Appendix

A. More Generation Samples

All samples are generated at a resolution of $256 \times 256 \times 3$ with 250 PLMS [29] steps. More samples can be found and generated in our code base.

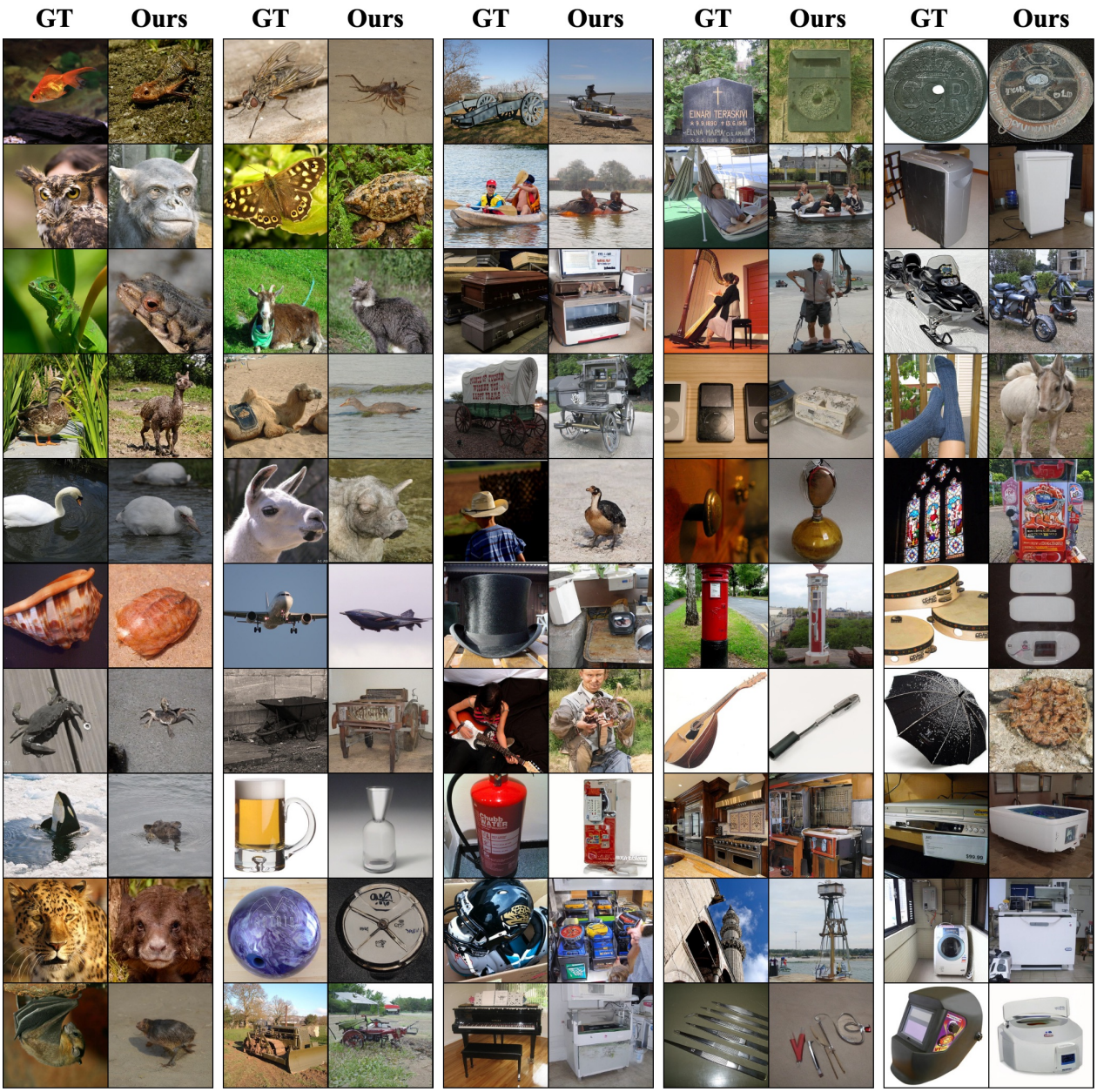


Figure A.1. Full Samples for Subject 3 in GOD.

Figure A.2. Full Samples for BOLD5000(Cont.).

