Seeing Beyond the Brain: Conditional Diffusion Model with Sparse Masked Modeling for Vision Decoding

Zijiao Chen[1*] Jiaxin Qing[2*] Tiange Xiang[3] Wan Lin Yue[1] Juan Helen Zhou[1†]

[1]National University of Singapore, [2]The Chinese University of Hong Kong, [3]Stanford University
https://mind-vis.github.io

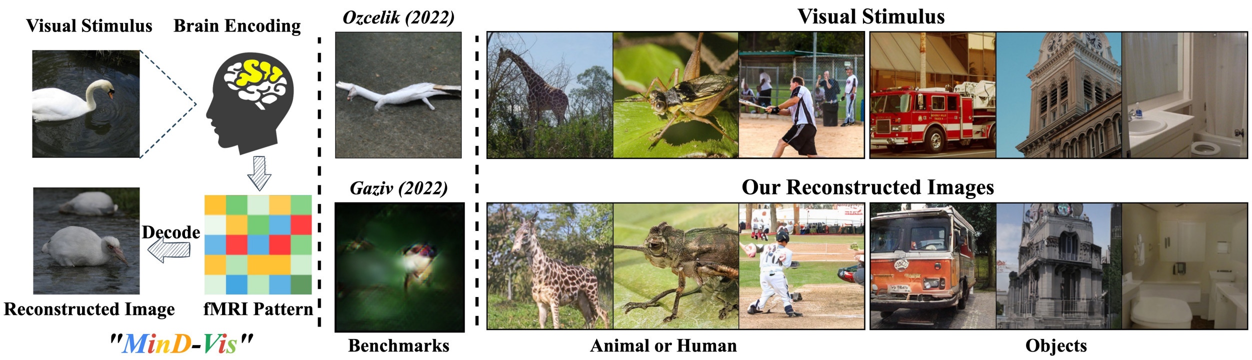

Figure 1. **Brain Decoding and Image Reconstruction**. For the first time, our proposed **MinD-Vis** is capable of decoding fMRI-based brain activities and reconstructing images with not only plausible details but also accurate semantics and image features (texture, shape, *etc*.), pushing this domain a considerable step forward. Left: Task overview. Middle: Comparison with benchmarks. Right: More reconstruction examples.

Abstract

*Decoding visual stimuli from brain recordings aims to deepen our understanding of the human visual system and build a solid foundation for bridging human and computer vision through the Brain-Computer Interface. However, reconstructing high-quality images with correct semantics from brain recordings is a challenging problem due to the complex underlying representations of brain signals and the scarcity of data annotations. In this work, we present **MinD-Vis**: Sparse **M**asked Bra**in** Modeling with Double-Conditioned Latent **D**iffusion Model for Human **Vis**ion Decoding. Firstly, we learn an effective self-supervised representation of fMRI data using mask modeling in a large latent space inspired by the sparse coding of information in the primary visual cortex. Then by augmenting a latent diffusion model with double-conditioning, we show that MinD-Vis can reconstruct highly plausible images with semantically matching details from brain recordings using very few paired annotations. We benchmarked our model qualitatively and quantitatively; the experimental results indicate that our method outperformed state-of-the-art in both semantic mapping (100-way semantic classification) and generation quality (FID) by* 66% *and* 41% *respectively. An exhaustive ablation study was also conducted to analyze our framework.*

*Equal contributions.
†Corresponding author (helen.zhou@nus.edu.sg)

1. Introduction

"What you think is what you see". Human perception and prior knowledge are deeply intertwined in one's mind [51]. Our perception of the world is determined not only by objective stimuli properties but also by our experiences, forming complex brain activities underlying our perception. Understanding these brain activities and recovering the encoded information is a key goal in cognitive neuroscience. Within this broad objective, decoding visual information is one of the challenging problems that are the focus of a large body of literature [22, 26, 34, 67].

As a non-invasive and effective method to measure brain activities indirectly, functional Magnetic Resonance Imaging (fMRI) is usually used to recover visual information, such as the image classes [21, 39]. With the help of recent deep learning models, it is intriguing if the original visual stimuli can be directly recovered from corresponding fMRI [2, 46], especially with the guidance of biological principles [43, 52]. However, due to the lack of fMRI-image pairs and useful biological guidance when decoding complex neural activity from fMRI directly, reconstructed images are usually blurry and semantically meaningless. Thus it is crucial to learn effective and biological-valid representations for fMRI so that a clear and generalizable connection between brain activities and visual stimuli can be established with a few paired annotations.

Moreover, individual variability in brain representations further complicates this problem. Individuals have unique brain activation patterns responding to the same visual stimulus (See Fig. 2). From the perspective of fMRI representation learning, a powerful brain decoding algorithm should robustly recognize features shared across the population over a background of individual variation [5, 21]. On the other hand, we should also expect decoding variances due to the variation in individual perceptions. Therefore, we aim to learn representations from a large-scale dataset with rich demographic compositions and relax the direct generation from fMRI to conditional synthesis allowing for sampling variance under the same semantic category.

Self-supervised learning with pretext tasks in large datasets is a powerful paradigm to distill the model with context knowledge. A domain-specific downstream task (*e.g.* classification) is usually adopted to finetune the pre-trained model further [36, 58], especially when the downstream dataset is small. Various pretext tasks are designed to benefit downstream tasks [23, 66]. Among these methods, Masked Signal Modeling (MSM) has achieved promising results in both vision [18, 62] and language understanding [8, 37] recently. At the same time, the probabilistic diffusion denoising model has shown its superior performance in content generation and training stability [9]. A strong generation ability is also desired in our task to decode faithful visual stimuli from various categories.

Driven by the above analysis, we propose **MinD-Vis**: Sparse Masked Brain Modeling with Double-Conditioned Latent Diffusion Model for Human Vision Decoding, a framework that exploits the power of large-scale representation learning and mimics the sparse coding of information in the brain [14], including the visual cortex [56]. Different from [18], we use a much larger representation-to-data-space ratio to boost the information capacity of learned representations. Our contributions are as follows:

- We propose Sparse-Coded Masked Brain Modeling (SC-MBM), designed under biological guidance as an effective brain feature learner for vision decoding.
- Augmenting the latent diffusion model with double conditioning (DC-LDM), we enforce stronger decoding consistency while allowing variance under the same semantics.
- Integrating the representation ability of SC-MBM with the generation ability of DC-LDM, **MinD-Vis** generates more plausible images with better preserved semantic information compared with previous methods.
- Quantitative and qualitative tests are performed on multiple datasets, including a new dataset that has not previously been used to evaluate this task.

