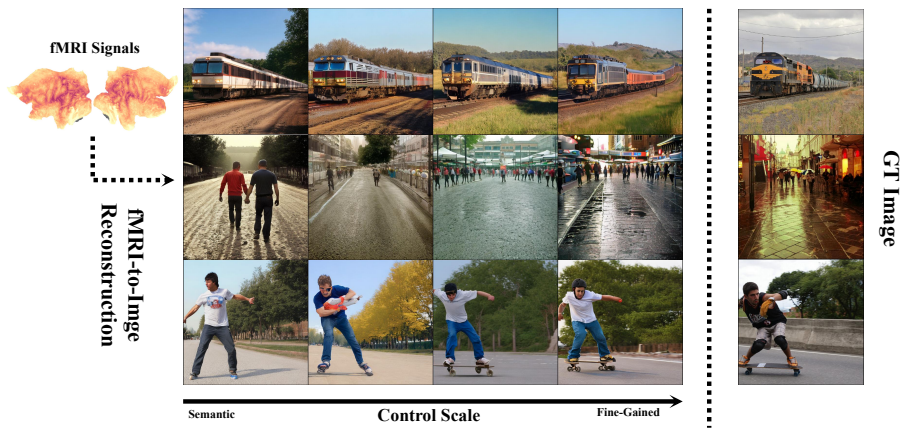# NeuroPictor: Refining fMRI-to-Image Reconstruction via Multi-individual Pretraining and Multi-level Modulation

Jingyang Huo[*], Yikai Wang[*], Yun Wang[*], Xuelin Qian, Chong Li, Yanwei Fu[†], and Jianfeng Feng

Fudan University, Shanghai, China
yanweifu@fudan.edu.cn

Fig. 1: NeuroPictor achieves precise control over decoding low-level structures from fMRI signals while preserving high-level semantics. The decoded images progress from reconstructing visual stimulus solely from high-level semantics to both high-level semantics and low-level structures as the influence increases from left to right.

Abstract. Recent fMRI-to-image approaches mainly focused on associating fMRI signals with specific conditions of pre-trained diffusion models. These approaches, while producing high-quality images, capture only a limited aspect of the complex information in fMRI signals and offer little detailed control over image creation. In contrast, this paper proposes to directly modulate the generation process of diffusion models using fMRI signals. Our approach, NeuroPictor, divides the fMRI-to-image process into three steps: i) fMRI calibrated-encoding, to tackle multi-individual pre-training for a shared latent space to minimize individual difference and enable the subsequent multi-subject training; ii) fMRI-to-image multi-subject pre-training, perceptually learning to guide diffusion model with high- and low-level conditions across different individuals; iii) fMRI-to-image single-subject refining, similar with step

---

[*] Equal contributions.
[†] Corresponding author.

ii but focus on adapting to particular individual. NeuroPictor extracts high-level semantic features from fMRI signals that characterizing the visual stimulus and incrementally fine-tunes the diffusion model with a low-level manipulation network to provide precise structural instructions. By training with about 67,000 fMRI-image pairs from various individuals, our model enjoys superior fMRI-to-image decoding capacity, particularly in the within-subject setting, as evidenced in benchmark datasets. Our code and model are available at `https://jingyanghuo.github.io/neuropictor/`.

Keywords: Neural Decoding · FMRI-to-Image · Diffusion Model

1   Introduction

Decoding visual stimuli perceived by the eyes and recorded by our brains is a fascinating task, serving as a bridge between two closely linked yet distinct fields: computer vision and visual neuroscience. In neuroscience, we rely on functional magnetic resonance imaging (fMRI), a non-invasive neuroimaging technique that detects brain activity by measuring changes in blood flow, offering valuable insights into cognitive processes and brain function [15]. Meanwhile, the generative power of the Diffusion model enables the creation of visually appealing images from prior conditions or random signals, making it intriguing to decode fMRI-based brain activities and reconstruct images, as explored in [38].

This paper explores the process of decoding fMRI signals into images. We outline the experimental pipeline in three stages: "See it. Say it. Sorted". Essentially, an individual unfamiliar with our task is presented with an image ("See it"). They then take a moment to reflect on this image ("Say it"), during which their brain activity is scanned and recorded using fMRI equipment. Finally, our algorithm uses the recorded fMRI signals to reconstruct the original image ("Sorted"). Most previous works [7,29], focus on individual-specific fMRI-to-image decoders. Unfortunately, due to the potential unique brain activation patterns each brain processes information during the "say it" phase, there is no direct empirical evidence to show that another individual's fMRI data would assist in decoding for a specific individual.

Furthermore, in the "sorted" stage, researchers often employ latent codes extracted from fMRI signals to guide image generation through the Stable Diffusion (SD) model. Nevertheless, these codes tend to prioritize high-level semantics over low-level details, as fMRI voxels are aligned with CLIP text [28] and image features [36] to capture semantic information. However, using latent codes directly for SD presents challenges due to the fundamental mismatch between the separately trained fMRI encoder/decoder for 1D signals and the SD model for 2D images. Consequently, imperfections in both the fMRI signal-to-latent and latent-to-image projections lead to inevitable information loss and error accumulation, resulting in blurry reconstructions with lacking clear structure and pose information.
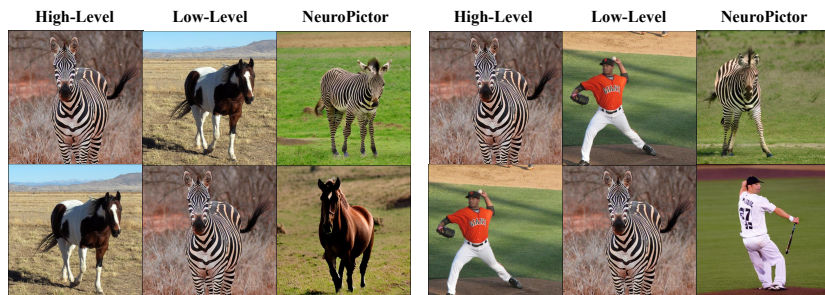
Fig. 2: NeuroPictor can swap high-level fMRI features to manipulate image semantics while maintaining structural consistency.

