Controllable Mind Visual Diffusion Model

Bohan Zeng[1]*, Shanglin Li[1]*, Xuhui Liu[1], Sicheng Gao[1]
Xiaolong Jiang[3], Xu Tang[3], Yao Hu[3], Jianzhuang Liu[4], Baochang Zhang[1,2,5]†
[1]Institute of Artificial Intelligence, Hangzhou Research Institute, Beihang University, China
[2]Nanchang Institute of Technology, Nanchang, China
[3]Xiaohongshu Inc
[4]Shenzhen Institute of Advanced Technology, Shenzhen, China
[5] Zhongguancun Laboratory, Beijing, China
{bohanzeng, shanglin, bczhang}@buaa.edu.cn

Abstract

Brain signal visualization has emerged as an active research area, serving as a critical interface between the human visual system and computer vision models. Diffusion-based methods have recently shown promise in analyzing functional magnetic resonance imaging (fMRI) data, including the reconstruction of high-quality images consistent with original visual stimuli. Nonetheless, it remains a critical challenge to effectively harness the semantic and silhouette information extracted from brain signals. In this paper, we propose a novel approach, termed as Controllable Mind Visual Diffusion Model (CMVDM). Specifically, CMVDM first extracts semantic and silhouette information from fMRI data using attribute alignment and assistant networks. Then, a control model is introduced in conjunction with a residual block to fully exploit the extracted information for image synthesis, generating high-quality images that closely resemble the original visual stimuli in both semantic content and silhouette characteristics. Through extensive experimentation, we demonstrate that CMVDM outperforms existing state-of-the-art methods both qualitatively and quantitatively. Our code is available at https://github.com/zengbohan0217/CMVDM.

Introduction

Understanding the cognitive processes that occur in the human brain when observing visual stimuli (e.g., natural images) has long been a primary focus for neuroscientists. Both objective visual stimuli and subjective cognitive activities can elicit the transmission of intricate neural signals in the visual cortex of the brain, thus laying the foundation for higher-order cognitive and decision-making processes. With the advancement of techniques such as functional magnetic resonance imaging (fMRI), it has become possible to capture real-time brain activity signals with greater accuracy and finer granularity, thereby accelerating the progress of neuroscientific research. Deciphering and reconstructing from these intricate signals remain a great challenge to both cognitive neuroscience and downstream applications like Brain-Computer Interfaces (BCI) (Nicolas-Alonso and Gomez-Gil 2012; Milekovic et al. 2018).

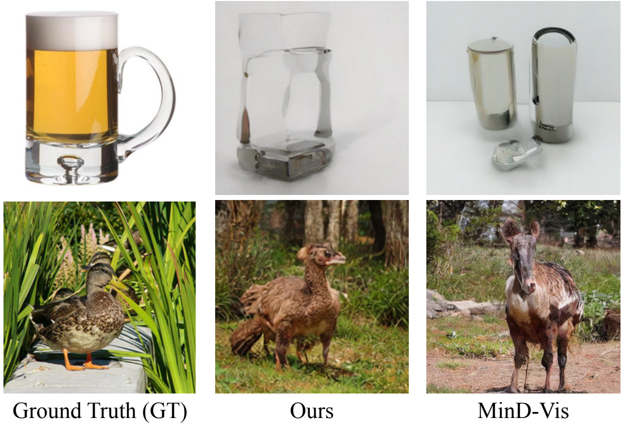

Figure 1: Illustration of synthesis results. A recent method MinD-Vis (Chen et al. 2023) can generate photo-realistic results, but they cannot well match the visual stimuli in terms of semantics and silhouette. Our method can generate better results more consistent with the GT visual stimuli.

Early attempts (Van Gerven et al. 2010; Damarla and Just 2013; Horikawa and Kamitani 2017; Akamatsu et al. 2020) at analyzing brain activity on visual tasks mainly focus on matching human subjects' brain activity with observed natural images, or reconstructing visual patterns of simple geometric shapes (Miyawaki et al. 2008; Schoenmakers et al. 2013; Van Gerven, De Lange, and Heskes 2010). These explorations demonstrate the feasibility of deriving semantic information for perceived images from brain signals, yet they have poor generalization to unseen semantic categories or complicated reconstruction tasks.

Recent studies (Beliy et al. 2019; Gaziv et al. 2022; Ozcelik et al. 2022; Chen et al. 2023; Takagi and Nishimoto 2023) have made significant progress in reconstructing visual stimuli from brain signals. (Beliy et al. 2019; Gaziv et al. 2022) can generate images that are similar in shape to the original visual stimuli, but the images suffer from severe distortion and blur issues. (Ozcelik et al. 2022; Chen

*These authors contributed equally.

†Corresponding author.

et al. 2023; Takagi and Nishimoto 2023) have employed commonly used generative models, such as Generative Adversarial Networks (GAN) or diffusion models, to generate high-quality RGB images that maintain semantic consistency with the original visual stimuli conditioned on corresponding fMRI signals. However, such methods struggle with positional inconsistency, as shown in Fig. 1. In general, existing methods have not effectively utilized the semantic and spatial features inherent in fMRI signals.

In this paper, we present a Controllable Mind Visual Diffusion Model (CMVDM) that enables the mind diffusion model with a control network to leverage the extracted faithful semantic and silhouette information for high-fidelity human vision reconstruction. Specifically, we first finetune a pretrained latent diffusion model (LDM) with a semantic alignment loss and pretrain a silhouette extractor to estimate accurate semantic and silhouette information of the fMRI data. Taking inspiration from ControlNet, we then introduce a control network, which takes the silhouette information as a condition, into the pretrained LDM to guide the diffusion process to generate desired images that match the original visual stimuli in terms of both semantic and silhouette information. Fig. 1 shows two examples where CMVDM outperforms the previous state-of-the-art approach, MinD-Vis.

In summary, the main contributions of this paper are as follows:

- We propose a novel Controllable Mind Visual Diffusion Model (CMVDM) that leverages both semantic and spatial visual patterns in brain activity to reconstruct photorealistic images. A control network is utilized to enable effective manipulation over the positions of generated objects or scenes in the reconstructed images, providing a much better structural similarity to the original visual stimuli.
- We design two extractors to extract semantic and silhouette attributes to provide accurate information for generating images that closely resemble the visual stimuli. Besides, we build a residual module to provide information beyond semantics and silhouette.
- We conduct comprehensive experiments on two datasets to evaluate the performance of our method. It achieves state-of-the-art qualitative and quantitative results compared to existing methods, demonstrating the efficacy of CMVDM for decoding high-quality and controllable images from fMRI signals.