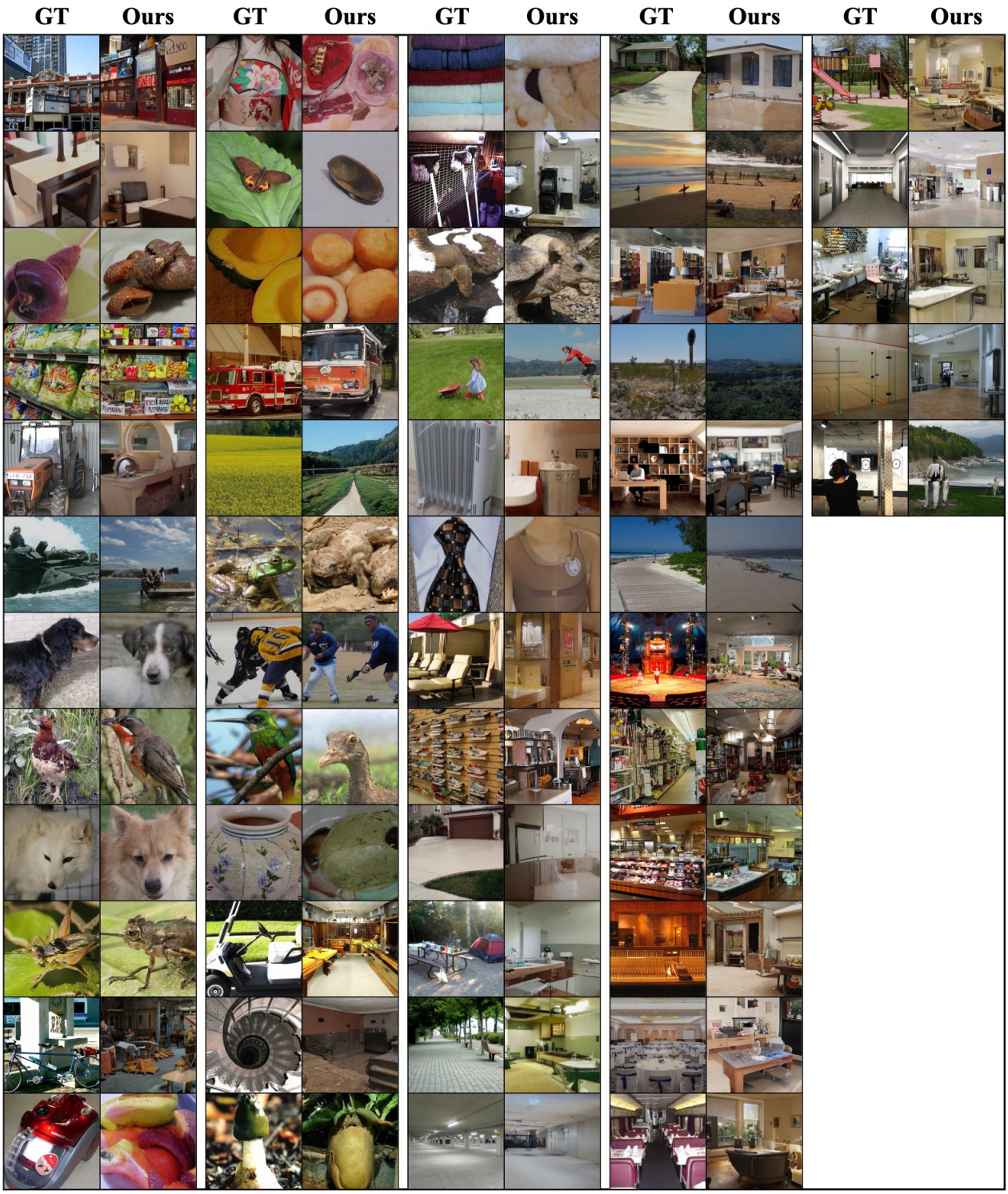
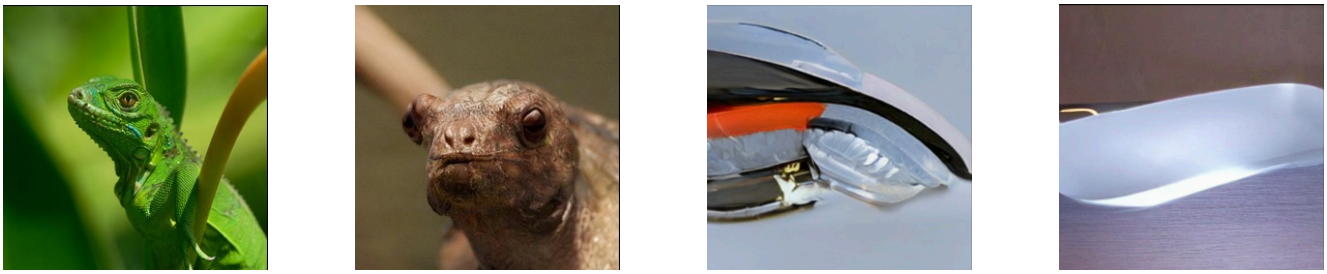


Figure A.3. Full Samples for BOLD5000.