To address these challenges, we introduce NeuroPictor[1], a novel framework designed to refine fMRI-to-image reconstruction through multi-individual pre-training and multi-level modulation. Unlike previous approaches [7, 29] focusing on individual-specific decoders, NeuroPictor significantly improves fMRI-to-image decoding for single individuals by leveraging pretraining on multiple-individual fMRI-image pairs. Our method establishes a universal fMRI latent space to capture diverse neural signal information across subjects. Within this framework, we incorporate both low-level manipulation and high-level guiding networks as in Fig. 1. The low-level manipulation network precisely adjusts the diffusion model's features, integrating complex brain signal information and eliminating the need for partial representation with intermediate images. Meanwhile, the high-level guiding network rectifies semantic gaps between text captions and visual images. These components form a unified pipeline, enabling training the full model from fMRI signals to visual stimulus directly via the fMRI-to-image reconstruction supervision and preventing information degradation and error accumulation from isolated processing. Moreover, disentangling high-level meaning and low-level structure in fMRI signals opens up intriguing possibilities. For example, we can swap the high-level features of one fMRI signal with another, enabling precise semantic manipulation on the reconstructed image while preserving the structures, as depicted in Fig. 2.

Formally, our NeuroPictor initially learns a universal fMRI latent space across individuals through multi-individual pre-training to capture all original signal information and individual differences. We use an fMRI encoder-decoder architecture to reconstruct the fMRI signal as 2D brain activation flatmaps. Afterward, we utilize the encoder to map the fMRI signal into latent codes. We introduce a High-Level Guiding Network (HLGN) to extract semantic-related signals essential in fMRI representations. The HLGN includes dual encoders aligning with text information and providing supplementary data to enhance generative model adaptability. Additionally, we incorporate a Low-Level Manipulation Net-

[1] The name combines "Neuro" (representing human neurons) with "Pictor" (an easel for painting), symbolizing the depiction of neuron status.

work (LLMN) to collaborate with the diffusion model by directly manipulating underlying features. Recent studies [25, 47] have shown advancements through feature map adjustments in generative models to control fine-grained visual details. Our LLMN combines a feature transformation technique with a U-net Encoder to guide image generation by updating feature maps. By decoupling the decoding process and separately controlling high-level semantics and low-level manipulation, our NeuroPictor achieves precise fMRI-to-image reconstruction.

We propose to learn fMRI-to-image decoding in three steps. i) The fMRI calibrated-encoding step. We pre-train the fMRI signal processing model for fMRI reconstruction, calibrating fMRI signals of different individuals to a shared latent space. ii) The fMRI-to-image multi-subject pre-training step. We pre-train our full model directly for fMRI-to-image decoding, leveraging about 67k fMRI-image pairs from different individuals. iii) For particular individual, we perform single-subject refining, using the same training strategy of step ii but focuses on the refinement on the single subject. Our results demonstrate the benefits of multi-individual pre-training for within-subject image reconstruction and the potential of the foundation fMRI decoding model. Extensive experiments validate the efficacy of our NeuroPictor.

Contribution. Our contributions are listed as follows: (1) We propose pre-training on multiple-individual fMRI-image pairs to enhence fMRI-to-image decoding for single individuals; (2) We capture diverse neural signal information and divide it into high-level and low-level guidance to supervise the diffusion generation process; (3) We integrate all the components as a unified model for multi-subject pre-training and single-individual refining, adapting the full model directly based on the supervision of fMRI-to-image decoding.

2   Related Work

fMRI-to-image Reconstruction. This task focuses on reconstructing visual images from recorded fMRI signals. The classical methods typically transform it into semantic content identification [10], image classification [18, 43] or retrieval tasks [21], or limited to simple images such as handwritten digits [35]. With the development of deep networks, Generative Adversarial Networks (GANs) and VAE are employed to better reconstruct natural images [17, 26, 31, 39]. Recently, diffusion models [32] have spurred studies [8, 12, 13, 28, 36, 46] in natural image decoding. Particularly, Chen *et al.* [8] improve visual quality and semantic consistency of reconstructed images using masked brain modeling and the Latent Diffusion Model (LDM). However, the SD text condition utilized in their approach cannot properly handle the learning of detailed spatial conditions. To achieve consistent reconstructions both semantically and visually, some works [28, 36] use CLIP text and image features along with VAE to generate blurry intermediate images, further refined with image-to-image Versatile Diffusion [45]. Other recent works employ depth maps [13] and silhouette maps [46] for low-level consistency, yet potential information gaps remain. In contrast, NeuroPictor, our

proposed approach, adapts the diffusion model to directly extract and decode fMRI signals without relying on potentially insufficient intermediate information.

Conditional Diffusion Model. The vanilla stable diffusion [32] generates images based on text prompts. Conditional diffusion models incorporate additional modalities like category [2], sketch [3], depth [25], normal [20], and semantic maps [46] to guide image generation. Efforts have focused on fine-tuning stable diffusion for tasks such as image editing [4] and personalization [14, 34]. T2i-adapter [25] modifies stable diffusion's feature maps directly for precise control, while ControlNet [46] learns a trainable-copy of SD to control generation with localized image conditions. Inspired by these methods, we employ a low-Level manipulation network, combining feature transformation techniques with a U-net Encoder, enabling intricate low-level detail learning from fMRI signals, leading to a fMRI-controllable generation model.

