Appendix

A.1 Reconstruction Performance

The bar chart Figure A.1 below shows a comparison between our proposed model and various baseline models on the GOD subjects 1, 2, 4, and 5. The performance of subject 3 has been detailed in the main text.

As shown in the figure, **our model substantially outperforms the previous state-of-the-art method (DC-LDM) on all GOD subjects**. Specifically, our model surpasses DC-LDM by around $110\%, 16.8\%, 24.7\%, 11.8\%$ in GOD subjects 1, 2, 4, and 5, respectively. To achieve DC-LDM's reported performance in its original paper [6], this method need signals from test set fMRI data. This is not a setting adopted by other baselines. To ensure a fair evaluation, we banned DC-LDM from tuning on the test set in the main paper. But we show here that, **our model still largely exceeds DC-LDM on GOD subjects even after DC-LDM is tuned on the test set fMRI data.** As depicted in Figure A.1, compared to DC-LDM-test-tuned, our model achieves an improvement in accuracy of $63.9\%, 36.1\%, 14.5\%, 22.8\%$ in GOD subject 1, 2, 4 and 5, respectively.

Additionally, we provide the performance of our model on BOLD5000 subjects 1, 2, 3, and 4 in Table A.1. Following previous work [6], all results are presented in 50-way-top-1 classification accuracy.

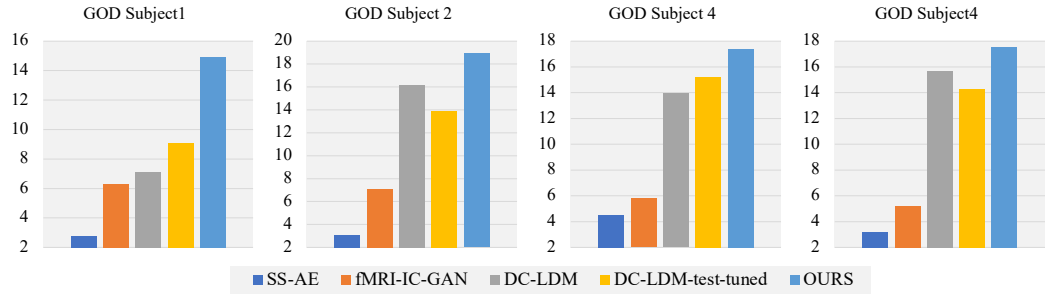

Figure A.1 Reconstruction performance of our model and other baselines on GOD subjects 1, 2, 4 and 5, measured by 50-way-top-1 classification accuracy

BOLD 5000	CSI 1	CSI 2	CSI 3	CSI 4
OURS	25	18.69	16.14	18.98

Table A.1 Reconstruction performance of our model on BOLD5000 subject CSI 1-4, measured by 50-way-top-1 classification accuracy.

A.2 Examples of Reconstructed Images

Figures A.2 and A.3 present images generated by our model using fMRI data from GOD and BOLD5000 datasets, respectively. We generated all images at a resolution of $256 \times 256 \times 3$ using 250 PLMS steps. More samples can be generated using our code base in the supplementary materials. The code will be open-sourced with the camera ready version of this paper.