Related Work

Diffusion Probabilistic Models. Diffusion models (DMs) were initially introduced by (Sohl-Dickstein et al. 2015) as a novel generative model that gradually denoises images corrupted by Gaussian noise to produce samples. Recent advances in DMs have demonstrated their superior performance in image synthesis, with notable models including (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Dhariwal and Nichol 2021; Vahdat, Kreis, and Kautz 2021; Rombach et al. 2022; Peebles and Xie 2022). DDGAN (Xiao, Kreis, and Vahdat 2022) is a model that reduces the number of sampling steps by directly predicting the ground truth in each timestep. DMs have also achieved state-of-the-art performance in other synthesis tasks, such as text-to-image generation with GLIDE (Nichol et al. 2021), speech synthesis with (Kong et al. 2020; Liu et al. 2021), and super-resolution with (Li et al. 2022a; Saharia et al. 2022; Gao et al. 2023). In addition, DMs have been applied to text-to-3D synthesis in (Poole et al. 2022; Lin et al. 2022), and other 3D object syntheses in (Anciukevičius et al. 2022; Li et al. 2022b; Luo and Hu 2021). Furthermore, DMs have found applications in video synthesis (Ho et al. 2022b,a), semantic segmentation (Baranchuk et al. 2021), text-to-motion generation (Tevet et al. 2022), face animation (Zeng et al. 2023), and object detection (Chen et al. 2022). (Kulikov et al. 2022; Wang et al. 2022) are models that generate diverse results by learning the internal patch distribution from a single image. ControlNet employs a control network on a pretrained text-conditioned LDM for controllable image synthesis. Overall, DMs have shown promising results and have been widely adopted in various synthesis tasks.

Neural Decoding of Visual Stimuli. Neural decoding of visual stimuli has been a topic of growing interest in recent years. Numerous studies have explored the possibility of using machine learning algorithms to decode visual information from patterns of neural activity in the human brain. For instance, (Naselaris et al. 2009) demonstrates that it is possible to reconstruct natural images from fMRI data using a linear decoder. Similarly, (Kay et al. 2008) shows that the orientation of gratings from patterns of activity in the early visual cortex can be decoded using a support vector machine. More recent studies have built on these findings by exploring more complex visual stimuli, such as natural scenes (Nishimoto et al. 2011) and faces (Kriegeskorte et al. 2007), and by developing more sophisticated machine learning algorithms, such as deep neural networks (Yamins et al. 2014). To enable decoding of novel scenarios, some works use an identification-based approach (Horikawa and Kamitani 2017; Akamatsu et al. 2020; Kay et al. 2008), where they model the relationship between brain activity and visual semantic knowledge such as image features extracted by a CNN (Horikawa and Kamitani 2017; Akamatsu et al. 2020). These studies provide valuable insights into the interpretation of human brain signals in the visual cortex, which can help the development of more effective decoding algorithms for a wide range of neuroimaging applications, such as Brain-Computer Interfaces. However, these methods require a large amount of paired stimuli-responses data that is hard to obtain. Therefore, decoding novel image categories accurately remains a challenge.

fMRI-to-Image Reconstruction With the remarkable advancements in generative models, recent studies have focused on the reconstruction of images from human brain activity. These studies employ various approaches, such as building an encoder-decoder structure to align image features with corresponding fMRI data, as demonstrated by (Beliy et al. 2019) and (Gaziv et al. 2022). To further enhance the quality of image reconstruction, researchers have turned to more sophisticated techniques, including genera-

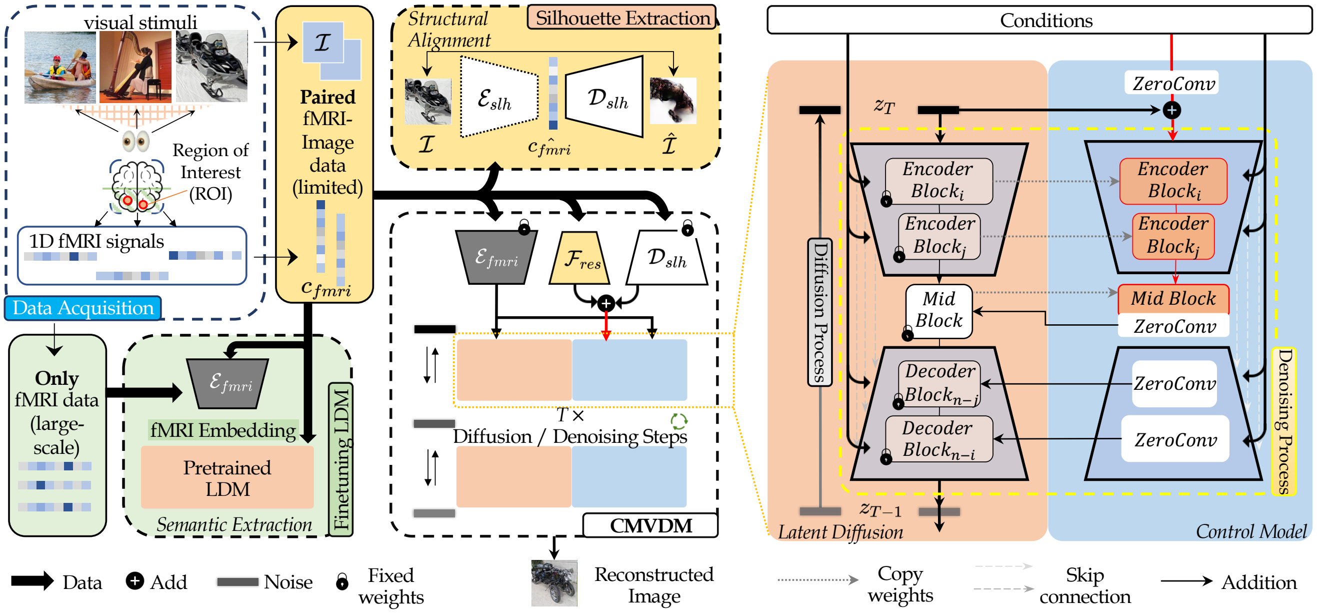

Figure 2: Overview of our proposed method. Initially, we train $\mathcal{E}_{fmri}$ and $\mathcal{D}_{slh}$ in the "Finetuning LDM" and "Silhouette Extraction" parts, respectively. Subsequently, we utilize $\mathcal{E}_{fmri}$, $\mathcal{D}_{slh}$, and $\mathcal{F}_{res}$ to extract semantic, silhouette, and supplement information from fMRI signals as conditions. Finally, we integrate the control network with the LDM to generate high-fidelity and controllable results tailored to the aforementioned conditions.

