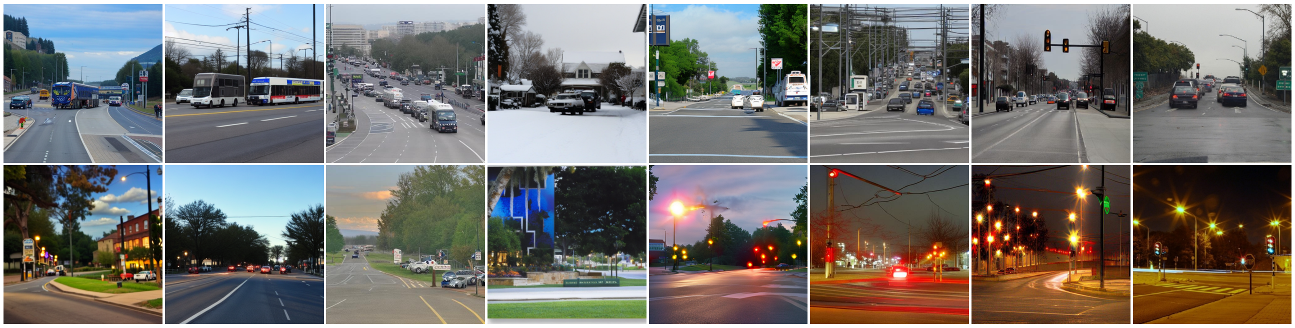
Cross-Subject Mind Decoding from Inaccurate Representations

Yangyang Xu[1]*, Bangzhen Liu[2,5], Wenqi Shao[3], Yong Du[4]*, Shengfeng He[5], Tingting Zhu[1]
[1]The University of Oxford [2]South China University of Technology [3]Shanghai AI Lab
[4]Ocean University of China [5]Singapore Management University



(a) BrainDiffuser (b) MindEye1 (c) MindBridge (d) MindEye2 (e) Neuropictor (f) Direct Decoding (g) Ours (h) Stimulus

Figure 1. We propose a framework for faithful cross-subject mind decoding. Unlike prior approaches that struggle with cross-subject generalization and accumulate errors in representation prediction, our method ensures more accurate and high-fidelity image reconstruction from fMRI signals.

Abstract

Decoding stimulus images from fMRI signals has advanced with pre-trained generative models. However, existing methods struggle with cross-subject mappings due to cognitive variability and subject-specific differences. This challenge arises from sequential errors, where unidirectional mappings generate partially inaccurate representations that, when fed into diffusion models, accumulate errors and degrade reconstruction fidelity. To address this, we propose the Bidirectional Autoencoder Intertwining framework for accurate decoded representation prediction. Our approach unifies multiple subjects through a Subject Bias Modulation Module while leveraging bidirectional mapping to better capture data distributions for precise representation prediction. To further enhance fidelity when decoding representations into stimulus images, we introduce a Semantic Refinement Module to improve semantic representations and a Visual Coherence Module to mitigate the effects of inaccurate visual representations. Integrated with ControlNet and Stable Diffusion, our method outperforms state-of-the-art approaches on benchmark datasets in both qualitative and quantitative evaluations. Moreover, our framework exhibits strong adaptability to new subjects with minimal training samples.

1. Introduction

The human visual cortex processes sensory stimuli and encodes them into brain signals, playing a fundamental role in shaping perceptual experience [11, 13, 37]. Decoding these brain signals back into the original stimuli has become a significant focus in neuroscience and computer science, presenting a challenging inverse problem with potential applications in brain-computer interfaces (BCIs) and cognitive science [7]. The functional Magnetic Resonance Imaging (fMRI), which captures changes in blood oxygenation, is widely used in BCIs for mind decoding due to its ability to reflect dynamic brain activity patterns [5, 21, 25, 35, 40].

Early methods for decoding fMRI data map voxel signals to the feature space of pre-trained Convolutional Neural Networks (CNNs) to classify object categories [9]. However, these methods were limited in reconstructing complex visual stimuli. To improve the realism of reconstructions, researchers introduced Generative Adversarial Networks (GANs) [3, 14, 47–49, 53] and diffusion models [33, 45, 46] for mind decoding, where fMRI data is typically mapped to image representations within generative models to guide image synthesis.