(a) GT (b) Full Model (c) w/o SC-MBM Small (d) w/o SC-MBM Large

(e) Optimize Encoder only (f) Single Conditioning (g) Text-to-Image Pre-trained LDM (LAION) (h) Layout-to-Image Pre-trained LDM (OpenImages)

Figure A.4. Samples for Ablation Study. (a) Ground-truth stimulus. (b) Full model: SC-MBM pre-training; optimize the fMRI encoder and cross-attention heads; double-conditioned; Label-to-Image pre-trained LDM (ImageNet). (c) Model with small fMRI encoder without SC-MBM pre-training. (d) Model with the same fMRI encoder as the Full model without SC-MBM pre-training. (e) Optimize the fMRI encoder only, keep the cross-attention heads untouched. (f) Single conditioning. All other parameters are the same with the Full Model. The samples are obtained after finetuned for 500 epochs. See Tab. 1 and Tab. 2 for quantitative results of full test samples.

Figure A.5. Typical Failure Cases of Our Method. As discussed in the main text, we assume the failure cases are related to two reasons. On one hand, the GOD training set and testing set have no overlapping classes. That is to say, the model could learn the geometric information from the training but cannot infer unseen classes in the testing set. On the other hand, subjects might have some other stimuli-unrelated thoughts, which could be captured by fMRI and decoded by our method

B. Dataset and fMRI Preprocessing Details

Human Connectome Project (HCP) 1200 Subject Release [55]: large-scale magnetic resonance imaging dataset used for pre-training. We utilized around 2000×15-min 3T resting state fMRI runs from 1091 subjects. The visual cortex (V1-V4) defined in [17] is used as the ROI, which gives approximately 4000 voxels.

Generic Object Decoding Dataset (GOD) [21]: human fMRI scans with 1250 distinct images from ImageNet as a visual stimulus. During the fMRI scan, subjects were instructed to fixate on a cross located at the center of the presented images. This dataset consists of 1250 natural images from 200 distinct classes from ImageNet, where 1200 images are used for training. The remaining 50 images from classes not present in the training data are used for testing. Each image in the training set is presented once to the subject during the scan, while each image in the testing set is presented 35 times. Following the preprocessing in [21], the 35 repetitions are averaged for each image to create a higher SNR fMRI sample for testing. This dataset is widely used in brain image decoding [2, 13, 16, 31, 45, 46]. We used the manually defined ROIs (V1-V4, FFA, LOC, HVC) from the functional localizer runs provided in [21]. Altogether, the ROIs have around 4500 voxels per subject, with some individual variance as shown in Fig. 2.

Brain, Object, Landscape Dataset (BOLD5000) [5]: human fMRI study with 5,254 fMRI-image pairs from 4,916 distinct natural images (including various objects and indoor/outdoor scenes) from Scene UNderstanding (SUN) [61], Common Objects in Context (COCO) [28] and ImageNet [7]. In this dataset, 4,803 images are presented once, and 113 images are repeated twice or three times. The repeated data are also averaged to construct the testing set as in the GOD dataset. To the best of our knowledge, this dataset is the first time is applied to an image reconstruction task. The author also provided manually annotated ROIs based on a functional localizer. As a result of different scanning resolutions and ROI definition methods, the number of voxels in the defined ROIs is approximately 1,500 for each subject. Nonetheless, we show in our results that the pre-trained encoder can be directly applied to this dataset despite this difference in ROI definition and size.

Pre-training dataset Our upstream pre-training dataset is comprised of fMRI recordings from HCP and GOD. Following the processing step in target dataset [21], we averaged every 8.64 seconds (*i.e.* 12 time frames) of scans from HCP, which gives 130,000 fMRI time points. Including the training and testing fMRI in GOD, we have a pure fMRI dataset of 136,000 samples for pre-training. This pre-training dataset is, by far, the largest pre-training fMRI dataset used in this task.