tive adversarial networks (GAN) (Ozcelik et al. 2022) and diffusion models (Takagi and Nishimoto 2023; Chen et al. 2023). These methods have shown promise in achieving more plausible image reconstruction. Nonetheless, the approaches described above have limitations in terms of image reconstruction quality and localization accuracy, resulting in unreliable reconstruction outcomes and inadequate utilization of the deep semantic and shallow positional information inherent in fMRI signals.

Method

In this section, we describe the CMVDM model, which combines attribute extractors and a control model to produce precise and controllable outcomes from fMRI signals. Fig. 2 illustrates the architecture of CMVDM.

Problem Statement and Overview of CMVDM

Let the paired {fMRI, image} dataset $\Omega = \{(c_{fmri,i}, \mathcal{I}_i)\}_{i=1}^{n}$, where $c_{fmri,i} \in \mathbb{R}^{1\times N}$ and $\mathcal{I}_i \in \mathbb{R}^{H\times W\times 3}$. The fMRI data is extracted as a 1D signal from the region of interest (ROI) on the visual cortex averaged across the time during which the visual stimuli are presented. N denotes the number of voxels of the extracted signal. We adopt the pretrained image encoder of the LDM (Rombach et al. 2022) to encode the observed image $\mathcal{I}$ into the latent code z. Our CMVDM aims to learn an estimation of the data distribution $p(z|c_{fmri})$ through a Markov chain with T timesteps. Following (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Rombach et al. 2022), we define the fixed forward Markov diffusion process q as:

$$\begin{aligned} q\left(z_{1:T} \mid z_0\right) &= \prod_{t=1}^{T} q\left(z_t \mid z_{t-1}\right), \\ q\left(z_t \mid z_{t-1}\right) &= \mathcal{N}\left(z_t \mid \sqrt{1-\beta_t} z_{t-1}, \beta_t \mathbf{I}\right), \end{aligned} \tag{1}$$

where z_0 denotes the latent code of an image. This Markov diffusion process propagates by adding Gaussian noise, with variances $\beta_t \in (0,1)$ in T iterations. Given z_0, the distribution of z_t can be represented by:

$$q\left(z_t \mid z_0\right) = \mathcal{N}\left(z_t \mid \sqrt{\gamma_t} z_0, (1-\gamma_t)\mathbf{I}\right), \tag{2}$$

where $\gamma_t = \prod_{i=1}^{t}(1-\beta_i)$. In the inference process, CMVDM learns the conditional distributions $p_\theta(z_{t-1}|z_t, c_{fmri})$ and conducts a reverse Markov process from Gaussian noise $z_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to a target latent code z_0 as:

$$\begin{aligned} p_\theta\left(z_{0:T} \mid c_{fmri}\right) &= p\left(z_T\right) \prod_{t=1}^{T} p_\theta\left(z_{t-1} \mid z_t, c_{fmri}\right), \\ p\left(z_T\right) &= \mathcal{N}\left(z_T \mid \mathbf{0}, \mathbf{I}\right), \\ p_\theta\left(z_{t-1} \mid z_t, c_{fmri}\right) &= \mathcal{N}\left(z_{t-1} \mid \mu_\theta\left(c_{fmri}, z_t, t\right), \sigma_t^2 \mathbf{I}\right), \end{aligned} \tag{3}$$

where $\sigma_t = \frac{1-\gamma_{t-1}}{1-\gamma_t}\beta_t$. The pretrained image decoder of the LDM (Rombach et al. 2022) turns the final latent code to an image.

Furthermore, we extract the attributes and control the generated results. Firstly, we extract the semantic and silhouette information by utilizing the fMRI encoder $\mathcal{E}_{fmri}$ and the silhouette estimating network $\mathcal{D}_{slh}$, respectively. This step enables us to accurately decouple the fMRI information c_{fmri}.

Subsequently, we utilize the control model $\mathcal{F}_{ctrl}$ to generate high-quality images that match the visual stimuli in terms of both semantic and silhouette information. $\mathcal{F}_{ctrl}$ is able to leverage the extracted information to produce better results. Besides, the residual module $\mathcal{F}_{res}$ is designed to provide information beyond semantics and silhouette.

Finetuning of the Pretrained LDM

Before extracting the silhouette information and controlling the generated results, we need to finetune the pretrained LDM (Rombach et al. 2022) to enable it to generate consistent images and extract the semantic information based on the input fMRI signals. Following MinD-Vis, we employ the fMRI encoder $\mathcal{E}_{fmri}$ pretrained on the HCP dataset (Van Essen et al. 2013) to encode the brain activity signals to the fMRI embeddings. Besides, we use the pretrained LDM to generate output images. By optimizing the fMRI encoder $\mathcal{E}_{fmri}$ and the cross-attention layers in the LDM, while freezing the other blocks during the finetuning process, we can obtain reliable consistent generated results. The finetuning loss is defined as follows:

$$\mathcal{L}_f = \mathbb{E}_{z_0,t,c_{fmri},\epsilon\sim\mathcal{N}(0,1)}[||\epsilon - \epsilon_\theta(z_t,t,\mathcal{E}_{fmri}(c_{fmri}))||_2^2], \tag{4}$$

where ϵ_θ is the denoising network of the LDM. In this way, the LDM can ensure the consistency of the generated results. Let $c_{ctx} = \mathcal{E}_{fmri}(c_{fmri})$ be the semantic information extracted from the fMRI signals. Due to the lack of direct semantic supervision, $\mathcal{E}_{fmri}$ may be insufficient for providing enough semantic information. Therefore, we design a noval alignment loss $\mathcal{L}_{align}$ to further enhance the semantic information c_{ctx}:

$$\mathcal{L}_{align} = e^{-\text{cosine}(f_{img},\text{MLP}(c_{ctx}))}, \tag{5}$$

where $\text{cosine}(\cdot,\cdot)$ denotes the cosine similarity, f_{img} is the image feature extracted by the CLIP image encoder (Radford et al. 2021), and MLP represents a trainable multilayer perceptron. After this training stage, the LDM can make the generated images consistent with the fMRI signals. Nonetheless, due to the absence of explicit positional condition guidance, it is still a challenge for the LDM to generate silhouette-matched results. In the next two sections, we will describe how to extract silhouette information from the fMRI signals and control the final results.