fMRI Representation. Learning a cross-subject representation for the fMRI data could enhance the down-stream tasks. In [7,9], the masked brain modeling is pretrained on the Brain, Object, Landscape Dataset (BOLD5000) [6] to grasp useful context knowledge of fMRI data. However, due to the varied lengths of flattened 1-D fMRI voxels across individuals, only a single model can be trained per subject for subsequent image decoding. Recently, Qian *et al.* [29] present a large-scale transformer-based fMRI autoencoder using the UK Biobank dataset (UKB) [24], transforming individual native-space fMRI signals into unified 2D brain activation images for multi-individual brain modeling in a large-scale latent space. Different from previous works [7, 29] learn an individual-specific fMRI-to-image decoder, our NeuroPictor shows that leveraging the pretraining on multiple-individual fMRI-image pairs, can actually significantly improve the fMRI-to-image decoding of one individual. Concurrent work [37] also aims to develop a shared-subject fMRI-to-image model. Different from their approach, which uses subject-specific ridge regression to map individuals' fMRI data into a common space, our method directly transforms native-space fMRI signals from any individual into unified 2D activation images. This eliminates the need for subject-specific encoding modules.

3   Method

Overview.   The framework of NeuroPictor is illustrated in Fig. 3. NeuroPictor is trained by three steps of fMRI calibrated-encoding, multi-subject pre-training and single-subject refining for fMRI-to-image decoding.

i) In the fMRI calibrated-encoding stage, a universal latent fMRI space is learned to represent fMRI signals across multi-individuals, addressing the individual differences and data scarcity. This latent space is trained via a transformer-based auto-encoder for reconstructing the fMRI brain surface maps (Sec. 3.1).

ii) In the multi-subject pre-training stage, we learn to extract high-level semantic features (Sec. 3.2) and low-level structure features (Sec. 3.3) from the above learned fMRI latent, and then use these features as conditions to guide the generation process of diffusion model. These components are integrated via a
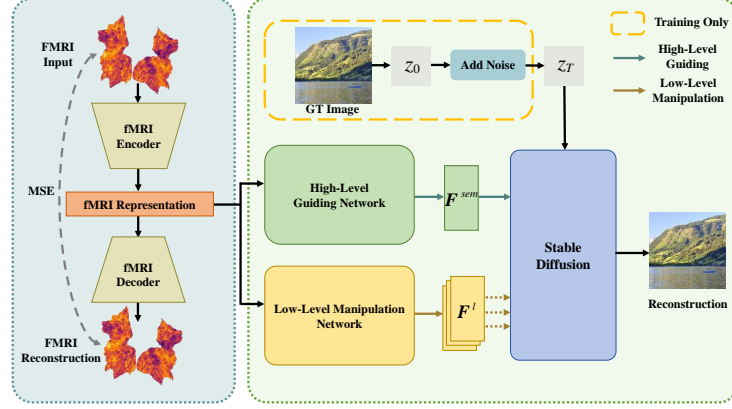
Fig. 3: Our NeuroPictor framework is trained in three steps for fMRI-to-image decoding. **i)** the fMRI calibrated-encoding stage, which establishes a universal latent fMRI space across multiple individuals; **ii)** the fMRI-to-image multi-subject pre-training stage, which achieves multi-level modulation through perceptual learning. **iii)** the fMRI-to-image single-subject refining stage, using the same strategy in step ii but focuses on refinement on particular subject.

unified model to directly train on fMRI-to-Image decoding task on multi-subject fMRI-image pairs.

iii) For the particular single-subject, we use the same training strategy as in step ii but focus on the single subject to refine our NeuroPictor.

3.1 fMRI Encoder

Inspired by the pre-processing of fMRI-PTE [29], we transform individual fMRI signals into unified 2D brain activation images, resulting in a 1-channel image S of size 256×256. This process unifies the fMRI representation across different individuals and thus enables training over multiple individuals to prevent missing signals of different individuals. However, fMRI-PTE [29] compresses fMRI signals using VQ-GAN, contrary to our goal of learning a universal latent representation without compression. Instead, we directly use the masked auto-encoder to reconstruct the fMRI surface map. In contrast to the random masking strategy of MAE, we mask all the patches in the latent space and introduce an additional *guide token* along with masked tokens to reconstruct the surface map. This forces the MAE learns to preserve all the information of fMRI signals in the *guide token*, which is beneficial to learn a compressed and universal fMRI latent. We train the autoencoder on the UKB dataset [24] learning a universal fMRI latent space capable of encoding information from different individuals.

After training, we discard the decoder and utilize the encoder $\mathrm{E}(\cdot)$ to convert the fMRI surface map into latent representation, described by:

$$\boldsymbol{S}^r = \mathrm{E}(S), \tag{1}$$

(a) High-Level Guiding Network.

(b) Low-Level Manipulation Network.
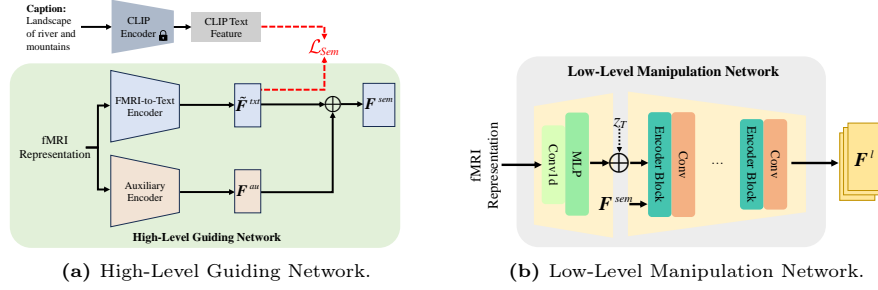
Fig. 4: Framework of our High-Level Guiding Network and Low-Level Manipulation Network.

where $\boldsymbol{S}^r \in \mathbb{R}^{L_r \times d_r}$, L_r and d_r represent the number of tokens and feature dimensions, respectively. $\boldsymbol{S}^r$ serving as the representation of fMRI data in the latent space. This fMRI latent is used to derive subsequent high-level semantic and low-level structural features. Please refer to Supplementary for more details.