2. Related Work

Conventional Decoding Methods Conventional methods rely on training with fMRI and corresponding hierarchical image features extracted by a pre-trained VGG [21, 46]. During testing, the predicted image features will either be used for classification or fed into a generative model like GAN [45] to reconstruct the original stimulus. Instead of directly learning the limited training pairs, [2] enabled unsupervised learning on unpaired fMRI and images with a reconfigurable autoencoder design. [16] further extended this method to images from diverse semantic categories. However, just as with conventional approaches, fMRI is used directly for training and decoding. In [31, 33], a regression model was used to extract latent fMRI representation, which was then used to finetune a pre-trained conditional bigGAN for image decoding. Mind Reader [27] encoded fMRI signals into a pre-aligned vision-language latent space and used StyleGAN2 for image generation. These methods generate more plausible and semantically meaningful images. We note that there is parallel work to ours by Takagi and Nishimoto [13], who proposed a method for image reconstruction from fMRI using Stable Diffusion. Their approach involves decoding brain activities to text descriptions and converting them to natural images using stable diffusion.

Masked Signal Modeling The power of MSM in learning representations from a large-scale dataset was first exploited in [8], which was later adapted to computer vision [18, 60, 62]. Successful applications to downstream tasks show that useful context knowledge is learned with MSM as a pretext task. In essence, MSM is a generalized denoising autoencoder that aims to recover the original data from the remaining after masking [4]. The portion of data to mask is different across data modalities, with an extremely high mask ratio (75%) usually used for visual signals [18]. In contrast, due to the disparity in information density, a low mask ratio (25%) is used in natural languages [8].

Diffusion Probabilistic Models Diffusion models [49] are emerging generative models that generate high-quality content. In its basic form [20], the diffusion model is a probabilistic model defined by a bi-directional Markov Chain of states. Two processes are transiting through the chain: *(i)* The forward diffusion process gradually adds noise to the data until it is fully destroyed to an isotropic Gaussian noise; *(ii)* The reverse process recovers the corrupted data by modeling a posterior distribution $p(x)$ at each state and eventually obtains a sample in the original data distribution [20, 49, 50]. Formally, assume a Markov Chain with a fixed length T, then the reverse conditional probability can be expressed as $q(x_{t-1}|x_t)$, where $t = 1, ..., T$ and x_t is obtained by corrupting the image x_{t-1} with Gaussian noise. After parameterization, this conditional probability can be learned by optimizing a variational lower

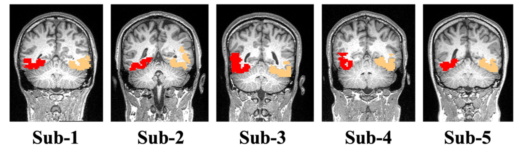

Figure 2. **Individual Differences in Regions Responding to Visual Stimuli.** Masks of the regions of interest activating during the same visual task differ in location and size across subjects. The primary visual cortex at the left (red) and the right (orange) hemisphere are shown.

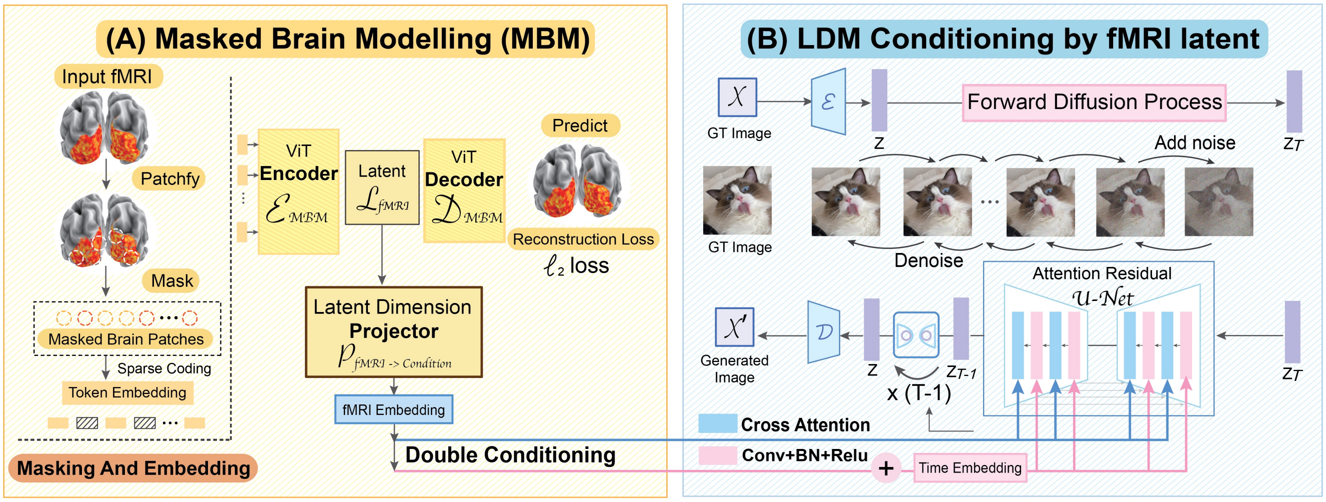

Figure 3. **MinD-Vis**. **Stage A (left):** Pre-train on fMRI with SC-MBM. We patchify, randomly mask the fMRI, and then tokenize them to large embeddings. We train an autoencoder ($\mathcal{E}_{\mathcal{MBM}}$ and $\mathcal{D}_{\mathcal{MBM}}$) to recover the masked patches. **Stage B (right):** Integration with the LDM through double conditioning. We project the fMRI latent ($\mathcal{L}_{fMRI}$) through two paths to the LDM conditioning space with a latent dimension projector ($\mathcal{P}_{fMRI\to Cond}$). One path connects directly to cross-attention heads in the LDM. Another path adds the fMRI latent to time embeddings. The LDM operates on a low-dimensional, compressed version of the original image (*i.e.* image latent), however, the original image is used in this figure for illustrations.

bound which can be simplified to the following objective [20]:

$$L_t^{simple} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon-\epsilon_\theta(x_t,t)\|_2^2\right], \tag{1}$$

where $\epsilon_\theta(x_t,t)$ is a set of denoising functions that are usually implemented as UNets [9, 41, 42]. We refer readers to [20] for detailed descriptions of the diffusion models.