Silhouette Extraction

In this section, we aim to extract silhouette information from fMRI signals. (Gaziv et al. 2022) uses a combination of self-supervised and supervised learning to reconstruct images similar to visual stimuli.

Despite the low fidelity of the image generation quality, their generated results demonstrate a notable ability to accurately replicate the silhouette of the visual stimuli (see Fig. 3). Based on this, we devise a silhouette estimation network that is capable of providing rough positional guidance for CMVDM.

Our silhouette estimation network consists of two components: an encoder $\mathcal{E}_{slh}$ and a decoder $\mathcal{D}_{slh}$. The encoder $\mathcal{E}_{slh}$ projects the input images to the fMRI signal space, while the decoder $\mathcal{D}_{slh}$ performs the inverse transformation.

Let $c_{fmri,i}$ be the ground truth (GT) fMRI signal, $\mathcal{I}_i$ be the corresponding GT image, and $c_{f\hat{mri},i} = \mathcal{E}_{slh}(\mathcal{I}_i)$ be the estimated fMRI signal. We define the encoder training loss $\mathcal{L}_e$ by a combination of the Mean Square Error (MSE) loss and cosine similarity:

$$\begin{aligned}\mathcal{L}_e = \frac{1}{|\Omega|}\sum_{i=1}^{|\Omega|}[&\alpha_1 \cdot \|c_{fmri,i} - c_{f\hat{mri},i}\|^2 \\ &+ \alpha_2 \cdot (1 - \text{cosine}(c_{fmri,i}, c_{f\hat{mri},i}))],\end{aligned} \tag{6}$$

where $\alpha_{i\in\{1,2\}}$ are the hyperparameters set empirically to $\alpha_1 = 1$ and $\alpha_2 = 0.3$.

After completing the training of $\mathcal{E}_{slh}$, we fix its parameters and train the reverse process for the decoder $\mathcal{D}_{slh}$. Due to the limited availability of paired {fMRI, image} data, mapping fMRI signals to images is challenging. Inspired by (Gaziv et al. 2022), we utilize semi-supervised training to extract intricate silhouette information. The self-supervised process can be simply represented as: $\hat{\phi_i} = \mathcal{D}_{slh}(\mathcal{E}_{slh}(\phi_i))$, where $\phi_i \in \Phi$ denotes the image from ImageNet (without corresponding fMRI data) (Deng et al. 2009), and $\hat{\phi_i}$ denotes the reconstructed image. By minimizing the disparity between ϕ_i and $\hat{\phi_i}$, the self-supervised process helps $\mathcal{E}_{slh}$ and $\mathcal{D}_{slh}$ to learn more generalized image representation. We employ the Structural Similarity (SSIM) loss besides the Mean Absolute Error (MAE) loss to penalize the spatial distances between the reconstructed images and the GT images. The two losses are:

$$\mathcal{L}_{mae} = \underbrace{\frac{1}{|\Omega|}\sum_{i=1}^{|\Omega|}|\hat{\mathcal{I}_i} - \mathcal{I}_i|}_{supervised} + \underbrace{\frac{1}{|\Phi|}\sum_{i=1}^{|\Phi|}|\hat{\phi_i} - \phi_i|}_{self-supervised}, \tag{7}$$

$$\mathcal{L}_{ssim} = 1 - \frac{(2\mu_{\mathcal{I}}\mu_{\hat{\mathcal{I}}} + C_1)(2\sigma_{\mathcal{I}\hat{\mathcal{I}}} + C_2)}{(\mu_{\mathcal{I}}^2 + \mu_{\hat{\mathcal{I}}}^2 + C_1)(\sigma_{\mathcal{I}}^2 + \sigma_{\hat{\mathcal{I}}}^2 + C_2)}, \tag{8}$$

where $\mu_{\hat{\mathcal{I}}}$, $\mu_{\mathcal{I}}$, $\sigma_{\hat{\mathcal{I}}}$, and $\sigma_{\mathcal{I}}$ represent the mean and std values of the reconstructed images $\hat{\mathcal{I}}$ and GT images $\mathcal{I}$, C_1 and C_2 are constants to stabilize the calculation.

The decoder loss $\mathcal{L}_d$ is defined as the combination of the two losses:

$$\mathcal{L}_d = \mathcal{L}_{mae} + \mathcal{L}_{ssim}. \tag{9}$$

After training, $\mathcal{D}_{slh}$ is able to generate images $\hat{\mathcal{I}}$ from c_{fmri} that provide positional guidance for CMVDM. To avoid confusion, we'll refer to $\hat{\mathcal{I}}$ as c_{slh} in the following section.

Training of Control Model

After obtaining the enhanced semantic information $c_{ctx} = \mathcal{E}_{fmri}(c_{fmri})$ and the reliable silhouette information $c_{slh} = \mathcal{D}_{slh}(c_{fmri})$ from c_{fmri}, we use them to control the generated results as shown in Fig. 2. Inspired by ControlNet, we design a control model to control the overall composition of the generated images. Specifically, we freeze all the parameters in the denoising network ϵ_θ and clone the U-Net encoder of ϵ_θ into the trainable $\mathcal{F}_{ctrl}(\cdot;\Theta_c)$ with a set of parameters

Method	GOD			BOLD5000		
	Acc (%)	PCC	SSIM	Acc(%)	PCC	SSIM
Beliy (2019)	4.288	0.48285	0.51795	/	/	/
Gaziv (2022)	9.128	0.68326	**0.64857**	/	/	/
IC-GAN (2022)	29.386	0.44857	0.54489	/	/	/
MinD-Vis (2023)	26.644	0.53159	0.52669	25.918	0.54486	0.52379
CMVDM (Ours)	**30.112**	**0.76751**	0.63167	**27.791**	**0.55691**	**0.53459**

Table 1: Quantitative comparison with four state-of-the-art (SOTA) methods. Bold results denote the best results and underlined results denote the second-best results.

