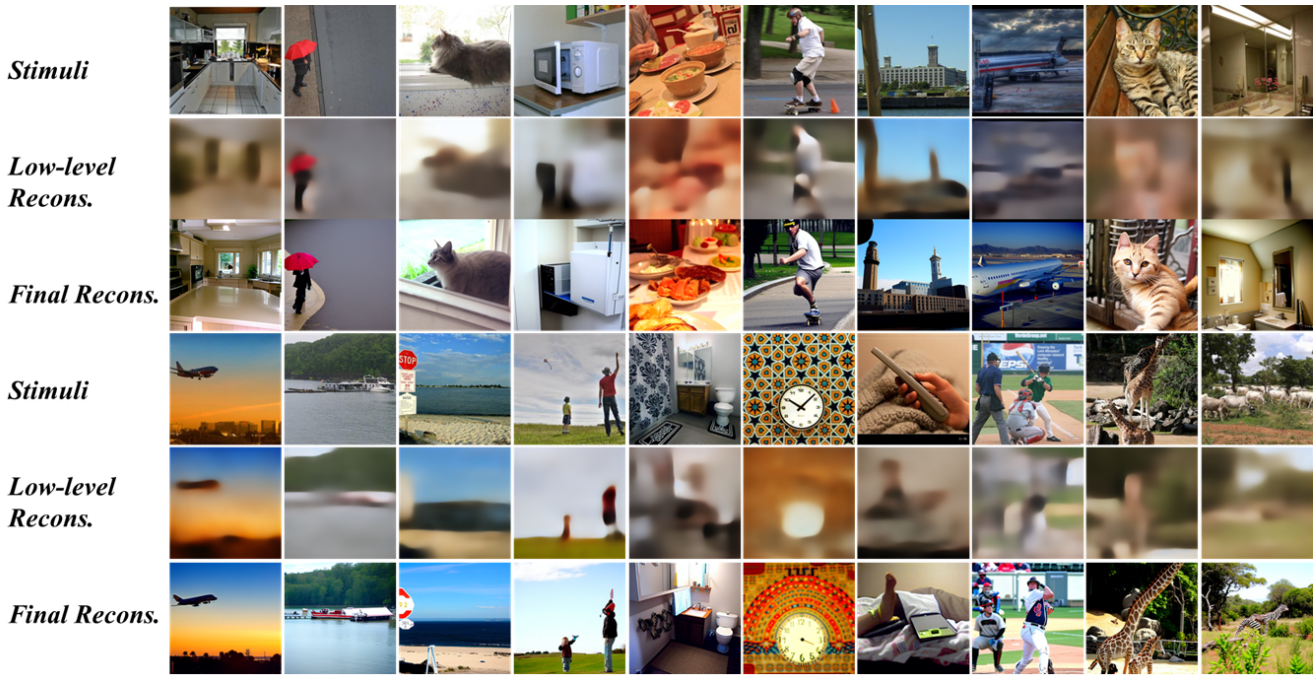

Figure 17. More examples of low-level and high-resolution reconstructions for NSD. Results are generated with diffusion prior.

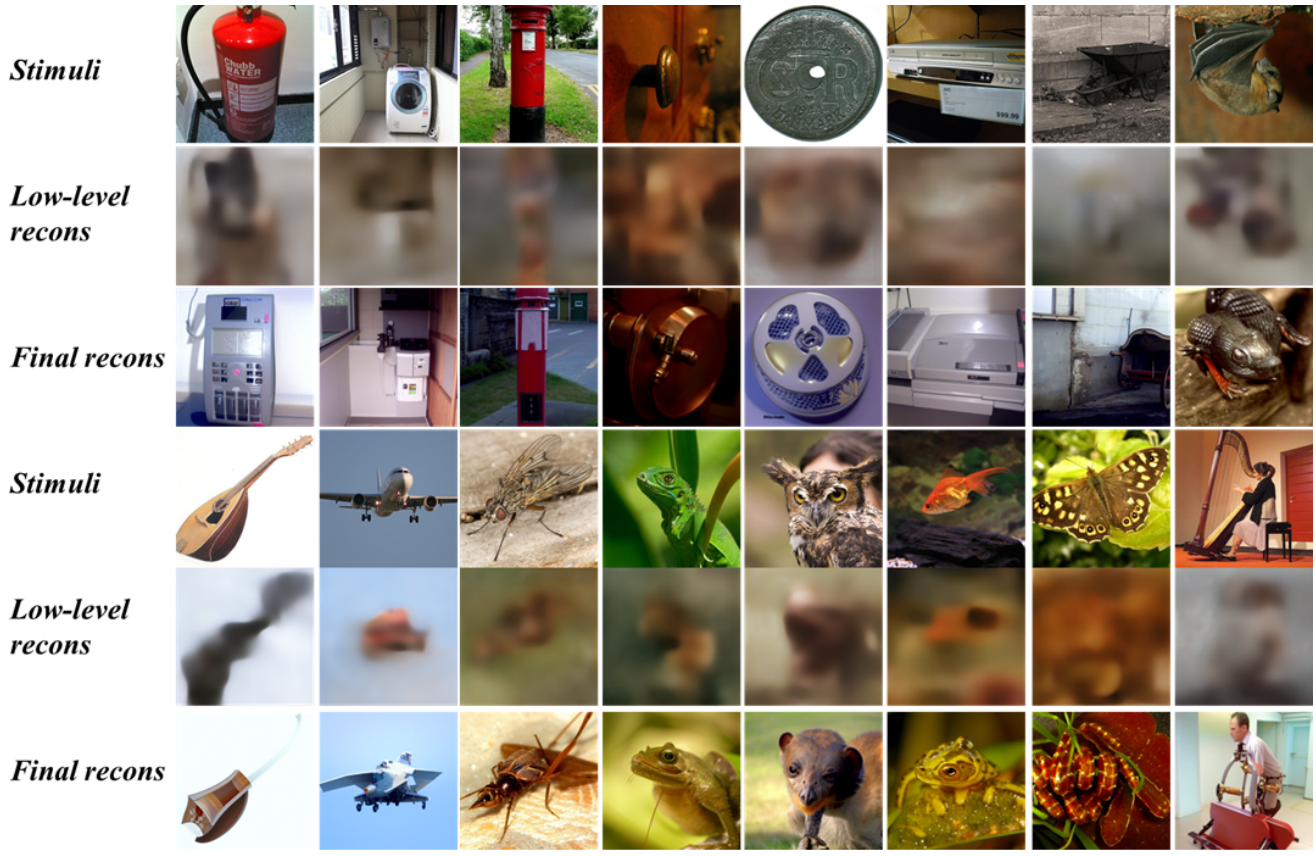

Figure 18. More examples of low-level and high-resolution reconstructions for GOD. Results are generated with diffusion prior.