3.2   High-Level Semantic Feature Learning

The vanilla SD model uses text prompts to generate images aligned with the given text. During the text condition stage, it utilizes a fixed CLIP encoder to encode the text prompts as $\boldsymbol{F}^{txt} \in \mathbb{R}^{L_T \times d_T}$. These prompts then guide the cross-attention layers in image generation. To maximize the potential of the high-level adapted cross-attention layers, we suggest substituting text conditions with high-level semantic fMRI features.

To fully upscale SD as a fMRI-to-image model, we replace the text encoder with a semantic feature learned from fMRI signals. Using the encoded fMRI feature $\boldsymbol{S}^r \in \mathbb{R}^{L_r \times d_r}$, we design a high-level guiding network to connect the fMRI latent space and text condition space. Illustrated in Fig. 4, the high-level guiding network processes fMRI representations through two parallel branches: the fMRI-to-text encoder aligns fMRI signals with CLIP Text features $\boldsymbol{F}^{txt}$ of image captions, while the auxiliary encoder provides additional semantic information to compensate for inaccuracies in the fMRI-to-text encoder and capture rich semantic details from images not explicitly described in captions.

In the fMRI-to-text encoder, two 1-D convolution layers are used to down-sample the token number L_r of fMRI representation $\boldsymbol{S}^r \in \mathbb{R}^{L_r \times d_r}$. Subsequently, MLP is employed to project it into text condition space:

$$\tilde{\boldsymbol{F}}^{txt} = \mathrm{E}_{txt}(\boldsymbol{S}^r), \tag{2}$$

where $\tilde{\boldsymbol{F}}^{txt} \in \mathbb{R}^{L_T \times d_T}$ is used to estimate the CLIP text feature of the ground truth image caption and E_{txt} is the fMRI-to-text encoder we introduced.

To further enforce the learned feature $\tilde{\boldsymbol{F}}^{txt}$ aligned with text condition, we use the caption of the original image to optimize it. Specifically, a frozen CLIP

text encoder is used to exact text feature $\boldsymbol{F}^{txt} \in \mathbb{R}^{L_T \times d_T}$ as the ground truth text condition. And we use mean square error as the semantic loss to optimize it, which can be formulated as:

$$\mathcal{L}_{sem} = \frac{1}{L_T} \sum_{i=1}^{L_T} \|\boldsymbol{F}_{(i)}^{txt} - \tilde{\boldsymbol{F}}_{(i)}^{txt}\|_2^2, \tag{3}$$

where $\boldsymbol{F}_{(i)}^{txt}$ and $\tilde{\boldsymbol{F}}_{(i)}^{txt}$ represents the i-th token feature of $\boldsymbol{F}^{txt}$ and $\tilde{\boldsymbol{F}}^{txt}$, respectively.

However, directly using the hard-mapped feature $\tilde{\boldsymbol{F}}^{txt}$ is not optimal. To address inaccurate semantic information learned in the fMRI-to-text branch, we introduce an auxiliary encoder to obtain an auxiliary semantic code $\boldsymbol{F}^{au}$. The architecture of the auxiliary encoder resembles that of the fMRI-to-text encoder, with the addition of a 1D zero convolutional layer in the final layer to initialize $\boldsymbol{F}^{au}$ as a zero tensor. This auxiliary semantic code is then residually added to the text code, resulting in the semantic representation $\boldsymbol{F}^{sem}$.

$$\boldsymbol{F}^{sem} = \tilde{\boldsymbol{F}}^{txt} + \boldsymbol{F}^{au}. \tag{4}$$

This adapted semantic feature $\boldsymbol{F}^{sem}$ is then fed into SD to guide the cross-attention layers.

3.3   Low-Level Manipulation Network

While learning general semantic information by mapping fMRI signals into a textual condition space is feasible, bridging the gap between fMRI and intricate images, rich in low-level details, poses challenges. To enhance low-level structure guidance, we introduce a Low-Level Manipulation Network (LLMN). This network injects low-level fMRI conditions by directly manipulating feature maps in the SD's U-net, adapting the generative model to fMRI signals.

Our LLMN consists of two components: a feature transformation module to convert fMRI representation to SD latent space, and a manipulator to incrementally trained to manipulate the SD latent.

Feature Transformation. Given an image of size $H \times W \times 3$, Stable Diffusion apply an Autoencoder similar to VQGAN [11] to convert the pixel-space image into latent space as $z_0 \in \mathbb{R}^{c \times h \times w}$. Then, a U-net [33] is used to progressively denoise images in the latent space.

To directly manipulate feature maps in SD's U-net, we introduce a feature transformation technique that simultaneously performs channel-wise and dimensional-wise feature learning on fMRI embeddings [2]. This connects fMRI representation learning with feature map learning, facilitating the subsequent manipulation process. Specifically, we employ two layers of 1D convolution to handle channel-wise correlations of fMRI embeddings and the MLP to conduct

---
[2] We use the terms "channel" and "dimension" to refer to the second and third dimensions of the fMRI latent tensor.

dimensional-wise feature learning. Finally, we use a reshape operation to form a feature map $\boldsymbol{F}_0^l \in \mathbb{R}^{c \times h \times w}$.

$$\boldsymbol{F}_0^l = \mathrm{E}_{ft}(\boldsymbol{S}^r). \tag{5}$$

The feature map $\boldsymbol{F}_0^l$ is then fed into the subsequent U-net encoder blocks, acting as condition controlling the generation process.