Despite these advances, current approaches are limited by individual cognitive variability and often rely on subject-specific models that require separate training for each individual [36, 42]. Such subject-specific models generally lack generalizability across individuals, reducing their scalability and applicability. Recently, cross-subject mind decod-

*Corresponding authors: Yangyang Xu (xuyangyang@hit.edu.cn) and Yong Du (csyongdu@ouc.edu.cn).

ing frameworks [28, 36, 42] have attempted to map fMRI voxels to shared representations across subjects. However, these frameworks face two sequential sources of error: i) unidirectional mappings often fail to capture the complex variability across subjects, producing partially inaccurate representations, and ii) these inaccurate representations are subsequently fed into pre-trained diffusion models without adjustments for errors, leading to compounded inaccuracies and frequently resulting in reconstructions with low fidelity and unrealistic details. As shown in Fig. 1, prior works fail to decode the night street scene.

To achieve high-fidelity mind decoding, we argue that both sources of inaccurate must be addressed. First, enhancing the accuracy of image representations during the initial mapping stage is essential to minimize inaccuracies. Second, the framework must be resilient to inaccurate in image representations during downstream processing to prevent error propagation and ensure that final reconstructions maintain high fidelity.

To address these challenges, we first propose a Bidirectional Autoencoder Intertwining (BAI) framework, which learns a bidirectional mapping between fMRI voxels and semantic/visual representations, capturing complex cross-subject relationships between these domains. By intertwining transformations in both directions, it not only improves fidelity in fMRI-to-image decoding but also enables the synthesis of fMRI-like data from semantic/visual inputs, yielding more reliable representations. However, as shown in Fig. 1f, decoding the representations predicted by BAI directly cannot guarantee the fidelity reconstruction, as it ignores the error in second step. To support inaccurate-tolerant decoding, we introduce the Semantic Refinement Module (SRM) and Visual Coherence Module (VCM), which mitigate the impact of representation errors on image reconstruction. Specifically, BAI employs two intertwined autoencoders for fMRI voxels and representations, achieving bidirectional mapping by swapping decoders. To reduce subject-specific biases, we incorporate a Subject Bias Modulation Module (SBMM) in the fMRI autoencoder, which applies statistical modulation. The Semantic Refinement Module refines the predicted semantic embedding, while the Visual Coherence Module optimally integrates visual representations to ensure output fidelity [20]. Combined with ControlNet [51], these modules reduce dependence on precise representations, preserving both semantic and visual consistency with the original stimuli. We evaluate our approach on the Natural Scenes Dataset [1], and results demonstrate that our framework outperforms state-of-the-art methods. Furthermore, our framework adapts effectively to new subjects with minimal additional samples.

In summary, our contributions are threefold:

- We propose a cross-subject mind decoding framework that learns bidirectional mappings between fMRI voxels and semantic/visual representations, capturing complex cross-subject relationships between these domains.
- We design Semantic Refinement and Visual Coherence modules to enhance reconstruction accuracy and consistency, reducing dependence on exact representations for high-fidelity mind decoding.
- Extensive experiments demonstrate that our framework significantly outperforms state-of-the-art methods in cross-subject mind decoding and adapts effectively to new subjects with minimal additional samples.

2. Related Works

Brain Decoding. Brain decoding aims to reconstruct stimuli from brain signals [4, 15, 44]. Early studies demonstrate that coarse visual information could be decoded from fMRI [6, 12, 41]. With advancements in deep learning, Tomoyasu *et al*. [9] map fMRI signals to CNN features. Recently, pre-trained generative models have shown powerful generative capabilities, and several studies leverage these models for brain decoding. Furkan *et al*. [26] and Milad *et al*. [24] extract representative features from fMRI and fine-tuned pre-trained BigGAN[3] for stimuli reconstruction. MindReader [17] maps fMRI signals to CLIP embeddings [29], then decoded stimuli images using conditional StyleGAN2 [14, 53]. Recent works introduce diffusion models [33, 46] for mind decoding by mapping fMRI signals to intermediate representations [5, 21, 35, 40, 43]. While promising results have been obtained, these approaches typically require separate model training for different subjects, limiting their broader applicability. Some recent works [28, 36, 42] have introduced cross-subject frameworks that unify different subjects within a single model. However, these approaches often suffer from inaccurate predicted representations due to the unidirectional mappings.