Latent Diffusion Model (LDM) Apart from the conventional diffusion models that generate samples in the original data space, another category of diffusion models that generate samples in the latent feature space has been proposed [41, 48]. Operating in the latent feature space reduces the computational cost and introduces less spatial downsampling, giving better image synthesis quality. The LDM proposed in [41] consists of two components: (i) Vector Quantization (VQ) regularized [12] autoencoder that compresses images into lower-dimensional latent features and then reconstructs the images from features in the same space; (ii) UNet-based denoising model with attention modules. Incorporating attention mechanisms into the UNet allows the flexibility to condition image generation through key/value/query vectors during the Markov Chain transitions.

3. Methodology

3.1. Motivation and Overview

In this subsection, we provide a detailed analysis of the fMRI data and elaborate on the motivations of our designs.

(i) fMRI measures the brain blood-oxygen-level-dependent (BOLD) changes as 3D voxels that serve as a proxy for the underlying changes in brain activity. Neighboring voxels often have similar amplitudes, indicating spatial redundancy in fMRI [53].

(ii) fMRI data is averaged across the time during which the stimulus is presented. A region of interest (ROI) of the averaged data is usually extracted as a **1D vector** of voxels (in the visual processing hierarchy). The ROI size (voxel number) is generally smaller than the image size (pixel number). For example, [21] has about 4500 voxels (visual cortex), which is much smaller than a 256×256 RGB image. This creates a large difference in dimensionality when transforming fMRI into images.

(iii) fMRI data from different datasets may have significant domain shifts due to experimental conditions and scanner setups. Even with the same scan conditions, ROI size and location mismatch persist due to individual differences (See Fig. 2).

Driven by this analysis, we propose **MinD-Vis**, designed with two sequential stages as outlined in Fig. 3. Briefly, in **Stage A**, fMRI representations are learned by an autoencoder trained in a large fMRI dataset with masked signal modeling as a pretext task. The learned representations will be used as a condition to guide the image-generation process in the next stage. In **Stage B**, the pre-trained fMRI encoder is integrated with the LDM through cross-attention and time-step conditioning for conditional synthesis. In this stage, the encoder is jointly finetuned with cross-attention heads in the LDM using paired annotations.

3.2. Stage A: Sparse-Coded MBM (SC-MBM)

Activity in the human brain involves non-linear interactions among 86 billion neuronal cells in the brain and are thus highly complex [32, 40]. The fMRI measuring the BOLD signals is an indirect and aggregate measure of neuronal activities, which can be analyzed hierarchically with functional networks [1, 6, 59]. These functional networks comprised of voxels of fMRI data have implicit correlations with each other in response to external stimuli [54, 68]. Therefore, learning these implicit correlations by recovering masked voxels will equip the pre-trained model with a deep contextual understanding of the fMRI data.

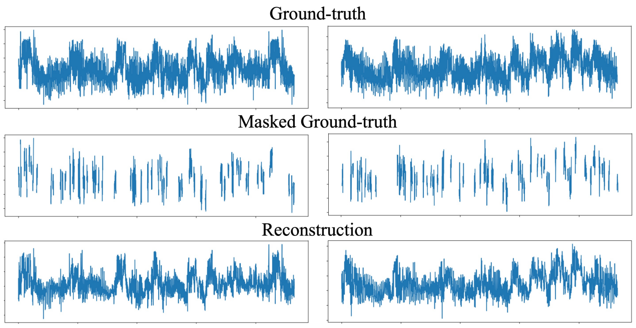

Figure 4. **Masked Brain Modeling.** Mask ratio 0.75; 4500 voxels

Following [18], we divide the vectorized voxels into patches which will be subsequently transformed into embeddings using a 1D convolutional layer with a stride equal to the patch size. The hemodynamic response and spatial smoothing functions in fMRI BOLD signal jointly cause spatial blurring, which creates spatial redundancy in fMRI data, like in natural images [11, 47]. Due to the spatial redundancy, fMRI data can still be recovered even if a large portion is masked (See Fig. 4). Thus, in the first stage of MinD-Vis, we can mask a large portion of the fMRI patches to save computations without losing the learning power of masked modeling.

Masked Image Modeling (MIM) uses the embedding-to-patch-size ratio around one [18], leading to a representation size similar to the original data size. However, we use a large embedding-to-patch-size ratio, which significantly **increases the information capacity** with a large fMRI representation space. This design also relates to the sparse coding of information in the brain, which has been proposed as a general strategy for the representation of sensory information [25].

We also adopt an asymmetric architecture as in [18]: the encoder is optimized to learn effective fMRI representations, while the decoder tries to predict the masked patches. Therefore, we make the decoder small in size, and it is discarded in Stage B as long as the pre-training converges.

Visual Encoding and Brain-Inspired Sparse Coding Here, we explain the biological basis of using SC-MBM to learn representations of visual stimuli in the brain from the perspective of visual encoding mechanisms. Theoretical and empirical studies suggest that visual stimuli are sparsely encoded in the primary visual cortex [32, 38, 56], with most natural images activating only a portion of the neurons in the visual cortex. This strategy increases information transmission efficiency and creates minimal redundancy in the brain [38]. As a result, visual information of natural scenes can be reconstructed from a small portion of data collected from the primary visual cortex via different imaging modalities, including fMRI [15, 64]. This observation is interesting for the computer vision community because the sparse coding could be an efficient way for vision encoding in computer vision as well [25, 63].

Sparse coding is an encoding strategy that in essence uses over-complete bases to represent data, where more locality is generally enforced to generate smoother representations [57, 65]. In SC-MBM, fMRI data are divided into patches to introduce locality constraints. Then each patch is encoded into a high-dimensional vector space with a size much larger than the original data space, thus creating an over-complete space for fMRI representation (See Appendix). Emulating the brain vision encoding, SC-MBM can be a biologically-valid and effective brain feature learner for fMRI decoding.

3.3. Stage B: Double-Conditioned LDM (DC-LDM)

After the large-scale context learning in Stage A, the fMRI encoder transforms fMRI data into sparsely coded representations with locality constraints. To further decode visual contents from this abstract representation and allow for sampling variance, we formulate the decoding task as a conditional synthesis problem and approach it with a pre-trained LDM.