To handle the different voxel numbers, all fMRI data are first padded to the maximum length in a wrap-around manner and then padded to the boundary of the patch size. Additionally, training fMRI is normalized to have zero mean and unit standard deviation. The testing samples are normalized with the mean and standard deviation from the training set.

C. Results on Different Subjects

The GOD consists of five different subjects, and the BOLD5000 consists of four subjects. The signal-to-noise ratio (SNR) is usually used to quantify the quality of a dataset. A higher SNR means better data quality. As reported by their authors respectively, the BOLD5000 has a much higher SNR than GOD. Within the GOD, the SNR differs among subjects as shown in Tab. C.1, where Subject 3 has a significantly higher SNR than the others. A higher SNR leads to better performance in our experiments, which has also been shown in various literature. Other than possible noise introduced during the scan, the SNR is also related to the subjects' on-line processing or information processing ability. Subjects with better information processing ability (*i.e.* better learners) will have a higher SNR during the scan under the same scanning conditions.

Dataset	GOD					BOLD5000			
Subject	Sub1	Sub2	Sub3	Sub4	Sub5	CSI1	CSI2	CSI3	CSI4
Acc (%)	9.1	13.9	27.4	15.2	14.3	34.5	18.5	21.0	20.9
FID	2.2	1.6	1.7	2.7	2.4	1.2	1.9	1.4	1.3
SNR	$0.064_{\pm 0.07}$	$0.061_{\pm 0.05}$	$0.10_{\pm 0.11}$	$0.092_{\pm 0.1}$	$0.065_{\pm 0.06}$	$4.65_{\pm 0.2}$	$5.20_{\pm 0.2}$	$5.55_{\pm 0.35}$	$5.40_{\pm 0.1}$

Table C.1. Full Results for All Subjects. The accuracy is obtained from the 1000-trials 50-way top-1 semantic classification test on the best-generated samples. SNR: signal-to-noise ratio. The voxel-wise mean SNR is obtained from [21] and [5] respectively.

D. More Implementation Details

D.1. Evaluation Metric Implementation

This algorithm performs the N-trial, n-way top-1 semantic classification test. It measures the semantic accuracy of generated images. We describe our evaluation method in Algorithm D.1, where the generated image and its corresponding ground-truth

image are denoted by x and $\hat{x}$ respectively, and y is for the class label. This metric relies on a pre-trained ImageNet classifier to determine whether x and $\hat{x}$ belong to the same class rather than using handcrafted features to represent each class. This method is thus reasonable and easily reproducible. We used a pre-trained ResNet as the classifier. We also showed that using other model based pre-trained classifiers will not change the result of this metric.

Algorithm D.1 N-Trials n-way Top-1 Accuracy Classification

Input pre-trained classifier $\mathcal{C}(\cdot)$, image pair (Generated Image x, Corresponding GT Image $\hat{x}$)
Output success rate $r \in [0,1]$
for N trials **do**
    $\hat{y} \leftarrow \mathcal{C}(\hat{x})$ get the ground-truth class
    $\{p_0,...,p_{999}\} \leftarrow \mathcal{C}(x)$ get the output probabilities
    $\{p_{\hat{y}},p_{y_1},...,p_{y_{n-1}}\} \leftarrow$ pick n-1 random classes
    success if $\underset{y}{\operatorname{argmax}}\{p_{\hat{y}},p_{y_1},...,p_{y_{n-1}}\} = \hat{y}$
end for
$r =$ number of success / N

D.2. SC-MBM Pre-training

In Masked Image Modeling (MIM) [18], images are divided into patches which are sequentially transformed into embeddings to adapt to a transformer-based architecture [10]. Following this practice, we divided fMRI voxels into patches which will be subsequently transformed into embeddings using a one-dimensional convolutional layer with a stride of the patch size.

A patch size of 16 and an embedding dimension of 1024 were used as the Full model. Notice that our embedding size to patch size ratio is much larger than that of MIM. For example in [18], the authors used a patch size of 16 and embedding dimension 768, which gave an embedding to patch dimension ratio: $768/(16 \cdot 16 \cdot 3) = 1$, compared to ours: $1024/(16) = 64$. This design largely expands the representation dimension of fMRI data, significantly boosting the information capacity of the fMRI representations. This design is justified by both considering the dimension gap between fMRI and natural images, as well as the hypothesis of sparse coding in the visual encoding process.