A.2.1 Reconstructed Images from GOD Dataset

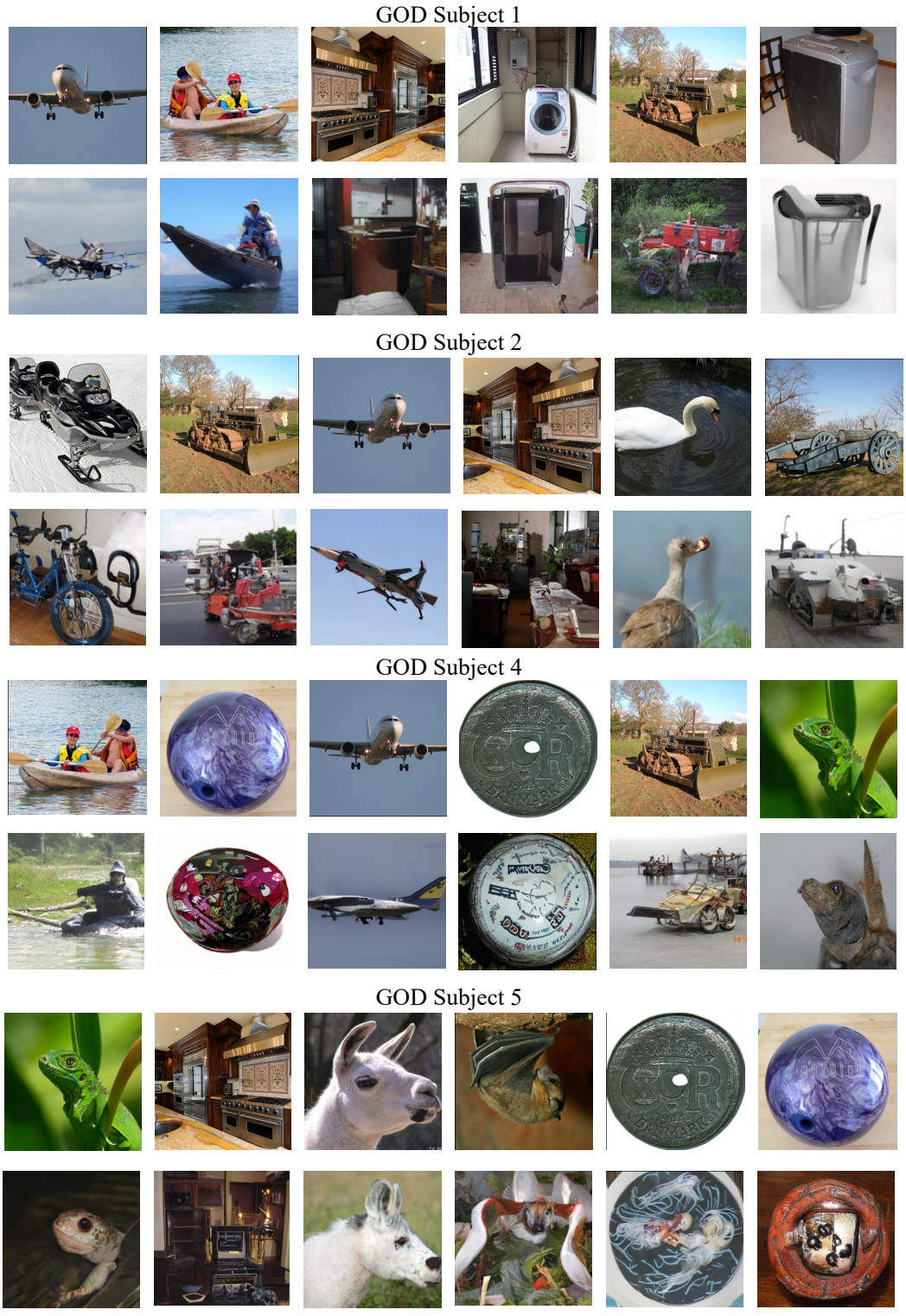

Figure A.2: Randomly selected reconstructed images from GOD subject 1, 2, 4 and 5. For each subject, the upper line shows the ground truth images while the lower line shows the reconstructed images by our method.