Θ_c (the red blocks of control model in Fig. 2). The inputs of $\mathcal{F}_{ctrl}$ include z_t, c_{ctx}, and the silhouette feature c_{slh}. The combined condition code $x'_{c,t}$ can be formulated as:

$$x'_{c,t} = \mathcal{Z}(\mathcal{F}_{ctrl}(z_t + \mathcal{Z}(c_{slh}), c_{ctx}; \Theta_c)), \tag{10}$$

where $\mathcal{Z}(\cdot)$ denotes the zero convolution operation (Zhang and Agrawala 2023). Furthermore, in order to compensate for the fMRI data loss during attribute extraction, we utilize a trainable residual block denoted as $\mathcal{F}_{res}$. This block is trained in conjunction with $\mathcal{F}_{ctrl}$. The final combined condition code $x_{c,t}$ is represented as:

$$\begin{aligned} x_{c,t} =& \mathcal{Z}(\mathcal{F}_{ctrl}(z_t + \\ & \mathcal{Z}(c_{slh} + \mathcal{Z}(\mathcal{F}_{res}(c_{fmri}))), c_{ctx}; \Theta_c)). \end{aligned} \tag{11}$$

Then the output features $x_{c,t}$ of the control model are added to the U-Net decoder features of the frozen ϵ_θ, as shown in Fig. 2.

Finally, we use the following loss $\mathcal{L}_{ctrl}$ to supervise the training of the control model and $\mathcal{F}_{res}$ in our CMVDM:

$$\begin{aligned} &\mathcal{L}_{ctrl} = \\ &\quad \mathbb{E}_{z_0,t,c_{fmri},\epsilon\sim\mathcal{N}(0,1)}[||\epsilon - \epsilon_\theta(z_t, t, c_{ctx}, x_{c,t})||_2^2]. \end{aligned} \tag{12}$$

Note that with their losses, the control model training, the pretrained LDM finetuning, and the $\mathcal{D}_{slh}$ training are independent. In our framework, we separately pretrained $\mathcal{E}_{fmri}$ and $\mathcal{D}_{slh}$ and froze their weights to jointly train $\mathcal{F}_{res}$ and $\mathcal{F}_{ctrl}$ (as depicted in Fig 2).

Experiments

Datasets and Implementation

Datasets. In this study, we employ two public datasets with paired fMRI signals and images: Generic Object Decoding (GOD) dataset (Horikawa and Kamitani 2017), and Brain, Object, Landscape Dataset (BOLD5000) (Chang et al. 2019). The GOD dataset is a well-known and extensively researched collection of fMRI-based brain signal decoding data. It comprises 1250 distinct images belonging to 200 different categories, with 50 images designated for testing. The BOLD5000 dataset is a rich resource for studying the neural representation of visual stimuli, as it contains diverse images from natural and artificial domains. The images are drawn from three existing datasets: SUN (Xiao et al. 2010), COCO (Lin et al. 2014), and ImageNet (Deng et al. 2009), which contain images of various categories of objects and animals. BOLD5000 was acquired from four subjects who underwent fMRI scanning while viewing 5,254 images in 15 sessions. The fMRI data were preprocessed and aligned to a common anatomical space, resulting in 4803 fMRI-image pairs for training and 113 for testing. The dataset provides a unique opportunity to investigate how the human brain encodes visual information across different levels of abstraction and complexity. Additionally, we use the large-scale fMRI data from Human Connectome Project (HCP) (Van Essen et al. 2013) in an unsupervised manner to pretrain the fMRI encoder $\mathcal{E}_{fmri}$ in our method, which aims to fully extract the features of fMRI signals.

Training Details. We adopt 1 A100-SXM4-40GB GPU for the training of $\mathcal{E}_{fmri}$ and the control model, and 1 V100-SXM2-32GB GPU for $\mathcal{D}_{slh}$ training. Both $\mathcal{E}_{fmri}$ and the control model are trained by the AdamW (Loshchilov and Hutter 2017) with $\beta = (0.9, 0.999)$ and $eps = 1e - 8$ for 500 epochs. $\mathcal{D}_{slh}$ is optimized using Adam (Kingma and Ba 2015) with a learning rate of $5e - 3$ and $\beta = (0.5, 0.99)$ for 150 epochs.

Evaluation Metrics

N-way Classification Accuracy (Acc). Following (Gaziv et al. 2022; Chen et al. 2023), we employ the n-way top-1 classification task to evaluate the semantic correctness of the generated results, where multiple trials for top-1 classification accuracies are calculated in $n - 1$ randomly selected classes with the correct class. Specifically, we follow MinD-Vis and use a pretrained ImageNet-1K classifier (Dosovitskiy et al. 2020) to estimate the accuracy. Firstly, we input the generated results and the ground-truth images into the classifier, and then check whether the top-1 classification matches the correct class.

Pearson Correlation Coefficient (PCC). The Pearson correlation coefficient (PCC) measures the degree of linear association between two variables. PCC is used to measure the correlation between the pixel values of the generated results and those of the ground truth, with +1 indicating a perfect positive linear relationship and -1 indicating a perfect negative linear relationship. The larger the PCC value, the stronger the relevance between visual stimuli and generated images.