Low-Level Manipulation. To regulate the generation results with fine-grained low-level details, inspired by [47], we incrementally fine-tune the SD to improve the coherence between fMRI signals and visual stimulus. Utilizing a residual connection network, we effectively bridge the gap between fMRI representations and real natural images. Specifically, we incrementally manipulate the blocks of SD U-Net by adding low-level fMRI conditions within the proposed manipulation network. As depicted in Fig. 3, our manipulation network additionally take as inputs the feature map $\boldsymbol{F}_0^l$ converted from the fMRI signal. Then, to manipulate the SD latent in a fine-grained manner. We utilize a series of transformer layers to manipulate different blocks of SD. Denoted as E_U, this series of feature maps are produced by:

$$\boldsymbol{F}^l = \mathrm{E}_\mathrm{U}(\boldsymbol{F}_0^l), \tag{6}$$

where $\boldsymbol{F}^l = \{\boldsymbol{F}_{(i)}^l | i = 1, \cdots, 13\}$ and $\boldsymbol{F}_{(i)}^l$ is the i-th feature map produced by i-th encoder block.

These feature maps are passed into zero convolution layers and residually added with the outputs of the middle block and decoder blocks of locked SD branch as,

$$\tilde{\boldsymbol{F}}^l = \boldsymbol{F}_{sd} + \alpha \mathcal{Z}(\boldsymbol{F}^l), \tag{7}$$

where $\mathcal{Z}$ represent the zero convolution layer, $\boldsymbol{F}_{sd}$ represent the latent codes in SD U-net, α is a hyperparameters that balance the high-level semantic guidance with low-level detail manipulation. In this way, the feature maps in SD U-net is updated. We pass the feature map $\tilde{\boldsymbol{F}}^l$ into the subsequent layers of locked SD model to control the generation process. As a result, with the aid of the SD decoder, we obtain the final reconstructed image corresponding to specific fMRI signals. This incremental fine-tuning strategy is more efficient and stable than fully fine-tuning the SD.

3.4   Training and Inference

Training Objectives. During the training process, our NeuroPictor takes fMRI-image pairs as inputs. The input natural image $\boldsymbol{X}$ of size 512×512 is firstly compressed into latent space as $\boldsymbol{z}_0 \in \mathbb{R}^{64 \times 64 \times 4}$. Then, the diffusion process produces a noisy version $\boldsymbol{z}_t$ of $\boldsymbol{z}_0$ by progressively adding noise for t time steps to the initial feature $\boldsymbol{z}_0$. In the denoising stage, we use a frozen U-Net in Stable Diffusion and a trainable LLMN to jointly predict a denoised variant of the noisy input $\boldsymbol{z}_t$, conditioning on time step $\boldsymbol{t}$, semantic condition $\boldsymbol{F}^{sem}$, and a low-level feature map $\boldsymbol{F}^l$ converted from the fMRI representaion $\boldsymbol{S}^r$. The learning objective of the latent denoising stage can be formulated as:

$$\mathcal{L}_{dif} = \mathbb{E}_{\boldsymbol{z}_0, \boldsymbol{t}, \boldsymbol{S}^r, \epsilon \sim \mathcal{N}(0,1)} \left[\| \epsilon - \epsilon_\theta(\boldsymbol{z}_t, \boldsymbol{t}, \boldsymbol{S}^r, \boldsymbol{F}^{sem})) \|_2^2 \right], \tag{8}$$

where ϵ_θ represent the the LLMN. For simplify, we do not include the parameters of fMRI encoder E and semantic mapper E_T in Eq. (8). But these models are all trainable, and the parameters will be optimized with backpropagation of the gradients of $\boldsymbol{S}^r$ and $\boldsymbol{F}^{sem}$.

We combine diffusion loss with semantic loss (Sec. 3.2) to guide the training process. The final objective function is:

$$\mathcal{L} = \mathcal{L}_{dif} + \lambda\mathcal{L}_{sem}, \tag{9}$$

where λ denotes the loss weight to balance both terms.

Training Strategy. Benefiting from the unified fMRI representation in Sec. 3.1, NeuroPictor can be trained using fMRI-image pairs from different individuals. Thus, we split the training into three stages: i) *fMRI calibrated-encoding*: We pretrain the calibrated-encoder to enable the cross-subject fMRI latent space; ii) *Multi-subject pre-training*: During pretraining, the Stable Diffusion branch and LLMN are initialized with Stable Diffusion v2.1 trained on a large-scale dataset. We jointly train the full model, excluding the auxiliary encoder, using about 67,000 fMRI-image pairs from 8 subjects in the Natural Scenes Dataset [1] for pretraining, conducting 100k iterations to obtain a generalized fMRI-to-image generation model for these 8 subjects. iii) *single-individual refining*: for better modeling individually independent perception patterns, we fine-tune the network for an additional 60k iterations.

Inference. To maintain visual consistency, we balance high-level semantic guidance and low-level detail manipulation using a control scale instead of integrating them too faithfully. Essentially, focusing too much on high-level aspects can result in lost image detail, while prioritizing low-level features may impact the semantic processing abilities of the SD model, causing generated images to prioritize structural accuracy over semantic consistency. Thus in the inference stage, we can adjust control scale α to weight the SD output and LLMN output in Eq. (7). We show how control scale influence the generation process in Fig. 1. As the control scale enlarges from left to right, the decoded images gradually achieve the reconstruction of visual stimuli, progressing from high-level semantics only to high-level semantics and low-level structures consistently.