Diffusion Models in Brain Decoding. Diffusion models [30–34, 46] have made significant progress in generating diverse and realistic images and videos. Many studies leverage the generative power of diffusion models for brain decoding. Specifically, versatile diffusion [46] unifies text-to-image and image-to-text synthesis within a single framework, and various works [35, 36, 42] employ versatile diffusion as a visual decoder by mapping fMRI signals into embeddings. Stable Diffusion (SD)[33] performs denoising in latent space, enabling high-quality text-to-image synthesis. Takagi *et al*. [40] map fMRI signals into SD's latent representation for stimuli reconstruction but produces blurry results. MindVis [5] learns discriminative features from fMRI and projects them into two conditions, controlling the generation process of SD via a cross-attention mechanism. DREAM [43] decodes semantics, depth, and color conditions from fMRI and guides the output of SD with T2I-adapter [23]. However, all these works overlook the impact of inaccurate representations on reconstruction fidelity. In

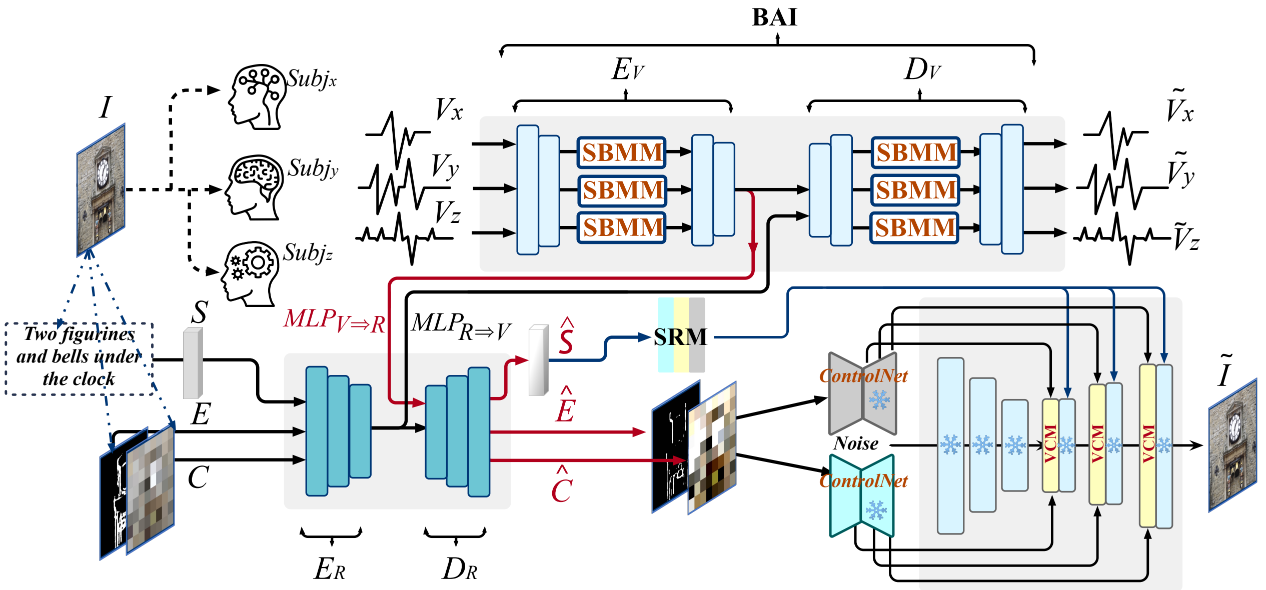


Figure 2. Overview of our framework. It consists of two autoencoders: one for fMRI voxels V_x and another for image representations (S, E, and C), supporting both reconstruction and bidirectional translation by swapping decoders. The Subject Bias Modulation Module (SBMM) is integrated into the fMRI encoder $\mathcal{E}_V$ and decoder $\mathcal{D}_V$ to eliminate inter-subject variance. Red lines illustrate the translation pipeline from fMRI voxels to image representations. To enhance reconstruction accuracy from inaccurate representations, we introduce the Semantic Refinement Module (SRM) and Visual Coherence Module (VCM) within the frozen ControlNet and Stable Diffusion (SD) components, reducing dependency on precise representations. Finally, the stimulus image $\tilde{I}$ is reconstructed using the DDIM sampler.