Structure Similarity Index Measure (SSIM). We adopt SSIM to evaluate the reconstruction faithfulness of the generated results. As analyzed in (Wang et al. 2004), the structural similarity of two images is measured by three different factors, brightness, contrast, and structure, where the mean

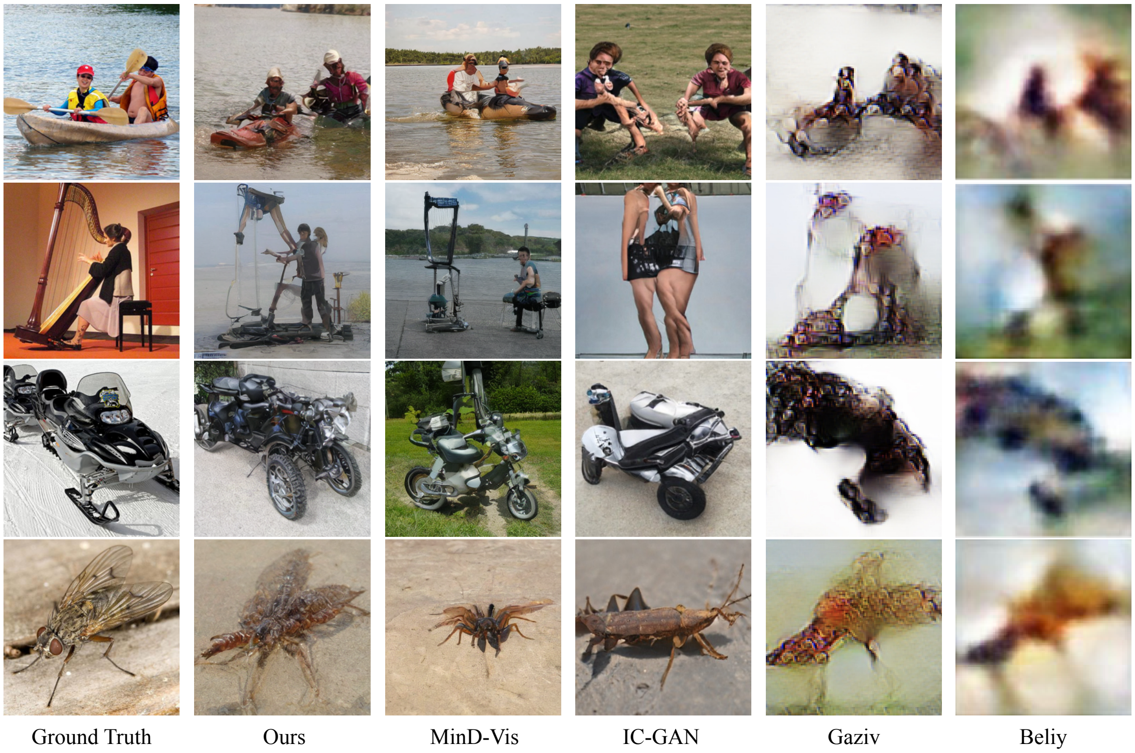

Figure 3: Comparison with four SOTA methods on the GOD dataset.

is used as the estimate of brightness, the standard deviation as the estimate of contrast, and the covariance as the measurement of structural similarity.

Comparison with State-of-the-Art Methods

Methods. We compare our CMVDM with four state-of-the-art (SOTA) methods: MinD-Vis, IC-GANs (Ozcelik et al. 2022), Gaziv (Gaziv et al. 2022), and Beliy (Beliy et al. 2019). We use their official pretrained models for all the comparisons, which are trained on the GOD dataset. For the BOLD5000 dataset, we only compare with the official pretrained MinD-Vis model, because other works (Beliy et al. 2019; Gaziv et al. 2022; Ozcelik et al. 2022) did not conduct experiments and release their models on BOLD5000.

Results on the GOD Dataset. We conduct a quantitative comparison between CMVDM and the four SOTA models using the testing dataset of GOD. Table 1 summarizes the results, revealing that CMVDM overall outperforms the other methods significantly. Compared to MinD-Vis and IC-GAN, both of which yield good results, CMVDM outperforms them significantly in terms of SSIM. This indicates that the images generated by CMVDM exhibit a higher degree of resemblance to the visual stimuli in terms of object silhouette and image structure. Additionally, Fig. 3 demonstrates that CMVDM generates visually impressive images with semantic and structural information closest to the visual stimuli. Gaziv achieves remarkable results in terms of SSIM, but their accuracy reported in Table 1 and visual results presented in Fig. 3 demonstrate that their method is not capable of generating high-fidelity images.

Method	Acc (%)	PCC	SSIM
MinD-Vis	26.644	0.53159	0.54489
MinD-Vis+$\mathcal{L}_{align}$	27.362	0.56686	0.52628
MinD-Vis+Control Model	28.438	0.75730	**0.63404**
CMVDM	**30.112**	**0.76751**	0.63167

Table 2: Ablation study of CMVDM's components.

Results on the BOLD5000 Dataset. We conduct a comparative analysis between our CMVDM and the most recent method MinD-Vis using the testing dataset of BOLD5000. As depicted in Table 1, it is evident that CMVDM consistently outperforms MinD-Vis across all evaluation metrics. Additionally, Fig. 4 provides visualizations of some results from both methods, clearly demonstrating that CMVDM generates more realistic outcomes that are more similar to the GT visual stimuli. Notably, the BOLD5000 dataset, being more complex than the GOD dataset, further validates the effectiveness of our proposed method.

Ablation Study

We further conduct experiments on the GOD dataset to analyze the effectiveness of each module of CMVDM. Specifically, we employ MinD-Vis as the baseline and design two comparison models: (1) adding the semantic align loss $\mathcal{L}_{align}$ to MinD-Vis, (2) adding the control model to MinD-Vis. The results, presented in Table 2, demonstrate the ef-

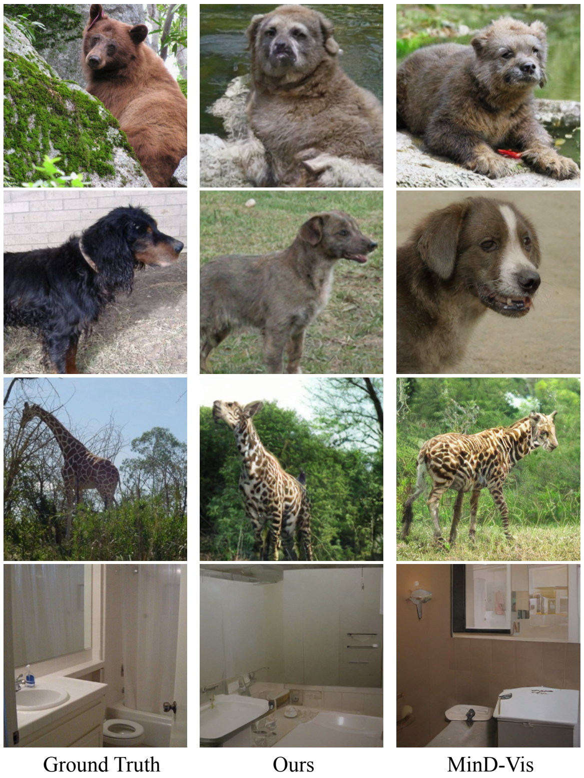

Figure 4: Comparison with MinD-Vis on the BOLD5000 dataset.