A.2.2 Reconstructed Images from BOLD5000 Dataset

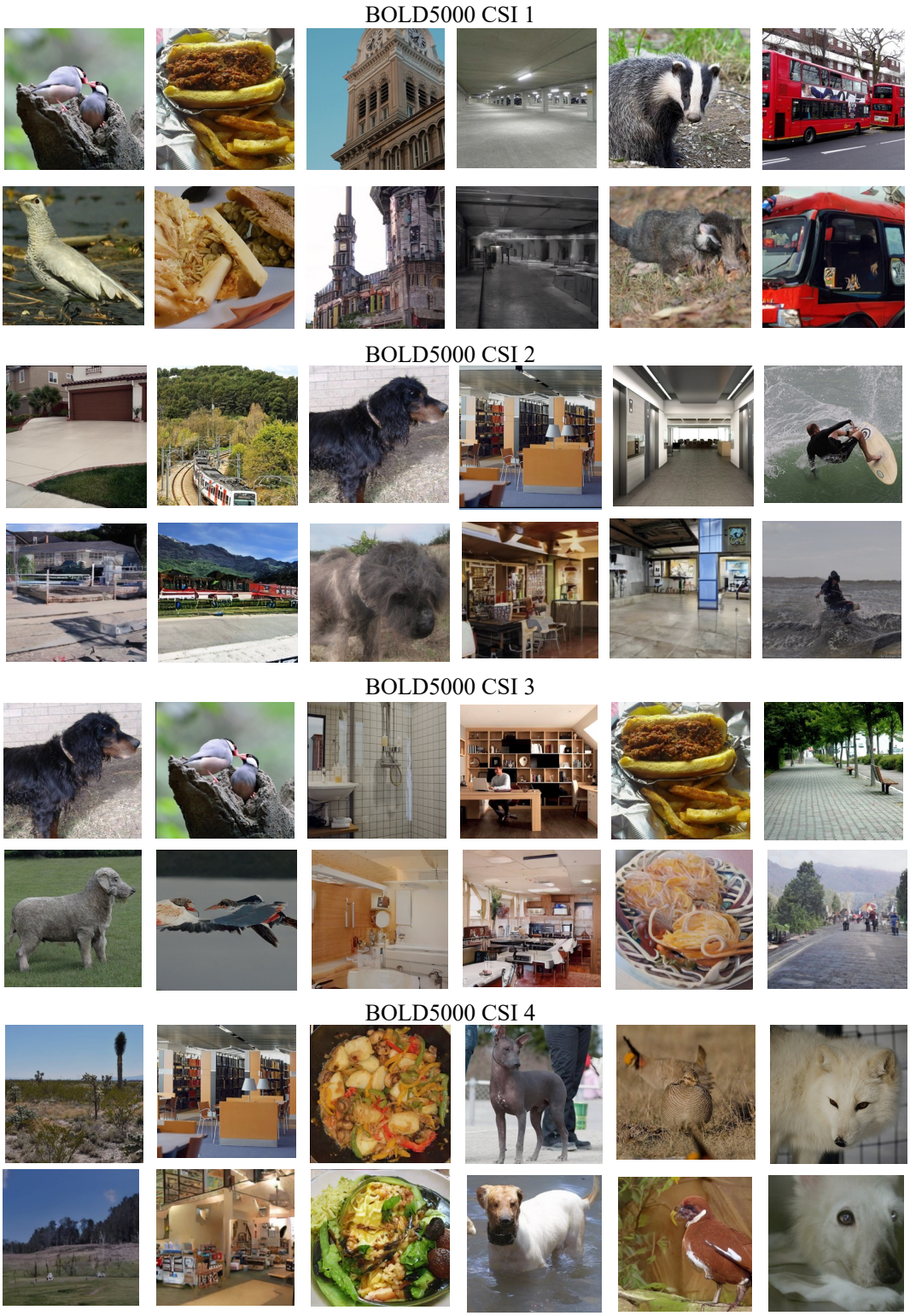

Figure A.3: Randomly selected reconstructed images from BOLD5000 CSI 1-4. For each subject, the upper line shows the ground truth images while the lower line shows the reconstructed images by our method.

A.3 FMRI Dataset Introduction

HCP The Human Connectome Project (HCP) originally serves as an extensive exploration into the connectivity of the human brain. It offers an open-source database of neuroimaging and behavioral data collated from 1,200 healthy young adults within the age range of 22-35 years. Currently, it stands as the largest public resource of MRI data pertaining to the human brain, providing an excellent foundation for the pre-training of brain activation pattern representations. Of the subjects involved, 1113 underwent scanning via a Siemens Skyra Connectom scanner for 3T MR, while a Siemens Magnetom scanner for 7T MR was utilized for the remaining 184. For the scope of this paper, we will predominantly focus on the data derived from the more populated 3T dataset.

GOD The Generic Object Decoding (GOD) Dataset is a specialized resource developed for fMRI-based decoding. It aggregates fMRI data gathered through the presentation of images from 200 representative object categories, originating from the 2011 fall release of ImageNet. The training session incorporated 1,200 images (8 per category from 150 distinct object categories). In contrast, the test session included 50 images (one from each of the 50 object categories). It is noteworthy that the categories in the test session were unique from those in the training session and were introduced in a randomized sequence across runs. The fMRI scanning was conducted on five subjects.

BOLD5000 The BOLD5000 dataset is a result of an extensive slow event-related human brain fMRI study. It comprises 5,254 images, with 4,916 of them being unique. This makes it one of the most comprehensive publicly available datasets in the field. The dataset's principal advantage is its high diversity, enabling the capture of the complexity and variability inherent in natural visual stimuli. The images in BOLD5000 were selected from three popular computer vision datasets: ImageNet, COCO, and Scenes. ImageNet provided 1,916 images primarily focusing on singular objects. Meanwhile, COCO contributed 2000 images featuring multiple objects, and Scenes contributed 1000 images depicting hand-crafted indoor and outdoor scenes. Four participants labeled CSI1 through CSI4, were involved in this study and underwent scanning via a 3T Siemens Verio MR scanner equipped with a 32-channel phased array head coil.