Following [62], we adopt an asymmetric architecture where the decoder is much smaller than the encoder. Before feeding patch embeddings to the encoder, a random portion is masked. We used a large mask ratio similar to the mask ratio used in MIM due to the similarity in information density between fMRI data and images. We additionally embed mask tokens and include positional embeddings along with the patch encodings at the end of the encoder and transform them into the decoder's embedding space via a linear projector. On the other hand, our decoder aims to recover the masked patches with the voxel value as the prediction target.

To train the data-hungry model like the ViT, we also applied random sparsification (RS) for data augmentation, where 20% of voxels in each fMRI were randomly selected and set to zero.

Hyperparameters used in the SC-MBM pre-training stage are listed in Tab. D.2. All other unlisted parameters are set to their defaults. The SC-MBM pre-training is performed on 8 RTX3090ti GPUs until the model converges. Examples of masked brain prediction are shown in Fig. D.6.

parameter	value	parameter	value	parameter	value	parameter	value
patch size	16	encoder depth	24	decoder embed dim	512	clip gradient	0.8
embedding dim	1024	encoder heads	16	max learning rate	2.5e-4	weight decay	0.05
mask ratio	0.75	decoder depth	8	warm-up epochs	40	batch size	500
mlp ratio	1.0	decoder heads	16	max epochs	500	optimizer	AdamW [30]

Table D.2. Hyperparameters used in the Full model for SC-MBM Pre-training.

D.3. DC-LDM Finetuning

The finetuning is performed by jointly optimizing the fMRI encoder and cross-attention heads in the LDM using the training set. Specifically, for an fMRI-image pair, the image will be encoded into the latent space via a VQ encoder, which will be subsequently used as an objective to train the fMRI encoder and cross-attention heads. In the forward pass, fMRI data is passed through the encoder, producing a patchified enlarged representation. Then this representation is projected into an intermediate space with a

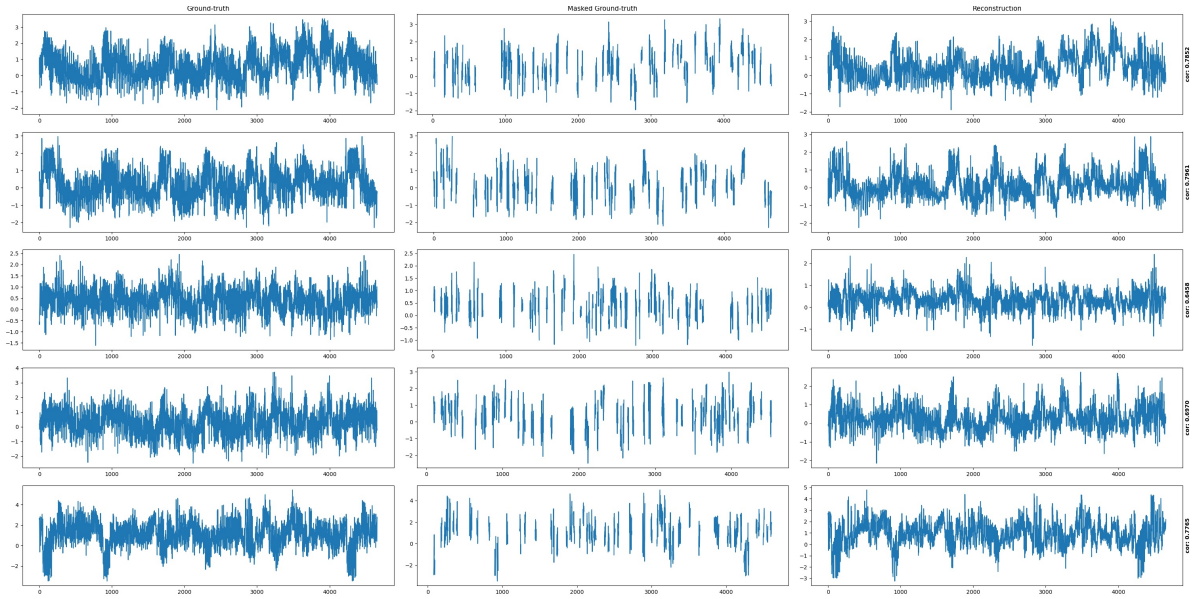


Figure D.6. Examples of masked brain prediction. First column: original fMRI data (Visual Cortex) flattened; Second column: masked fMRI; Third column: data recovered from SC-MBM decoder. Mask ratio: 0.75. The correlations between the original and recovered fMRI are also shown.

channel size of M. This intermediate representation will be used as the key and value in cross-attention modules in the UNet and will also be added to the time embedding used in the UNet. The UNet tries to denoise a Gaussian noise with the fMRI representation as a condition, mimicking the reverse transitions through a Markov Chain. L2 loss is used in training. During the training, only the fMRI encoder and cross-attention modules in the LDM are optimized. Other parts are kept intact.