The LDM operates on the image latent space denoted by $\mathcal{E}(x)$ where x is an image in pixel space and $\mathcal{E}(\cdot)$ is a VQ encoder. In our setting, we omit $\mathcal{E}(x)$ and use x directly to represent the latent variable of LDM for simplicity. Specifically, given the fMRI data z, we aim to learn the reverse diffusion process formulated by $q(x_{t-1}|x_t,z)$. As proposed in [41], conditional information is applied through cross-attention heads in the attention-based UNet, where $\text{CrossAttention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)$, with

$$Q = W_Q^{(i)}\varphi_i(x_t),\ K = W_K^{(i)}\tau_\theta(z),\ V = W_V^{(i)}\tau_\theta(z).$$

Here, τ_θ is the fMRI encoder with a suitable dimension projector, $\varphi_i(x_t)$ denotes intermediate values of the UNet and $W_Q^{(i)}$, $W_K^{(i)}$, $W_V^{(i)}$ are projector matrices with learnable parameters.

Diversity and consistency are two opposite objectives when sampling a conditional generative model. Sampling diversity across various modalities such as label-to-image and text-to-image is very important in many image-generation tasks. However, the fMRI-to-image transition relies more on **generation consistency**—decoded images from similar brain activities are expected to be semantically similar. Thus, a stronger conditioning mechanism is desired to ensure such generation consistency, especially for probabilistic diffusion models.

In this way, we integrate the cross-attention conditioning with another conditioning method called the *time steps conditioning* [9] to provide stronger guidance for our task. In time steps conditioning, we add $\sigma_\theta(\tau_\theta(z))$ to time step embeddings, where $\sigma_\theta(\cdot)$ is another suitable dimension projector. Time step embeddings are used in intermediate layers of the UNet, thus we have $\varphi_i(x_t) = \varphi_i(x_t, \sigma_\theta(\tau_\theta(z)))$. We further reformulate the optimization objective Eq. (1) to a *double conditioning* alternation:

$$L_t^{cond} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t, \tau(z), \sigma(\tau(z)))\|_2^2\right]. \quad (2)$$

We omit the parameterization symbol θ in $\tau(\cdot)$ and $\sigma(\cdot)$ for simplicity. Additionally, we have $\tau(z) \in \mathbb{R}^{M\times d_\tau}$ and $\sigma(\tau(z)) \in \mathbb{R}^{1\times d_t}$, where d_τ and d_t are the latent dimensions and time embedding dimension respectively, and M is a tunable parameter.

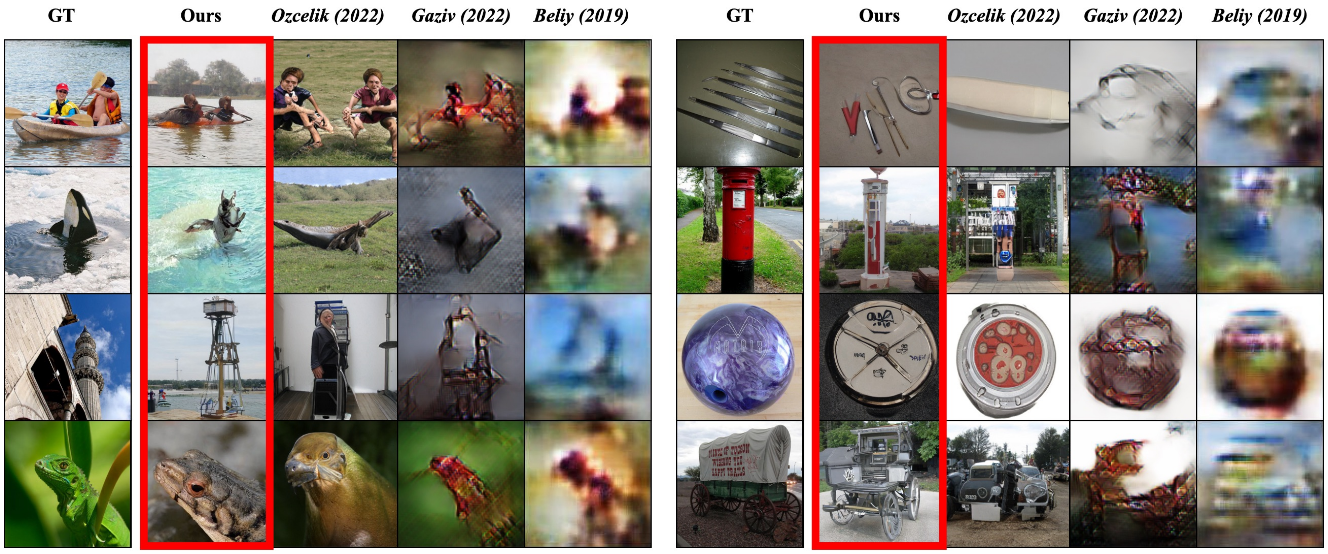

Figure 5. **Decoding Performance Comparisons on GOD Test Set.** The ground truth, images reconstructed by MinD-Vis and images reconstructed from three other methods are shown for comparison. MinD-Vis decoded the most accurate and plausible images with semantically similar details.

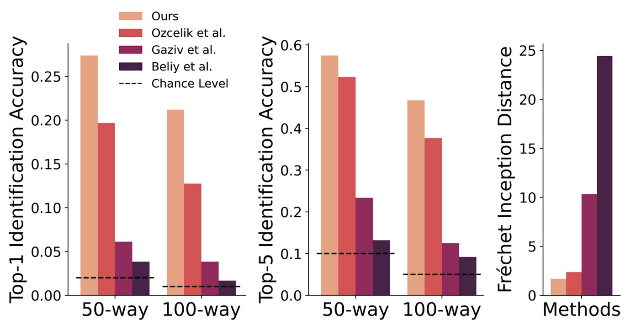

Figure 6. **Quantitative Performance Comparisons on GOD Test Set.** Performance is evaluated in terms of semantic correctness (*1000-trial n-way top-k classification accuracy; the higher the better*) and generation quality (*FID; the lower the better*).

Finetuning After the fMRI encoder is pre-trained with SC-MBM, it is integrated with a pre-trained LDM through double conditioning. Commonly, the encoder's output is averaged, or a *cls* token is appended to produce a pooled 1D feature vector for downstream tasks [8, 18]. This strategy is effective for tasks like prediction and classification, where learned knowledge is expected to be distilled, producing distinguishable features. However, pooling into a 1D vector is inappropriate for retaining fMRI representations' sparsity and information capacity. Instead, we used convolution layers to pool the encoder's output into a latent dimension of $\mathbb{R}^{M \times d_\tau}$ as described in Eq. (2).