this paper, we learn bidirectional mappings between fMRI voxels and semantic/visual representations, enabling more accurate representation predictions. Additionally, we introduce two modules to reduce dependency on exact representations, further enhancing fidelity in reconstruction.

Controllable Diffusion Models. To better leverage the generative capabilities of diffusion models, various works [19, 45, 50–52, 54] enhance the controllability of pre-trained diffusion models. ControlNet [51] introduces zero-initialized layers to control the generation process under conditions, such as pose, edge, depth, and more. T2i-adapter [23] incorporates multiple adapters for pre-trained SD. Uni-ControlNet [52], UniControl [27], and ControlNet++[22] unify various control conditions within a single framework, ensuring that output images strictly adhere to multiply conditions. LooseControl[2] introduces rougher conditions to foster greater creative flexibility, while SmartControl [20] analyzes ControlNet's control mechanisms and proposes a module that resolves conflicts between text prompts and multiple control conditions. In our work, we aim to reconstruct realistic and faithful stimuli images despite the presence of both inaccurate semantic and visual conditions.

3. Methodology

Given fMRI voxels $V_x \in \mathbb{R}^d$ collected from subject x viewing a stimulus image $I \in \mathbb{R}^{h\times w\times 3}$, our goal is to develop a model that reconstructs the visual stimulus image $\hat{I}$ from V_x, independent of the specific subject. We follow the pipeline [25, 35, 42, 43] that map fMRI voxels to the representations of an image, and decode the representations using a pre-trained diffusion model. Our framework improves the reconstruction fidelity by improving the precision of mapped representations in the first stage, and the tolerance to inaccurate representations in next decoding stage. Inspired by DREAM [43], we divide a natural image into three representations: semantic embedding S for high-level information, edge map E for structure, and color palette C for low-level appearance. The overview of our method is shown in Fig. 2, we first introduce a bidirectional mapping between fMRI voxels and semantic/visual representations, and then introduce two modules that handle inaccurate representations. Finally, the stimulus image is reconstructed with ControlNet and SD.

3.1. Bidirectional Autoencoder Intertwining

We first predict three representations from fMRI voxels using the bidirectional autoencoder intertwining framework, which supports bidirectional mapping by intertwining transformations in both directions. In contrast to the widely used unidirectional mapping, bidirectional mapping encourages the model to capture complex relationships between fMRI voxels and image representations, resulting in more accurate predictions. Additionally, it introduces unsupervised cycle consistency between the two domains, providing additional supervision [55]. Specifically, our framework consists of two autoencoders: one for fMRI voxels and one for image representations. It supports both translation and reconstruction between the two domains. The reconstruction pipelines are represented as $V_x \Rightarrow \tilde{V}_x$, and $\{S, E, C\} \Rightarrow \{\tilde{S}, \tilde{E}, \tilde{C}\}$, where $\{\tilde{\cdot}\}$ denotes the reconstructed results. The translation pipelines are represented as $V_x \Rightarrow \{\hat{S}, \hat{E}, \hat{C}\}$ and $\{S, E, C\} \Rightarrow \hat{V}_x$, where $\{\hat{\cdot}\}$ denotes the predicted results.

Reconstruction. Given fMRI voxels V_x from subject x, they are encoded into a latent space shared across different subjects [42] by the fMRI encoder $\mathcal{E}_V$. Features in this shared latent space are subject-invariant, meaning they can be used not only for voxel reconstruction but also for translation to image representations. The reconstruction is performed by decoding the features with the decoder $\mathcal{D}_V(\cdot)$:

$$\tilde{V}_x = \mathcal{D}_V(\mathcal{E}_V(V_x)). \tag{1}$$

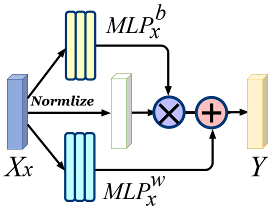


Figure 3. Structure of SBMM.