4   Experiments

4.1   Experimental setup

Dataset. We use Natural Scenes Dataset (NSD) [1] for both training and evaluation. NSD contains visual image stimulus and corresponding fMRI recordings of 8 subjects, with each subject viewing 8,000-9,000 images. The original images are collected from MS-COCO dataset [23], which are consisted of complex natural images. Following [27], we use the corresponding captions of the images in COCO dataset for training. For both training and evaluation, we average three trials of fMRI signal of the same images. Previous works only use the data of subjects 1, 2, 5 or 7 to train a model for each subject. Because these subjects
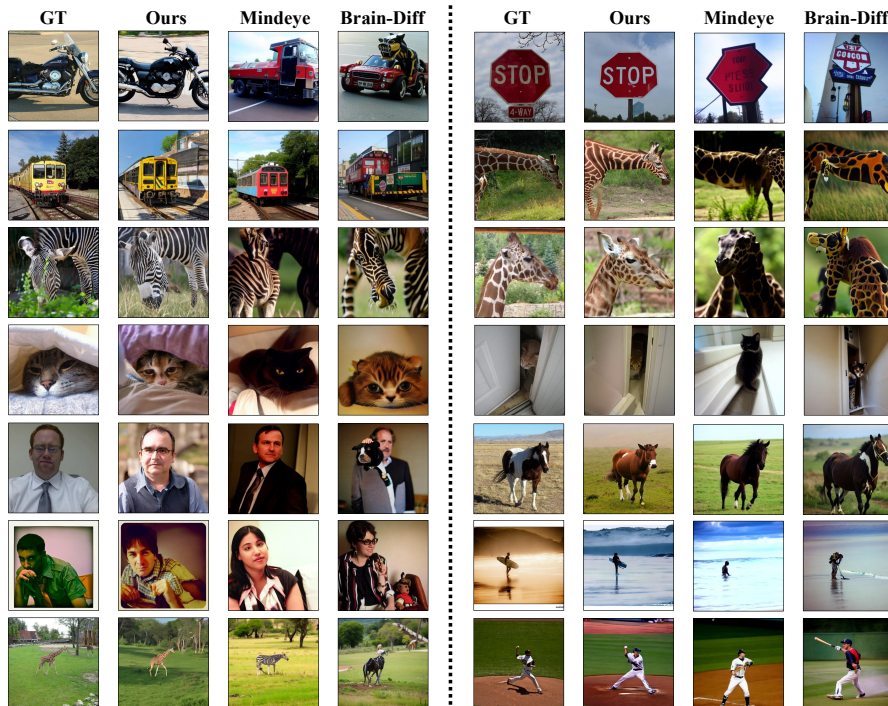
Fig. 5: Qualitative comparision of our NeuroPictor and previous state-of-the-art methods. Compared with other methods, our NeuroPictor achieves both high-level semantics and low-level structures consistency.

complete all scanning sessions, sharing the same 982 images as testing data. As our unified fMRI representations enable training across different person, we use all training data of 8 subjects to conduct a pretraining for modeling shared activity of all subjects. This scales up our training data to about 67k pairs. Then, following the setup of previous work, we finetune our NeuroPictor using the train split of subject 1, 2, 5 and 7, respectively, and evaluate our model using the shared test split.

Evaluation Metrics. We follow the metrics of Mindeye [36] to evaluate both high-level and low-level consistency. On the low-level aspect, we use pixelwise correlation, Structural Similarity Index Metric (SSIM) [44], AlexNet(2), and AlexNet(5). High-level metrics are calculated by extracting features using specific networks, including EffNet-B [42], SwAV [5], Inception [40], and CLIP [30]. Please refer to Supplementary for more details.

Implementation Details. The pretraining experiments are conducted using 6 NVIDIA RTX A6000 GPUs, with a batch size set to 96. During this stage, we exclude the auxiliary encoder and semantic loss to learn a base model. The entire network is trained for 100,000 iterations to achieve convergence. In the finetuning stage, 2 NVIDIA RTX A6000 GPUs are employed to finetune the

Table 1: Quantitative comparison of within-subject brain decoding of our NeuroPictor and the previous state-of-the-art mothods on Natural Scenes Dataset.

Method	Low-Level				High-Level			
	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
Lin et al. [22]	–	–	–	–	78.2%	–	–	–
Takagi... [41]	–	–	83.0%	83.0%	76.0%	77.0%	–	–
Gu et al. [16]	.150	.325	–	–	–	–	.862	.465
Brain-Cap [13]	**.353**	.327	89.0%	97.0%	84.0%	90.0%	–	–
Brain-Diff [27]	.254	.356	94.2%	96.2%	87.2%	91.5%	.775	.423
MindEye [36]	.309	.323	94.7%	97.8%	93.8%	**94.1%**	.645	.367
NeuroPictor (w/o ft)	.141	.349	91.4%	95.7%	88.3%	88.9%	.722	.417
NeuroPictor	.229	**.375**	**96.5%**	**98.4%**	**94.5%**	93.3%	**.639**	**.350**

Table 2: Ablation study on Subject-1. "Random init" means the fMRI Encoder starts without pre-trained weights from Sec. 3.1. Instead, it's trained from scratch using a standard Xavier uniform initialization. "Frozen" means the fMRI Encoder remains unchanged during fMRI-to-image training. "Full" initializes the fMRI Encoder with pre-trained weights, making it trainable during fMRI-to-image training. "Multi Pret." indicates whether fMRI-to-image multi-subject pre-training (subjects 1-8) is done. If not, the entire network is trained directly on Subject-1 data.

Method			Low-Level				High-Level			
Model	fMRI Enc.	Multi Pret.	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
NeuroPictor	random init		.128	.323	91.2%	95.1%	84.9%	86.0%	.769	.448
NeuroPictor	[CLS]		.139	.313	90.8%	94.2%	87.9%	87.6%	.736	.437
NeuroPictor	frozen		.220	.342	97.5%	98.8%	93.8%	92.8%	.657	.360
NeuroPictor	full		.265	.368	98.2%	99.1%	94.7%	93.4%	.641	.342
W/o Finetune	full	✓	.167	.350	95.1%	97.7%	89.1%	90.7%	.698	.399
W/o LLMN	full	✓	.111	.307	75.4%	88.4%	89.4%	89.9%	.746	.460
W/o Au. Enc.	full	✓	.221	.276	91.1%	97.5%	94.5%	91.0%	.668	.387
W/o $\mathcal{L}_{sem}$	full	✓	.255	.373	98.0%	98.9%	94.3%	93.0%	.647	.350
NeuroPictor	full	✓	**.277**	**.385**	**98.8%**	**99.3%**	**96.2%**	**94.5%**	**.619**	**.334**

model for each subject, spanning 60,000 iterations. The batch size for finetuning is set to 32. In Eq. (9), the parameter λ is set to 0.1. Additionally, we enables NeuroPictor to adapt to Classifier-Free Guidance [19] by randomly replacing 5% of the semantic features with unconditional CLIP embeddings corresponding to empty characters in the end of each stage. During evaluation, the unconditional guidance scale is set to 5.0.