Operating in the image latent space, the computations needed for DC-LDM finetuning are small. All finetunings in our experiments are performed with a single RTX3090ti GPU for 500 epochs. The detailed hyperparameters are shown in Tab. D.3. All other unlisted parameters are set to their defaults. Please see [41] for the detailed model architecture of the LDM.

parameter	value	parameter	value	parameter	value	parameter	value
batch size	5	diffusion steps	1000	image latent dim	$64\times64\times3$	learning rate	5.3e-5
image resolution	$256\times256\times3$	optimizer	AdamW	pre-trained type	Label-to-Image	M	77

Table D.3. Hyperparameters used in the Full model for DC-LDM Finetuning.

E. Other Ablation Studies

Patch Sizes In [10], an image is divided into sixteen 16×16 patches which can be considered 16 words. Analogous to the fMRI data, the more words are used to describe the data, the higher accuracy of the resulting representation will be. Therefore, smaller patches will lead to better results if the number of voxels remains unchanged. This claim is justified by Tab. E.4. A continuous decrease in accuracy can be observed when the patch size is increased from 16 to 64. However, the minimal patch size applicable is constrained by available memory, as the number of patches increases drastically with smaller patch size.

Encoder Depth The fMRI encoder depth is set to 24 in our Full model similar to the ViT-Large [10]. However, different depths lead to a different number of parameters and encoding capabilities. Usually, a deeper model is appreciated, but it comes with the need for more training samples as well. Therefore, considering the limited data, we explore whether a smaller model would have better results. To maintain an asymmetrical architecture, the depth of the SC-MBM decoder is kept at half of the encoder's depth. A deeper fMRI encoder (as deep as 24 transformer blocks) gives the best result as shown in Tab. E.4.

Mask Stragtey In [18], different masking strategies are tested for images, and the authors conclude that random masking is the best strategy for images. We explore in our ablation if it is the case for fMRI learning. For images, there are different strategies such as center masking and grid masking due to the geometric correlations among pixels in an image. For fMRI data, brain activities

are reflected by the connectivity among groups of voxels (functional networks). Seven networks in the visual cortex are used in our study (*i.e.* V1-V4, FFA, PPA, and LOC), in which the V1, the primary visual cortex, consists of the most voxels and is the first stage of visual processing. Therefore we design a focus masking strategy similar to the center masking in images. In the center masking, pixels at the center of an image will be masked the most because the center of an image usually contains the richest information. Learning to recover the center potentially is beneficial to learning the underlying semantics of an image. Similar to the center masking, our focus masking in fMRI masks more patches in the V1 region than in other regions. However, in our experiments, the focus masking does not outperform the random masking strategy as shown in Tab. E.4.

Pretext Tasks As discussed, since fMRI voxels are correlated reflecting the underlying brain activities, masked modeling is a suitable learner for fMRI representations. With SC-MBM as a pretext task, self-supervised learning can be performed in a large unpaired fMRI dataset. On the other hand, considering a small part of paired fMRI are available in the training set. Therefore, it is intriguing if we can use the paired information in fMRI for the pre-training as well. So we include the image feature as another pretext task together with the masked modeling to guide the context learning. Specifically, the training set will be divided into two parts: the part with paired images; and the part without paired images. To construct a mini-batch, we randomly sample fMRI from these two parts. In this design, another decoder is added to decode image features. The image features extracted from the second layer of a pre-trained VGG will be used as a target for this decoder. During training, the image feature reconstruction loss will be added to the MBM loss with a regularization term. However, adding the image guidance does not outperform the MBM only pre-training as shown in Tab. E.4.