ficacy of both $\mathcal{L}_{align}$ and the control model within our CMVDM. MinD-Vis with $\mathcal{L}_{align}$ yields improved results in terms of ACC and PCC, which illustrate that $\mathcal{L}_{align}$ can improve the capability of CMVDM to obtain semantic information. Furthermore, MinD-Vis+Control Model outperforms MinD-Vis+$\mathcal{L}_{align}$ in each metric, particularly in SSIM, indicating that the silhouette contains valuable semantic information that is used in the control model.

Consistency Analysis

To further verify the generative stability of CMVDM, we conduct an analysis to compare the consistency of two diffusion-based methods. As shown in Fig. 5, we sample three images reconstructed by CMVDM and MinD-Vis from the same fMRI signal. The images generated by CMVDM demonstrate a high degree of consistency to GT images both semantically and structurally. However, the results generated by MinD-Vis are capable of reproducing GT images semantically but are not consistent in structure.

Further Analysis

The impact of using the residual module $\mathcal{F}_{res}$ in our CMVDM is significant on the BOLD5000 dataset, as demonstrated in Table 3. However, the effect of $\mathcal{F}_{res}$ on the GOD dataset is not as pronounced. We believe that there are two reasons for this discrepancy. Firstly, the voxels of a single fMRI signal provided by the BOLD5000 dataset are much less than that provided by the GOD dataset, making it more challenging to extract valid semantic and silhouette information from BOLD5000. Therefore, $\mathcal{F}_{res}$ is necessary to compensate for the information gap. Secondly, compared to GOD, BOLD5000 has more diverse images, including scenes that are not present in GOD. The semantic judgment and position alignment of the images in BOLD5000 are more complex than those in GOD. Therefore, we utilize $\mathcal{F}_{res}$ to provide more information and improve the reconstruction performance.

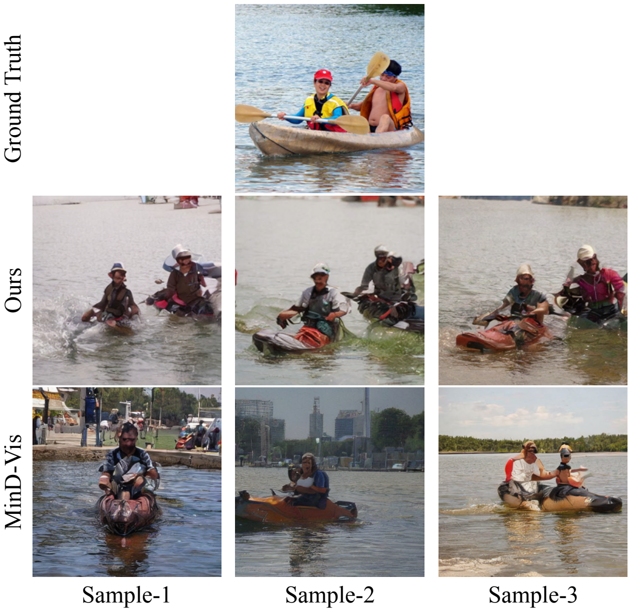

Figure 5: Consistency analysis of the generated results.

Dataset	Method	Acc(%)	PCC	SSIM
BOLD5000	w/o $\mathcal{F}_{res}$	25.393	0.54184	0.52951
	w $\mathcal{F}_{res}$	**27.791**	**0.55691**	**0.53459**
GOD	w/o $\mathcal{F}_{res}$	29.436	0.75837	**0.63894**
	w $\mathcal{F}_{res}$	**30.112**	**0.76751**	0.63167

Table 3: Quantitative analysis of the residual block in CMVDM.

Conclusion

In this paper, we propose a Controllable Mind Visual Diffusion Model (CMVDM) for decoding fMRI signals. Firstly, we simultaneously train a semantic encoder and perform finetuning on a pretrained latent diffusion model to generate semantically consistent images from fMRI signals. Secondly, we incorporate a silhouette extractor to derive reliable position information from fMRI signals. Furthermore, we design a control model to ensure CMVDM generates semantically-consistent and spatially-aligned images with the original visual stimuli. Extensive experiments demonstrate that our approach achieves state-of-the-art performance in generating high-quality images from fMRI signals.

Acknowledgements

This research was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LD24F020007, Beijing Natural Science Foundation L223024, National Natural Science Foundation of China under Grant 62076016. The work was also supported by the National Key Research and Development Program of China (Grant No. 2023YFC3300029) and "One Thousand Plan" projects in Jiangxi Province Jxsg2023102268 and ATR key laboratory grant 220402.

References

Akamatsu, Y.; Harakawa, R.; Ogawa, T.; and Haseyama, M. 2020. Brain decoding of viewed image categories via semi-supervised multi-view Bayesian generative model. *IEEE Transactions on Signal Processing*.

Anciukevičius, T.; Xu, Z.; Fisher, M.; Henderson, P.; Bilen, H.; Mitra, N. J.; and Guerrero, P. 2022. RenderDiffusion: Image Diffusion for 3D Reconstruction, Inpainting and Generation. *arXiv:2211.09869*.

Baranchuk, D.; Voynov, A.; Rubachev, I.; Khrulkov, V.; and Babenko, A. 2021. Label-efficient semantic segmentation with diffusion models. In *ICLR*.

Beliy, R.; Gaziv, G.; Hoogi, A.; Strappini, F.; Golan, T.; and Irani, M. 2019. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fMRI. In *NeurIPS*.

Chang, N.; Pyles, J. A.; Marcus, A.; Gupta, A.; Tarr, M. J.; and Aminoff, E. M. 2019. BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Scientific data*.

Chen, S.; Sun, P.; Song, Y.; and Luo, P. 2022. Diffusiondet: Diffusion model for object detection. *arXiv:2211.09788*.