4.2   Main Results

In this section, we comprehensively evaluate NeuroPictor against existing methods, using both quantitative and qualitative comparisons.

Competitors. We use the same test set as [36] and follow the values reported in [36]. Additionally, we include a new method [13] for comparision, which is also trained and tested on NSD.

Quantitative results. The results of quantitative comparison of our NeuroPictor and previous state-of-the-art methods are displayed in Tab. 1. We generate

one sample for each subject and calculate the low-level and high-level metrics following the official code of [36]. We average the metrics across 4 subjects (*i.e.* Subject-1, 2, 5 and 7) and report both the pre-trained (*i.e.* "w/o ft" in line 7 of Tab. 1) and fine-tuned results. The results show that our NeuroPictor outperforms other methods in six out of eight metrics, demonstrating that our reconstruction results can maintain semantic consistency while preserving low-level details present in the original images. We find that our method falls behind in pixel correlation metrics compared to three other models. This could potentially be attributed to the fact that other methods employ an image-to-image diffusion model and use a blurred image produced by VAE as an initialization. As reported in [13, 31], such a blurred image inherently exhibits higher pixel correlation than finely reconstructed results. Furthermore, our non-finetuned model achieves results comparable to other models, except for Mindeye. This validates that our training on multi-individual data effectively captures shared perceptual features across different individuals.

Qualitative results. We conduct qualitative comparisons with previous state-of-the-art works, namely Mindeye and Brain-diffuser, through visualization. We utilize the official codes to generate images for comparison. As shown in Fig. 5, our reconstruction results exhibit high visual consistency with the ground truth images, particularly in capturing some of the underlying details such as object structures, positions, and human poses. For instance, the appearance and poses of black motorcycles and trains, the positions of animals, as well as the gesture of people holding skateboards or engaged in throwing actions. We successfully reconstruct many challenging cases, such as a red "STOP" sign (row 1), animals like zebras and giraffes showing only their heads (row 3), and cats peeking out from under blankets or through door gaps (row 4). Additionally, our NeuroPictor successfully decodes the white borders (row 6) in visual stimuli, while other methods fails, highlighting the effectiveness of our direct low-level manipulation. Our reconstruction results maintain certain local matches with the original images, attributed to our end-to-end training fashion and pretraining on fMRI-image pairs.

4.3   Ablation Study

In this section, we perform ablation studies to evaluate the effects of our NeuroPictor framework techniques, focusing on Subject-1 for efficiency. Results are shown in Tab. 2, with additional discussion on failure cases in the Supplementary. **Ablation of fMRI Encoder.** We investigate the fMRI encoder's pretraining and feature selection impacts. Compared with training from scratch using a Xavier uniform initialization, pretraining an autoencoder on the UKB dataset before starting fMRI-to-image reconstruction significantly enhances performance, as shown in Tab. 2. Furthermore, fine-tuning the pretrained fMRI encoder, instead of freezing it during the fMRI-to-image stage, leads to improvements in all metrics. Switching the fMRI encoder's output from a high-dimensional feature to a [CLS] token results in reduced reconstruction quality.

Fig. 6: Interpolating the control scale between 0 and 1 transitions the reconstructed image from semantic consistency to fine-grained control.

Ablation of Multi-Subject Pre-training. We validate the impact of both our pretraining and finetuning in fMRI-to-image reconstruction stage by removing one at a time. The results indicate that pretraining on 8 subjects and finetuning on a single subject simultaneously improve the model's performance in both high-level and low-level aspects.

Ablation of High-Level Guiding Network and Low-Level Manipulation Network. Testing the High-Level Guiding Network without semantic loss shows its importance for achieving high-level consistency in reconstructions. Removing the auxiliary encoder and relying solely on the fMRI-to-Text Encoder for alignment with CLIP text features significantly degrades performance, highlighting the challenge of bridging semantic gaps between text captions and visual stimuli. Additionally, the presence or absence of the Low-Level Manipulation Network (LLMN) indicates its primary function in refining details, while also secondarily aiding in the preservation of high-level semantic coherence. As shown in Fig. 6, we also visualize generated results with a control scale varying between 0 and 1 to confirm the efficacy of both low-level and high-level pipelines in NeuroPictor.

5   Conclusions

In conclusion, this paper introduces NeuroPictor, a novel framework that tackle fMRI-to-image task by directly modulating the generation process of diffusion models using fMRI signals. Unlike previous approaches that focus on associating fMRI signals with pre-trained diffusion model conditions, NeuroPictor offers detailed control over image creation by dividing the fMRI-to-image process into three key steps: fMRI calibrated-encoding, fMRI-to-image single-subject refining and fMRI-to-image single-subject refining. NeuroPictor learns to exact high-level semantic features from fMRI signals and incrementally fine-tunes the diffusion model with a low-level manipulation network, providing precise structural instructions. Through training on over 60,000 fMRI-image pairs from diverse individuals, our model demonstrates superior fMRI-to-image decoding capabilities, particularly in within-subject scenarios.