To obtain subject-invariant features, we reduce subjective bias in V_x by introducing a Subject Bias Modulation Module (SBMM) within the fMRI encoder for each subject. As shown in Fig. 3, the SBMM consists of multiple Multi-Layer Perceptrons (MLPs) that modulate the mean and variance of each subject's fMRI features. Let X_x denotes the input feature of the SBMM. The SBMM uses two MLPs to predict the statistical characteristics of each subject, which are then used to modulate the normalized input. The operation in SBMM can be represented as:

$$Y = \mathcal{MLP}_x^W(X_x)\left(\frac{X_x - \mu(X_x)}{\sigma(X_x)}\right) + \mathcal{MLP}_x^B(X_x), \tag{2}$$

where Y denotes the output of the SBMM, $\mathcal{MLP}_x^W(\cdot)$ represents the MLP for modulating the variance of the input, and $\mathcal{MLP}_x^B(\cdot)$ is the MLP for mean modulation. This module effectively reduces subject-specific bias. With this module, the framework can be efficiently adapted to new subjects by training only the SBMM for the novel subject. More discussion about this module can be seen in Sec. 4.4.

The decoder $\mathcal{D}_V(\cdot)$ has a mirrored structure to the encoder and is also integrated with SBMM, enabling it to decode subject-invariant features back into subject-specific fMRI voxels. We define an $\mathcal{L}_2$ distance between the input and reconstructed fMRI voxels for reconstruction:

$$\mathcal{L}_V^{\text{Rec}} = \|\tilde{V}_x - V_x\|_2. \tag{3}$$

Similarly, the autoencoder for representations receives the three representations simultaneously. Each representation is processed with a specific encoding head, and the resulting features are concatenated. The concatenated feature is then decoded into reconstructed representations using distinct decoding heads. Note that SBMM is not applied in this autoencoder, as the representations are subject-invariant. The reconstruction pipeline for representations is represented as:

$$\tilde{S}, \tilde{E}, \tilde{C} = \mathcal{D}_R(\mathcal{E}_R(S, E, C)), \tag{4}$$

where $\mathcal{E}_R(\cdot)$, $\mathcal{D}_R(\cdot)$ denote the encoder and decoder for representations respectively, and $\tilde{S}$,$\tilde{E}$,$\tilde{C}$ denote the reconstructed representations respectively. This autoencoder is trained with reconstruction loss of three representations:

$$\mathcal{L}_E^{\text{Rec}} = \texttt{BCE}(\tilde{E}, E), \quad \mathcal{L}_C^{\text{Rec}} = \|\tilde{C} - C\|_2, \tag{5}$$

$$\mathcal{L}_S^{\text{Rec}} = 1 - \texttt{cos}(\tilde{S}, S) + \|\tilde{S} - S\|_2, \tag{6}$$

where $\mathcal{L}_E^{\text{Rec}}$, $\mathcal{L}_C^{\text{Rec}}$, and $\mathcal{L}_S^{\text{Rec}}$ denote the reconstruction loss for edge map, color palette, and semantic embedding respectively. $\texttt{cos}(\cdot, \cdot)$ calculates the cosine similarity of two inputs, and $\texttt{BCE}(\cdot, \cdot)$ is the binary cross-entropy loss.

Bidirectional Mapping. The BAI framework supports flexible bidirectional translation by simply swapping the decoders. As illustrated by the red lines in Fig. 2, the translation from fMRI voxels to representations is performed by learning an MLP, $\mathcal{MLP}_{V \Rightarrow R}(\cdot)$, that maps the encoded fMRI features to representation features, which are then decoded using $\mathcal{D}_R$:

$$\hat{S}, \hat{E}, \hat{C} = \mathcal{D}_R(\mathcal{MLP}_{V \Rightarrow R}(\mathcal{E}_V(V_x))). \tag{7}$$