Chen, Z.; Qing, J.; Xiang, T.; Yue, W. L.; and Zhou, J. H. 2023. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *CVPR*.

Damarla, S. R.; and Just, M. A. 2013. Decoding the representation of numerical values from brain activation patterns. *Human Brain Mapping*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. In *NeurIPS*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gao, S.; Liu, X.; Zeng, B.; Xu, S.; Li, Y.; Luo, X.; Liu, J.; Zhen, X.; and Zhang, B. 2023. Implicit Diffusion Models for Continuous Super-Resolution. *arXiv preprint arXiv:2303.16491*.

Gaziv, G.; Beliy, R.; Granot, N.; Hoogi, A.; Strappini, F.; Golan, T.; and Irani, M. 2022. Self-supervised natural image reconstruction and large-scale semantic classification from brain activity. *NeuroImage*.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022a. Imagen video: High definition video generation with diffusion models. *arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *NeurIPS*.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022b. Video diffusion models. In *NeurIPS*.

Horikawa, T.; and Kamitani, Y. 2017. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*.

Kay, K. N.; Naselaris, T.; Prenger, R. J.; and Gallant, J. L. 2008. Identifying natural images from human brain activity. *Nature*.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Kong, Z.; Ping, W.; Huang, J.; Zhao, K.; and Catanzaro, B. 2020. Diffwave: A versatile diffusion model for audio synthesis. *arXiv:2009.09761*.

Kriegeskorte, N.; Formisano, E.; Sorger, B.; and Goebel, R. 2007. Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences*, 104(51): 20600–20605.

Kulikov, V.; Yadin, S.; Kleiner, M.; and Michaeli, T. 2022. SinDDM: A Single Image Denoising Diffusion Model. *arXiv:2211.16582*.

Li, H.; Yang, Y.; Chang, M.; Chen, S.; Feng, H.; Xu, Z.; Li, Q.; and Chen, Y. 2022a. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*.

Li, M.; Duan, Y.; Zhou, J.; and Lu, J. 2022b. Diffusion-SDF: Text-to-Shape via Voxelized Diffusion. *arXiv:2212.03293*.

Lin, C.-H.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.-Y.; and Lin, T.-Y. 2022. Magic3D: High-Resolution Text-to-3D Content Creation. *arXiv:2211.10440*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Liu, J.; Li, C.; Ren, Y.; Chen, F.; Liu, P.; and Zhao, Z. 2021. Diffsinger: Diffusion acoustic model for singing voice synthesis. *arXiv:2105.02446*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. In *ICLR*.

Luo, S.; and Hu, W. 2021. Diffusion probabilistic models for 3d point cloud generation. In *CVPR*.

Milekovic, T.; Sarma, A. A.; Bacher, D.; Simeral, J. D.; Saab, J.; Pandarinath, C.; Sorice, B. L.; Blabe, C.; Oakley, E. M.; Tringale, K. R.; et al. 2018. Stable long-term BCI-enabled communication in ALS and locked-in syndrome using LFP signals. *Journal of Neurophysiology*, 120(7): 343–360.

Miyawaki, Y.; Uchida, H.; Yamashita, O.; Sato, M.-a.; Morito, Y.; Tanabe, H. C.; Sadato, N.; and Kamitani, Y.

2008. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5): 915–929.

Naselaris, T.; Prenger, R. J.; Kay, K. N.; Oliver, M.; and Gallant, J. L. 2009. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6): 902–915.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv:2112.10741*.

Nicolas-Alonso, L. F.; and Gomez-Gil, J. 2012. Brain computer interfaces, a review. *Sensors*, 12(2): 1211–1279.

Nishimoto, S.; Vu, A. T.; Naselaris, T.; Benjamini, Y.; Yu, B.; and Gallant, J. L. 2011. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19): 1641–1646.

Ozcelik, F.; Choksi, B.; Mozafari, M.; Reddy, L.; and VanRullen, R. 2022. Reconstruction of perceived images from fMRI patterns and semantic brain exploration using instance-conditioned GANs. In *IJCNN*.

Peebles, W.; and Xie, S. 2022. Scalable Diffusion Models with Transformers. *arXiv:2212.09748*.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv:2209.14988*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022. Image super-resolution via iterative refinement. *TPAMI*.

Schoenmakers, S.; Barth, M.; Heskes, T.; and Van Gerven, M. 2013. Linear reconstruction of perceived images from human brain activity. *NeuroImage*, 83: 951–961.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv:2010.02502*.

Takagi, Y.; and Nishimoto, S. 2023. High-resolution image reconstruction with latent diffusion models from human brain activity. In *CVPR*.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-Or, D.; and Bermano, A. H. 2022. Human motion diffusion model. *arXiv:2209.14916*.

Vahdat, A.; Kreis, K.; and Kautz, J. 2021. Score-based generative modeling in latent space. In *NeurIPS*.

Van Essen, D. C.; Smith, S. M.; Barch, D. M.; Behrens, T. E.; Yacoub, E.; Ugurbil, K.; Consortium, W.-M. H.; et al. 2013. The WU-Minn human connectome project: an overview. *NeuroImage*.

Van Gerven, M. A.; Cseke, B.; De Lange, F. P.; and Heskes, T. 2010. Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. *NeuroImage*.

Van Gerven, M. A.; De Lange, F. P.; and Heskes, T. 2010. Neural decoding with hierarchical generative models. *Neural Computation*, 22(12): 3127–3142.

Wang, W.; Bao, J.; Zhou, W.; Chen, D.; Chen, D.; Yuan, L.; and Li, H. 2022. SinDiffusion: Learning a Diffusion Model from a Single Natural Image. *arXiv:2211.12445*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *TIP*.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*.

Xiao, Z.; Kreis, K.; and Vahdat, A. 2022. Tackling the Generative Learning Trilemma with Denoising Diffusion GANs. In *ICLR*.

Yamins, D. L.; Hong, H.; Cadieu, C. F.; Solomon, E. A.; Seibert, D.; and DiCarlo, J. J. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23): 8619–8624.

Zeng, B.; Liu, X.; Gao, S.; Liu, B.; Li, H.; Liu, J.; and Zhang, B. 2023. Face Animation with an Attribute-Guided Diffusion Model. *arXiv preprint arXiv:2304.03199*.

Zhang, L.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*.