The fMRI encoder, cross-attention heads, and projection heads are jointly optimized, while other parts are fixed. Finetuning the cross-attention heads is critical for bridging the pre-trained conditioning space and fMRI latent space. The fine-tuning is performed end-to-end with fMRI-image pairs, during which a clearer connection between the fMRI and image features will be learned through the large-capacity fMRI representations.

4. Experiments

4.1. Datasets and Implementation

Datasets Three public datasets were used in this study: Human Connectome Project (HCP) 1200 Subject Release [55]; Generic Object Decoding Dataset (GOD) [21]; and Brain, Object, Landscape Dataset (BOLD5000) [5]. Our upstream pre-training dataset comprised fMRI data from HCP and GOD. Combining these two, we obtained 136,000 fMRI segments from 340 hours of fMRI scan, which is, by far, the largest fMRI pre-training dataset in the fMRI-image decoding task. The HCP dataset is commonly used in neuroscience research, containing only fMRI data. While the GOD is an fMRI-image paired dataset designed for fMRI-based decoding. The pairs in GOD were used for finetuning in our main analysis. The GOD consists of 1250 different images from 200 distinct classes, in which 1200 images were used as the training set, and the remaining 50 images were used as the testing set. The training set and testing set have no overlapping classes. The BOLD5000 dataset was used as the validation dataset in our study. It consists of 5254 fMRI-image pairs from 4916 distinct images, 113 images of which are used for testing. This is the first time that the BOLD5000 is used for fMRI decoding tasks.

Implementation The fMRI pre-training model is similar to ViT-Large [10] with a 1D patch embedder. We used a patch size of 16, embedding dimension of 1024, encoder depth of 24, and mask ratio of 0.75 as our Full model setting with an ImageNet class-conditioned pre-trained LDM. Different parameter choices are explored in our ablation study. Unless stated otherwise, the Full model is pre-trained for 500 epochs and finetuned for another 500. Results from the best model are reported. Images are generated at a resolution of 256×256 with 250 PLMS steps [29]. See Appendix for dataset and implementation details.

4.2. Evaluation Metric

N-way Classification Accuracy Following [16], we used the n-way top-1 and top-5 accuracy classification task to evaluate the semantic correctness of our results, where for multiple trials, top-1 and top-5 classification accuracies were calculated in $n-1$ randomly selected classes plus the correct one. *Note that we did not consider the pixel-level metrics as we aimed to recover the semantically correct images in this work.*

In [16], the authors generated a typical feature for each class selected and compared the distance between the reconstructed images and the typical features. However, this metric in [16] is hard to reproduce, and the semantic classification result largely depends on how the features are computed. Therefore, we propose a more straightforward and reproducible method, where a pre-trained ImageNet1K classifier [10, 35] is used to determine the semantic correctness of generated images rather than handcrafted features. We describe this evaluation method in Algorithm D.1. Specifically, both ground-truth and generated images are input to the classifier first. Then we check for the generated image if the top-k classification in n selected classes matches the ground-truth classification. This metric does not require the ground-truth image to be from the ImageNet 1k classes. As long as semantic classification results of the ground-truth and the generated image match, it will be considered to be correct.

Fréchet inception distance (FID) The FID [19] is a commonly used metric to assess image generation quality. In our experiments, we measured the FID between ground-truth images and generated images in the testing set. Note that FID is only used as a reference in our experiments due to the limited number of images available in GOD, which may lead to an underestimated distribution.

5. Results

Our main results are based on GOD which has no overlapping classes in the training and testing set. The training and testing were performed on the same subject, as individual differences remain a barrier when decoding at the group level [2, 16, 21, 31, 33]. To compare with the literature, we report results from Subject 3 here and leave other subjects in the Appendix.

We compared our results with Ozcelik *et al*. [33], Gaziv *et al*. [16] and Beliy *et al*. [16]. Gaziv *et al*. and Beliy *et al*. used the conventional method, which decoded images with higher pixel similarity but less plausibility and semantic details. On the other hand, Ozcelik *et al*. generated more plausible and semantically meaningful images using a pre-trained GAN. Based on the best-reconstructed samples of these methods (resized to 256×256), we performed a 1000-trial, n-way top-k accuracy identification task as described in Algorithm D.1. The experiment is repeated for $n=50{,}100$ and $k=1{,}5$ in the GOD testing set.

GT **Reconstructed Samples**

Figure 7. **Generation Consistency of MinD-Vis**. Images generated by our method were consistent across different samplings trials, sharing similar low-level features and semantics.

From Fig. 6, our identification accuracy outperformed the Ozcelik *et al*. in the 50-way top-1 accuracy task by 39% and in the 100-way top-1 accuracy task by 66%, achieving a success rate of 0.274 and 0.212 respectively. The generated images from Gaziv *et al*. and Beliy *et al*. were close to the ground-truth at the pixel level but contained few semantically meaningful details, as could be observed in Fig. 5. For example, our method generated plausible details such as water and waves in the first and second images, drawings on the bowling ball, wheels of the carriage, *etc*., which were not present in the previous decoded images. The image quality is also reflected by the FID, where we achieved 1.67 with our best samples, while Ozcelik *et al*. and others achieved 2.36 or more with the best samples generated by their method. Interestingly, color mismatches are observed in some cases with the color difference well preserved. It can be explained with [3] which suggests the color category information is processed in the frontal lobes as a cognitive process, while the visual cortex only recognizes the difference in colors.

5.1. Generation Consistency

The consistency of our method was tested by decoding the same fMRI data multiple times with different random states. Five samplings with different random states were performed in the testing set for each fMRI. In the 50-way and the 100-way top-1 accuracy identification tasks, we achieved an average success rate across the five samplings of $0.2385_{\pm 0.030}$ and $0.1736_{\pm 0.029}$ respectively, which are statistically higher than the best sampling results from Ozcelik *et al*. by 21% and 35%. Regarding image quality, we achieved an average FID of $2.22_{\pm 0.3}$ across the five samplings. The standard deviations across 5 samplings indicate that the generated images will always be in the same semantic category. It can also be seen in Fig. 7 where isomorphic samplings share similar details such as shape, color, texture, and semantics, matching with the ground-truth across trials.