On the other hand, we can also mimic the human visual system by mapping the representations to the corresponding fMRI voxels using another MLP, $\mathcal{MLP}_{R \Rightarrow V}(\cdot)$, which maps representations features to the fMRI feature:

$$\hat{V}_x = \mathcal{D}_V(\mathcal{MLP}_{R \Rightarrow V}(\mathcal{E}_R(S, E, C))). \tag{8}$$

The translation pipeline is trained using fMRI-representation pairs. Similar to the reconstruction pipeline, the translation losses for fMRI voxels and the three representations are denoted as $\mathcal{L}_V^{\text{Tr}}$, $\mathcal{L}_S^{\text{Tr}}$, $\mathcal{L}_E^{\text{Tr}}$, and $\mathcal{L}_C^{\text{Tr}}$, respectively.

Moreover, our bidirectional mapping framework introduces cycle consistency through cyclic mapping. Specifically, we map the real representations to fMRI voxels using Eq. 8, and then reconstruct the representations back from the mapped voxels using Eq. 7: $\{S, E, C\} \Rightarrow \hat{V}_x, \hat{V}_x \Rightarrow \{\hat{\hat{S}}, \hat{\hat{E}}, \hat{\hat{C}}\}$,, where $\{\hat{\hat{\cdot}}\}$ denotes the cyclic reconstructed results. The cycle consistency loss for the three representations is computed by minimizing the distance between the input and cyclically reconstructed representations:

$$\mathcal{L}_K^{\text{Cyc}} = \|K - \hat{\hat{K}}\|_2^2, \quad K \in \{S, E, C, V\}. \tag{9}$$

Similarly, we can achieve cycle reconstruction for fMRI voxels with $V_x \Rightarrow \{\hat{S}, \hat{E}, \hat{C}\}, \{\hat{S}, \hat{E}, \hat{C}\} \Rightarrow \hat{\hat{V}}_x$. The BAI

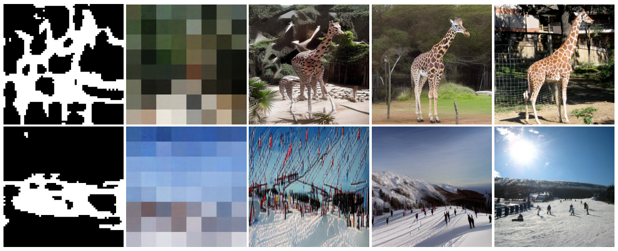

(a) Prd. Edge (b) Prd. Color (c) Direct Decod. (d) Our Decod. (e) Stimulus

Figure 4. "Direct Decod." denotes the decodes the predicted representations using diffusion models directly, which loses realism and fidelity. While using our decoded results are more faithful with stimulus GT.

framework is trained with the combined losses:

$$\mathcal{L} = \lambda_1 \underbrace{\mathcal{L}_S^{\text{Rec}} + \mathcal{L}_E^{\text{Rec}} + \mathcal{L}_C^{\text{Rec}} + \mathcal{L}_V^{\text{Rec}}}_{\text{Reconstruction}} + \lambda_2 \underbrace{\mathcal{L}_S^{\text{Tr}} + \mathcal{L}_E^{\text{Tr}} + \mathcal{L}_C^{\text{Tr}} + \mathcal{L}_V^{\text{Tr}}}_{\text{Translation}} + \lambda_3 \underbrace{\mathcal{L}_S^{\text{Cyc}} + \mathcal{L}_E^{\text{Cyc}} + \mathcal{L}_C^{\text{Cyc}} + \mathcal{L}_V^{\text{Cyc}}}_{\text{Cycle-Consistency}}, \tag{10}$$

where λ_s denote the balance factors, and we set λ_1 = 1, λ_2 = 1, and λ_3 = 0.5 empirically.