Acknowledgements

References

1. Allen, E.J., St-Yves, G., Wu, Y., Breedlove, J.L., Prince, J.S., Dowdle, L.T., Nau, M., Caron, B., Pestilli, F., Charest, I., et al.: A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. Nature neuroscience **25**(1), 116–126 (2022)
2. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation (2023)
3. Bashkirova, D., Lezama, J., Sohn, K., Saenko, K., Essa, I.: Masksketch: Unpaired structure-guided masked image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1879–1889 (2023)
4. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
5. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems **33**, 9912–9924 (2020)
6. Chang, N., Pyles, J.A., Marcus, A., Gupta, A., Tarr, M.J., Aminoff, E.M.: Bold5000, a public fmri dataset while viewing 5000 visual images. Scientific data **6**(1), 49 (2019)
7. Chen, Z., Qing, J., Xiang, T., Yue, W.L., Zhou, J.H.: Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. arXiv preprint arXiv:2211.06956 (2022)
8. Chen, Z., Qing, J., Xiang, T., Yue, W.L., Zhou, J.H.: Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22710–22720 (2023)
9. Chen, Z., Qing, J., Zhou, J.H.: Cinematic mindscapes: High-quality video reconstruction from brain activity. arXiv preprint arXiv:2305.11675 (2023)
10. Cox, D.D., Savoy, R.L.: Functional magnetic resonance imaging (fmri)"brain reading": detecting and classifying distributed patterns of fmri activity in human visual cortex. Neuroimage **19**(2), 261–270 (2003)
11. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 12873–12883. Computer Vision Foundation / IEEE (2021)
12. Fang, T., Zheng, Q., Pan, G.: Alleviating the semantic gap for generalized fmri-to-image reconstruction. Advances in Neural Information Processing Systems **36** (2024)

13. Ferrante, M., Ozcelik, F., Boccato, T., VanRullen, R., Toschi, N.: Brain captioning: Decoding human brain activity into images and text. arXiv preprint arXiv:2305.11560 (2023)
14. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
15. Glover, G.H.: Overview of functional magnetic resonance imaging. Neurosurgery Clinics **22**(2), 133–139 (2011)
16. Gu, Z., Jamison, K., Kuceyeski, A., Sabuncu, M.: Decoding natural image stimuli from fmri data with a surface-based convolutional network. arXiv preprint arXiv:2212.02409 (2022)
17. Gu, Z., Jamison, K.W., Khosla, M., Allen, E.J., Wu, Y., St-Yves, G., Naselaris, T., Kay, K., Sabuncu, M.R., Kuceyeski, A.: Neurogen: activation optimized image synthesis for discovery neuroscience. NeuroImage **247**, 118812 (2022)
18. Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P.: Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science **293**(5539), 2425–2430 (2001)
19. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
20. Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., Zhou, J.: Composer: Creative and controllable image synthesis with composable conditions. arXiv preprint arXiv:2302.09778 (2023)
21. Kay, K.N., Naselaris, T., Prenger, R.J., Gallant, J.L.: Identifying natural images from human brain activity. Nature **452**(7185), 352–355 (2008)
22. Lin, S., Sprague, T., Singh, A.K.: Mind reader: Reconstructing complex images from brain activities. Advances in Neural Information Processing Systems **35**, 29624–29636 (2022)
23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
24. Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L., et al.: Multimodal population brain imaging in the uk biobank prospective epidemiological study. Nature neuroscience **19**(11), 1523–1536 (2016)
25. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
26. Mozafari, M., Reddy, L., VanRullen, R.: Reconstructing natural scenes from fmri patterns using bigbigan. In: 2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020. pp. 1–8. IEEE (2020)
27. Ozcelik, F., VanRullen, R.: Brain-diffuser: Natural scene reconstruction from fmri signals using generative latent diffusion. arXiv preprint arXiv:2303.05334 (2023)
28. Ozcelik, F., VanRullen, R.: Natural scene reconstruction from fmri signals using generative latent diffusion. Scientific Reports **13**(1), 15666 (2023)
29. Qian, X., Wang, Y., Huo, J., Feng, J., Fu, Y.: fmri-pte: A large-scale fmri pretrained transformer encoder for multi-subject brain activity decoding. arXiv preprint arXiv:2311.00342 (2023)

30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

31. Ren, Z., Li, J., Xue, X., Li, X., Yang, F., Jiao, Z., Gao, X.: Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning. NeuroImage **228**, 117602 (2021)

32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)

33. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)

34. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)

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

36. Scotti, P.S., Banerjee, A., Goode, J., Shabalin, S., Nguyen, A., Cohen, E., Dempster, A.J., Verlinde, N., Yundler, E., Weisberg, D., et al.: Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. arXiv preprint arXiv:2305.18274 (2023)

37. Scotti, P.S., Tripathy, M., Villanueva, C.K.T., Kneeland, R., Chen, T., Narang, A., Santhirasegaran, C., Xu, J., Naselaris, T., Norman, K.A., et al.: Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. arXiv preprint arXiv:2403.11207 (2024)

38. Seung, H.S., Lee, D.D.: The manifold ways of perception. science **290**(5500), 2268–2269 (2000)

39. Shen, G., Horikawa, T., Majima, K., Kamitani, Y.: Deep image reconstruction from human brain activity. PLoS computational biology **15**(1), e1006633 (2019)

40. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)

41. Takagi, Y., Nishimoto, S.: High-resolution image reconstruction with latent diffusion models from human brain activity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14453–14463 (2023)

42. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)

43. Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J.B., Lebihan, D., Dehaene, S.: Inverse retinotopy: inferring the visual content of images from brain activation patterns. Neuroimage **33**(4), 1104–1116 (2006)

44. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)

45. Xu, X., Wang, Z., Zhang, G., Wang, K., Shi, H.: Versatile diffusion: Text, images and variations all in one diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7754–7765 (2023)

46. Zeng, B., Li, S., Liu, X., Gao, S., Jiang, X., Tang, X., Hu, Y., Liu, J., Zhang, B.: Controllable mind visual diffusion model. arXiv preprint arXiv:2305.10135 (2023)
47. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)