5.2. SC-MBM Design

This section will discuss the ablation study on the SC-MBM pre-training stage with various important parameters. Results are summarized in Tab. 1. For all experiments in this section, the 50-way, top-1 accuracy semantic identification task was performed with the best models obtained from the finetuning of 500 epochs. Average results over five samplings were reported.

Testing Without SC-MBM To show that useful representations were learned with SC-MBM, we trained two models directly using the fMRI-image pairs without the SC-MBM pre-training. The first model consisted of an untrained fMRI encoder with the same architecture as the Full model. The second model consisted of an untrained fMRI encoder with a depth of only 2. The second model was designed to have fewer parameters, making it less likely to overfit the data. All the other settings were the same. The results correspond to Model 1 and 2 in Tab. 1, where the Full model significantly outperformed the other two models without the SC-MBM pre-training, showing that the pre-training is crucial. In fact, without SC-MBM these two models even failed to generate sensible images (See Appendix).

Model	Embedding Dim	Mask Ratio	Params	Acc (%)
Full	**1024**	**0.75**	**303M**	**23.9** ±3.00
1	w/o SC-MBM + same Encoder		**303M**	2.6 ±1.39
2	w/o SC-MBM + smaller Encoder		**25M**	3.4 ±0.86
3	**32**	0.75	0.3M	5.4 ±1.50
4	**64**	0.75	1.2M	6.9 ±1.10
5	**128**	0.75	4.7M	14.8 ±1.78
6	**256**	0.75	18.9M	15.9 ±1.70
7	**512**	0.75	75.6M	17.9 ±2.58
8	**768**	0.75	170M	17.7 ±1.42
9	**1280**	0.75	472M	15.5 ±3.83
10	1024	**0.35**	303M	19.6 ±3.40
11	1024	**0.45**	303M	20.0 ±1.89
12	1024	**0.55**	303M	18.1 ±2.87
13	1024	**0.65**	303M	21.7 ±3.61
14	1024	**0.85**	303M	16.1 ±1.00

† $p < 0.0001$ (purple); $p < 0.01$ (pink); $p < 0.05$ (yellow); $p > 0.05$ (green)

Table 1. **SC-MBM Ablation Results**. Params: trainable parameters in the fMRI encoder; Cell colors reflect statistical significance differences (two-sample t-test) in accuracy compared with the Full model.

Patch Embedding Dimension Boosting the fMRI representation size using a large patch embedding matches the sparse coding mechanism of underlying visual information processing in the brain. Moreover, using a large patch embedding increases the information capacity of the representation. But larger embedding means more training parameters leading to a more data-hungry model. To balance this tradeoff, we tested SC-MBM models with different patch embedding dimensions ranging from 32 to 1280 (Model 3-9 in Tab. 1). We found that the accuracy generally increased as patch dimension increased, and accuracy peaked at 23.9% with 1024 patch embedding dimensions (full model), after which accuracy decreased as patch dimensions increased further.

Mask Ratios We used a high mask ratio in SC-MBM due to high spatial redundancy in fMRI data. In Tab. 1 Model 10-14, we show that a high mask ratio does not impair the decoding performance initially, with the highest average accuracy achieved with a relatively high mask ratio of 0.75. Importantly, using a high mask ratio significantly reduces memory consumption since the encoder only operates over unmasked patches. This is an important consideration for fMRI as SC-MBM is more memory-intensive than MIM due to the higher embedding-to-patch-size ratio.

5.3. DC-LDM Finetuning Design

This section will discuss the ablation study on the DC-LDM finetuning designs from three perspectives: conditioning methods, optimization designs, and pre-trained LDMs. Here, all ablations used the same pre-trained fMRI encoder as the Full model. Only important parameters in the finetuning stage were varied. The 1000-trial, 50-way, top-1 semantic identification test was performed. The results are summarized in Tab. 2, where five different samplings were averaged for each condition.

Model	Condition	Finetune	Pre-trained LDM	Acc (%)
Full	**C + T**	**E + A**	**Label2Image**	**23.9** ±3.00
1	**C only**	E + A	Label2Image	15.6 ±0.69
2	C+T	**E only**	Label2Image	13.76 ±2.60
3	C+T	E + A	**Text2Image**	13.42 ±3.00
4	C+T	E + A	**Layout2Image**	15.99 ±3.00

† $p < 0.0001$ (purple); $p < 0.01$ (pink); $p < 0.05$ (yellow); $p > 0.05$ (green)

Table 2. **DC-LDM Ablation Results**. 1: cross-attention condition only; 2: optimizing fMRI encoders only; 3: LDM pre-trained on text conditions (LAION); 4: LDM pre-trained on layout conditions (OpenImages). Abbr.: C (Cross-attention condition); T (Time condition); E (Encoder); A (Cross-attention heads). Cell colors reflect statistical significance (two-sample t-test) in accuracy compared with the Full model.

Conditioning Methods Here, we showed that the double conditioning method increased the conditioning strength in Tab. 2, where using only cross-attention conditioning achieved an identification accuracy of 15.6% (Model 1), which was significantly lower than the full method.

Optimizing LDM We proposed to finetune the fMRI encoder and the cross-attention heads jointly because the LDM was pre-trained in a different conditioning space. For example, for the ImageNet class-conditioning pre-trained LDM, the cross-attention heads were pre-trained to receive the class label information. To justify this choice, we tested on a model with the fMRI encoder finetuned and the cross-attention heads untouched. As shown in Model 2 in Tab. 2, the average identification accuracy dropped to 13.7% when only the fMRI encoder was finetuned, indicating stronger semantic guidance with the double conditioning. The visual quality and correspondence to the ground-truth of the generated images also decreased significantly (See Appendix).

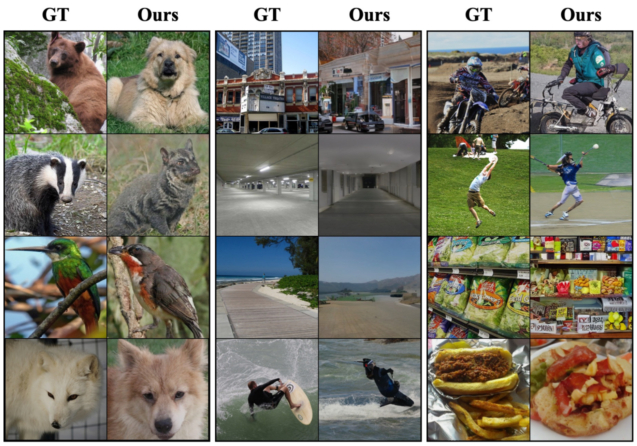

Figure 8. **Replication Dataset (BOLD5000)**. It achieved similar quantitative results as the GOD dataset. 50-way top-1 identification accuracy: 34%; FID: 1.2 (Subject 1).