Unequal Length Handle Due to individual variability, even in the same dataset, the voxel numbers of individuals are different. We need to handle this unequal length to include different subjects in the pre-training. The two most intuitive ways are considered: pad to the maximum length with a constant; cut to the minimum length. Besides padding with a constant, we pad the data in a wrap-around manner. From Tab. E.4 we can see that cutting the data gives the worst performance and wrap-around padding gives the best performance.

Crop Ratio In the finetuning, images are randomly center-cropped for augmentations. We tested different crop ratios, *i.e.* from 0 to 0.4. It is found that a crop ratio of 0.2 gives the best performance. Random cropping is an efficient augmentation in our task because the subjects' perceptions may be focused on different parts of the figure, even though they were instructed to fixate at the center of the image.

Patch size	16	32	64
Acc (%)	23.9	18.2	16.4

(a) Patch Size

Encoder depth	24	8	2
Acc (%)	23.9	14.8	13.6

(b) fMRI Encoder Depth

Strategy	random	focus
Acc (%)	23.9	16.3

(c) Mask Strategy

Task	SC-MBM	SC-MBM+image
Acc (%)	23.9	16.1

(d) Pretext Tasks

Strategy	wrap	constant	cut
Acc (%)	23.9	19.6	14.8

(e) Unequal Length Handle

Ratio	0	0.1	0.2	0.4
Acc (%)	14.9	17.9	23.9	15.2

(f) Crop Ratio

Table E.4. Other Design Ablations. The 1000-trial 50-way top-1 semantic classification accuracy is reported. All ablations are pre-trained for 500 epochs and then finetuned on GOD for another 500 epochs. Settings used in the Full model are colored in gray.

F. Extra Notes on Sparse-Coded Masked Brain Modeling

In our design, we use a large embedding-size-to-patch-size ratio to increase the information capacity of fMRI representations, which mimics the sparse coding mechanism underlying the encoding procedure of the visual cortex. Here, we provide a formal definition of the information capacity and explain the connection with the sparse coding mechanism.

Definition 1 *(Data Representation) For a piece of data given by a one-dimensional code vector $x \in \mathbb{R}^L$, let f be an injective function that maps x from the data domain to a representation domain, namely $f(x) = y$, where $y \in \mathbb{R}^{\tilde{L}}$ is a representation of x.*

Definition 2 *(Information Capacity) For a random variable X, the Shannon entropy of X is upper bounded by its cardinality, which is given by $H(X) \leq \log(|\mathcal{X}|)$. We define $\log(|\mathcal{X}|)$ as the information capacity of random variable X.*

The inequality in Definition 2 can be easily proved with Jensen's inequality regardless of the distribution of X. Obviously, for a representation Y, if the dimension of Y is larger, the representation space will be larger. Hence, Y will have a larger information capacity. To measure the change of information capacity after the representation mapping, we can simply divide the representation dimension by the data dimension, namely, $R = \tilde{L}/L$. In the context of masked modeling, we refer to R as the embedding-size-to-patch-size ratio. The

essence of sparse coding is to use sets of over-complete bases to efficiently represent data [25]. Analogous to this over-completeness, we use a representation space that is much larger than the data space, namely higher R, to learn the representations of the fMRI. Data locality is included in the representations by dividing the fMRI time series into patches and transforming patches into embeddings.

G. Pixel-level metrics

We also performed the pixel-level metrics (MSE & LPIPS) for additional evaluation (Table below). Semantic-oriented methods, Ozelik [33] and our approach outperformed the others in semantic metrics but not in pixel-level metrics. Pixel and semantic-level decodings recover visual stimuli from two perspectives, where the **trade-off between fidelity and meaningfulness** needs to be considered. In this work, we **prioritize the recovery of visual semantics** in fMRI, which is crucial for understanding the complex mechanism of human perception.

Method	MSE↓	LPIPS↓
Ours	101	0.69
Ozelik [33]	102	0.69
Gaziv [16]	**99**	**0.68**
Beliy [2]	105	0.81

Table G.5. Comparison of Pixel-level Metrics Using MSE and LPIPS Benchmarks.

H. 2-way and 5-way metrics

We also performed 2-way and 5-way metrics for additional evaluation (Table below). Our approach outperformed the others in both these two metrics.

Method	2-way↑	5-way↑
Ours	**0.86**	**0.63**
Ozelik [33]	0.84	0.61
Gaziv [16]	0.71	0.39
Beliy [2]	0.56	0.24

Table H.6. Comparison of 2-way and 5-way Classification Metrics.