A.4 Implementation Details

A.4.1 FMRI Representation Learning (FRL)

For both FRL Phase 1 and Phase 2, the fMRI auto-encoder is the same ViT-based masked auto-encoder (MAE). We divided fMRI voxels into patches and transformed them into embeddings using a one-dimensional convolutional layer with a patch-size stride. We employed an asymmetric architecture for the fMRI auto-encoder, in which the decoder is considerably smaller with 8 layers than the encoder with 24 layers. We used a larger embedding-to-patch size ratio, specifically a patch size of 16 and an embedding dimension of 1024 for our model. Our design choice expands the representation dimension of fMRI data, which increases the information capacity of the fMRI representations. To address the data-hungry nature of models like the Vision Transformer (ViT), we used random sparsification (RS) as a form of data augmentation, randomly selecting and setting 20% of voxels in each fMRI to zero.

FRL Phase 1 In Phase 1, we train the masked ViT-based fMRI auto-encoder with contrastive loss. For GOD subject 1,4,5 and BOLD5000 CSI 1,2, self-contrastive (γ_s) and cross-contrastive (γ_c) loss weights are both 1. The masking ratio is 0.5. For GOD subject 2,3 and BOLD5000 CSI 3,4, $\gamma_s = 1$ and $\gamma_c = 0.5$, masking ratio is 0.75. Optimizing contrastive losses prefers a larger batch size. So we set the batch size to 250 and train for 140 epochs on one NVIDIA A100 GPU. We train with 20-epoch warming up and an initial learning rate of 2.5e-4. We optimize with AdamW and weight decay 0.05.

FRL Phase 2 In Phase 2, we tune the fMRI autoencoder jointly with an image auto-encoder, which is a pre-trained ViT-based MAE released by [47]. The image auto-encoder has a 12-layer encoder with a 768 hidden size and a 6-layer decoder with a 512 hidden size. We set the batch size to be 16 and train for 60 epochs. We train with 2-epoch warming up. The initial learning rate is 5.3e-5. We optimize with AdamW and weight decay 0.05. We freeze the parameters of the decoder of the image-autoencoder and only tune the encoder.

A.4.2 Fine-tuning LDM

In this stage, we jointly optimize the parameters of LDM cross-attention heads and the fMRI encoder, while keeping other parameters of LDM unchanged. Given an fMRI-image pair, we first use the pre-trained VQ encoder to encode the image to obtain the latent representation which is further used as an objective to guide the joint training of the fMRI encoder and LDM cross-attention heads. During training, the fMRI data passes through the fMRI encoder trained using FRL, producing a patchified representation. This representation is then projected into key and value representation of cross-attention modules in the UNet of LDM. Furthermore, it is added to the time embedding to conduct double conditioning. The training follows the regular training pipeline of the diffusion model, where the model is optimized to learn to predict the Gaussian noise added to the image latent representation at each time step with the guidance of the given conditioning information. Here, we use the output of the fMRI encoder as the conditioning information. We conduct training with the following parameters: the batch size of 5, diffusion steps of 1000, the AdamW optimizer, a learning rate of $5.5e-5$, and an image resolution of $256 \times 256 \times 3$.

A.5 Evaluation Metrics

We use the common N-trial, n-way top-1 semantic classification as the main evaluation metrics. This evaluation method is summarized in the algorithm below:

Algorithm 1 Iterative Reasoning Module

Input:
pre-trained image classifier F, generated image $\hat{x}$, corresponding ground truth (GT) image x_{gt}
Output:
success rate $sr \in [0, 1]$
for $trail = 1$ to N **do**
$y_{gt} = F(x_{gt})$ get the prediction of GT image
$pred = F(\hat{x})$ get the output probabilities of generated image
$p = \{p_g, p_{y_1}, ..., p_{y_{n-1}}\}$ generate probabilities set contains $n-1$ randomly selected from $pred$ and y_{gt}
Success if $\arg\min_y = y_{gt}$
end for
return sr = number of success / N