Pre-trained LDM The pre-trained LDM determines the model's generative ability and the conditioning latent space to which the fMRI encoder would adapt. We considered three pre-trained LDM provided by [41], which were trained on datasets with different conditioning tasks, *i.e*. ImageNet (label conditioning), LAION (text conditioning) [44] and OpenImages (layout conditioning) [24]. As shown in Model 3-4 Tab. 2, the ImageNet pre-trained LDM (used in the full model) showed the best performance in the same decoding task. Notably, images generated by models pre-trained on LAION and OpenImages were less visually favorable and plausible (See Appendix). This result is surprising because both LAION and OpenImages contain diverse images from various categories. We attribute the main reason for their poor performance to the complexity of their conditioning latent space. With limited training pairs, the class-conditioning latent space is easier to adapt to, compared with the latent space of the text-conditioning model and the layout-conditioning model.

5.4. Replication Dataset

We validated our method on BOLD5000 using the same pre-trained fMRI encoder. Similarly, the pre-trained encoder was firstly finetuned for 20 epochs in the testing set of BOLD500 with wrap-around paddings to compensate for the unequal ROI size from the pre-training set, after which the model is further tuned with the fMRI-image training pairs in BOLD5000. All other settings were the same as the Full model. For the four subjects in BOLD5000, we achieved a 19% to 34% best accuracy in the 1000-trial, 50-way, top-1 accuracy semantic identification task (See Appendix). The generated images matched the ground-truth stimulus in both semantics and low-level features (Fig. 8). Our model accurately reconstructs images containing objects and animals, architecture, and landscapes.

Interestingly, we reconstructed similar images for some natural scenes with extra details that do not exist in the ground-truth stimulus. These extra details, for example, the river and the blue sky in Fig. 9, may reflect imagined scenery in the subject's mind when viewing the visual stimuli, which is captured in their brain activities. As reported in [21,46], features of imaginary images can also be decoded from the visual cortex.

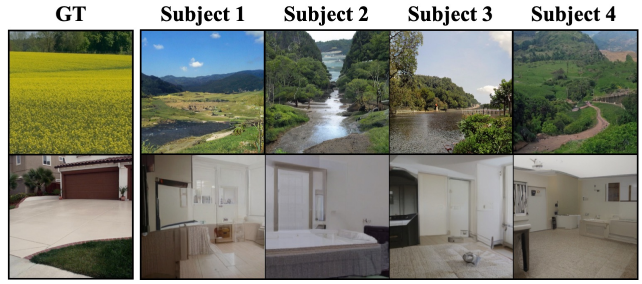

Figure 9. **Extra Features Decoded.** Imagery-related details can be decoded with our method. *e.g*. the river and blue sky were decoded with natural scenery stimulus (top row); similar interior decorating of indoor environments was decoded when a house was presented (bottom row).

To the best of our knowledge, this is the first work that performs fMRI decoding on BOLD5000. Additionally, adapting the same pre-trained model to this dataset shows that the SC-MBM pre-training indeed learns useful representations of brain recordings even when distinct domain shifts exist. These learned representations are shared and generalizable to datasets with different scanning protocols and preprocessing pipelines.

6. Discussion and Conclusion

Limitations MinD-Vis, in its current form, lacks strong pixel-level guidance and interpretation analysis, which limits its pixel-level performance (see G.5) and the biological understanding of the features learned by MBM.

Future Work Similar to all previous work, MinD-Vis focuses on individual decoding using the visual cortex only. But as a complex cognitive process, human vision may be affected by regions beyond the visual cortex. Therefore, future studies should extend to cross-subject generalization and also the incorporation of other brain regions. Additionally, the two-stage decoupling design of MinD-Vis allows us to explore the potential of emerging large-scale models and representation learning techniques in cognitive neuroscience, which is also subject to future studies.

Conclusion We proposed a two-stage framework MinD-Vis to decode visual stimuli using only a few paired fMRI-image annotations from brain recordings. In Stage A, we employ an fMRI pre-training scheme with masked modeling to learn generalizable context knowledge from a large-scale unlabeled fMRI dataset. In Stage B, we use a latent diffusion model with double conditioning to generate plausible seen images from learned fMRI representations. We validated the decoding results of MinD-Vis on multiple datasets and showed that our model generates more plausible and semantically similar images compared to previous methods, pushing the state-of-the-art a considerable step forward.

References

[1] Teddy J Akiki and Chadi G Abdallah. Determining the hierarchical architecture of the human brain using subject-level clustering of functional networks. *Scientific reports*, 9(1):1–15, 2019.

[2] Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. *Advances in Neural Information Processing Systems*, 32, 2019.

[3] Chris M Bird, Samuel C Berens, Aidan J Horner, and Anna Franklin. Categorical encoding of color in the brain. *Proceedings of the National Academy of Sciences*, 111(12):4590–4595, 2014.

[4] Shuhao Cao, Peng Xu, and David A Clifton. How to understand masked autoencoders. *arXiv preprint arXiv:2202.03670*, 2022.

[5] 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):1–18, 2019.

[6] Dietmar Cordes, Vic Haughton, John D Carew, Konstantinos Arfanakis, and Ken Maravilla. Hierarchical clustering to measure connectivity in fmri resting-state data. *Magnetic resonance imaging*, 20(4):305–317, 2002.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR'09*, pages 248–255. Ieee, 2009.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[11] Stephen A Engel, Gary H Glover, and Brian A Wandell. Retinotopic organization in human visual cortex and the spatial precision of functional mri. *Cerebral cortex (New York, NY: 1991)*, 7(2):181–192, 1997.

[12] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proc. CVPR'21*, pages 12873–12883, 2021.

[13] Tao Fang, Yu Qi, and Gang Pan. Reconstructing perceptive images from brain activity by shape-semantic gan. *Advances in Neural Information Processing Systems*, 33:13038–13048, 2020.

[14] Peter Foldiak. Sparse coding in the primate cortex. *The handbook of brain theory and neural networks*, 2003.

[15] Jeremy Freeman, Corey M Ziemba, David J Heeger, Eero P Simoncelli, and J Anthony Movshon. A functional and perceptual signature of the second visual area in primates. *Nature neuroscience*, 16(7):974–981, 2013.

[16] Guy Gaziv, Roman Beliy, Niv Granot, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. Self-supervised natural image reconstruction and large-scale semantic classification from brain activity. *NeuroImage*, 254:119121, 2022.

[17] Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proc. CVPR'22*, pages 16000–16009, 2022.

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conf. on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[21] Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):1–15, 2017.

[22] Asim Iqbal, Phil Dong, Christopher M Kim, and Heeun Jang. Decoding neural responses in mouse visual cortex through a deep neural network. In *IJCNN'19*, pages 1–7. IEEE, 2019.

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[24] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *IJCV*, 128(7):1956–1981, 2020.

[25] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. CVPR'06*, volume 2, pages 2169–2178. IEEE, 2006.

[26] Wenyi Li, Shengjie Zheng, Yufan Liao, Rongqi Hong, Weiliang Chen, Chenggnag He, and Xiaojian Li. The brain-inspired decoder for natural visual image reconstruction. *arXiv preprint arXiv:2207.08591*, 2022.

[27] Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind reader: Reconstructing complex images from brain activities. *arXiv preprint arXiv:2210.01769*, 2022.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV'14*, pages 740–755. Springer, 2014.

[29] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022.

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[31] Milad Mozafari, Leila Reddy, and Rufin VanRullen. Reconstructing natural scenes from fmri patterns using bigbigan. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2020.

[32] Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[33] 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 *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2022.

[34] Nikhil Parthasarathy, Eleanor Batty, William Falcon, Thomas Rutten, Mohit Rajpal, EJ Chichilnisky, and Liam Paninski. Neural networks for efficient bayesian decoding of natural images from retinal neurons. *Advances in Neural Information Processing Systems*, 30, 2017.

[35] PyTorch. Models and pre-trained weights, 2022. https://pytorch.org/vision/stable/models.html, Accessed September 26, 2022.

[36] 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 *ICML'21*, pages 8748–8763. PMLR, 2021.

[37] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[38] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999.

[39] Pieter R Roelfsema, Damiaan Denys, and P Christiaan Klink. Mind reading and writing: The future of neurotechnology. *Trends in cognitive sciences*, 22(7):598–610, 2018.

[40] Edmund T Rolls and Martin J Tovee. Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *Journal of neurophysiology*, 73(2):713–726, 1995.

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR'22*, pages 10684–10695, 2022.

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conf. on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[43] Fredrik Sandin and Mattias Nilsson. Synaptic delays for insect-inspired temporal feature detection in dynamic neuromorphic processors. *Frontiers in Neuroscience*, 14:150, 2020.

[44] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[45] Guohua Shen, Kshitij Dwivedi, Kei Majima, Tomoyasu Horikawa, and Yukiyasu Kamitani. End-to-end deep image reconstruction from human brain activity. *Frontiers in Computational Neuroscience*, page 21, 2019.

[46] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1):e1006633, 2019.

[47] Amir Shmuel, Essa Yacoub, Denis Chaimow, Nikos K Logothetis, and Kamil Ugurbil. Spatio-temporal point-spread function of fmri signal in human gray matter at 7 tesla. *Neuroimage*, 35(2):539–552, 2007.

[48] Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. D2c: Diffusion-decoding models for few-shot conditional generation. *Advances in Neural Information Processing Systems*, 34:12533–12548, 2021.

[49] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis R. Bach and David M. Blei, editors, *Proc. of the 32nd International Conf. on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conf. Proc.*, pages 2256–2265. JMLR.org, 2015.

[50] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

[51] Chunyue Teng and Dwight J Kravitz. Visual working memory directly alters perception. *Nature human behaviour*, 3(8):827–836, 2019.

[52] Fabian David Tschopp, Michael B Reiser, and Srinivas C Turaga. A connectome based hexagonal lattice convolutional network model of the drosophila visual system. *arXiv preprint arXiv:1806.04793*, 2018.

[53] Kamil Uğurbil, Junqian Xu, Edward J Auerbach, Steen Moeller, An T Vu, Julio M Duarte-Carvajalino, Christophe Lenglet, Xiaoping Wu, Sebastian Schmitter, Pierre Francois Van de Moortele, et al. Pushing spatial and temporal resolution for functional and diffusion mri in the human connectome project. *Neuroimage*, 80:80–104, 2013.

[54] Martijn P Van Den Heuvel and Hilleke E Hulshoff Pol. Exploring the brain network: a review on resting-state fmri functional connectivity. *European neuropsychopharmacology*, 20(8):519–534, 2010.

[55] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

[56] William E Vinje and Jack L Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276, 2000.

[57] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *Proc. CVPR'10*, pages 3360–3367. IEEE, 2010.

[58] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022.

[59] Yanlu Wang and Tie-Qiang Li. Analysis of whole-brain resting-state fmri data using hierarchical clustering approach. *PloS one*, 8(10):e76315, 2013.

[60] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proc. CVPR'22*, pages 14668–14678, 2022.

[61] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proc. CVPR'10*, pages 3485–3492. IEEE, 2010.

[62] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proc. CVPR'22*, pages 9653–9663, 2022.

[63] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Proc. CVPR'09*, pages 1794–1801. IEEE, 2009.

[64] Takashi Yoshida and Kenichi Ohki. Natural images are reliably represented by sparse and variable populations of neurons in visual cortex. *Nature communications*, 11(1):1–19, 2020.

[65] Kai Yu, Tong Zhang, and Yihong Gong. Nonlinear learning using local coordinate coding. *Advances in neural information processing systems*, 22, 2009.

[66] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[67] Yijun Zhang, Tong Bu, Jiyuan Zhang, Shiming Tang, Zhaofei Yu, Jian K Liu, and Tiejun Huang. Decoding pixel-level image features from two-photon calcium signals of macaque visual cortex. *Neural Computation*, 34(6):1369–1397, 2022.

[68] Juan Zhou, Michael D Greicius, Efstathios D Gennatas, Matthew E Growdon, Jung Y Jang, Gil D Rabinovici, Joel H Kramer, Michael Weiner, Bruce L Miller, and William W Seeley. Divergent network connectivity changes in behavioural variant frontotemporal dementia and alzheimer's disease. *Brain*, 133(5):1352–1367, 2010.