3.2. Reconstruction from Inaccurate Representations

The BAI maps fMRI voxels to image representations. A common solution for mind decoding involves using ControlNet [51] or T2I-adapter [23], which controls the output using these representations [43]. In this case, let F_i denote the feature of the i-th layer of the decoder $\mathcal{D}^{\text{UNet}}$ in the diffusion model's UNet, and the control process is performed by adding the representation features directly to F_i:

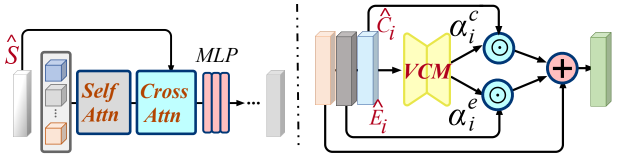
$$F_{i+1} = \mathcal{D}_i^{\text{UNet}}(F_i + \hat{E}_i + \hat{C}_i, \hat{S}), \tag{11}$$

where $\hat{E}_i$ and $\hat{C}_i$ represent edge and color features obtained from $\hat{E}$ and $\hat{C}$, respectively.

However, this operation requires extremely precise representations for accurate reconstruction, and the predicted representations often do not meet the strict requirements. As shown in Fig. 4c, decoding the predicted representations using diffusion models directly suffers from low fidelity due to inaccuracies. To mitigate this, we introduce two modules: the Semantic Refinement Module (SRM) and the Visual Coherence Module (VCM), which respectively handle errors in semantic and visual representations.

The Semantic Refinement Module (SRM) refines imprecise semantic representations. It is a transformer-based structure (see in Fig. 5a), trained by minimizing the distance between its output and the ground truth semantic embeddings:

$$\mathcal{L}_{SRM} = 1 - \cos(\text{SRM}(\tilde{S}), S) + \|\text{SRM}(\tilde{S}) - S\|_2, \tag{12}$$



(a) Semantic Refinement Module (b) Visual Coherence Module

Figure 5. Detailed structure of our SRM and VCM.

while this module use the similarly losses in training of BAI, this module tolerates imprecise semantic representations, thereby improving the fidelity during decoding stage.

We introduce VCM that harmonizes the imprecise visual representation features to original diffusion features. As indicated by [20], the weights of different features control the influence of visual conditions to the output image, our VCM is designed for predicting the weights that cohere different features. The structure of VCM is shown in Fig. 5b, it takes the concatenation of three features as input and produces two weights α_e and α_c with the same spatial size of representations:

$$\alpha_i^e, \alpha_i^c = \text{VCM}(\texttt{concat}(F_i, \hat{E}_i, \hat{C}_i)), \tag{13}$$

where $\texttt{concat}(\cdot)$ is the concatenate operation. The controlling process can be rewritten as:

$$F_{i+1} = \mathcal{D}_i^{\text{UNet}}(F_i + \alpha_i^e \odot \hat{E}_i + \alpha_i^c \odot \hat{C}_i, \hat{S}), \tag{14}$$

and the VCM is trained with:

$$\mathcal{L}_{\text{VCM}} = \|\epsilon - \epsilon_\theta(z_t, t, \text{SRM}(\tilde{S}), \tilde{C}, \tilde{E})\|_2^2. \tag{15}$$

Finally, the stimuli image $\tilde{I}$ can be reconstructed faithfully from imprecise representations with DDIM sampler [38].

4. Experiments

4.1. Experimental Settings

Dataset. We follow previous works [8, 28, 40, 42, 43] that conduct experiments on the largest mind decoding dataset, the Natural Scenes Dataset (NSD)[1]. NSD comprises 7-Tesla fMRI scans collected from eight subjects as they viewed thousands of stimulus images from the MS-COCO dataset[18]. Following prior studies [8, 40, 42], we use data from four subjects (Subj01, Subj02, Subj05, and Subj07) in our experiments. Each subject's training set consists of 8,859 fMRI-stimuli-caption pairs, while the test set includes 982 images viewed by all four subjects. The Regions of Interest (ROIs) in fMRI signals vary in size across subjects. To standardize these variations, we adopt the adaptive max pooling function used in MindBridge [42], which resizes the ROI representations to a fixed dimension of 8,192. We obtain semantic embeddings by encoding the image captions with the CLIP text encoder [29]. Additionally, edge maps are extracted from stimulus images using PidiNet [39]. To obtain

Table 1. Qualitative comparisons with related works on NSD dataset. All metrics are calculated as the average across 4 subjects.

Method	Cross Subject?	Low-Level				High-Level			
		PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
MindReader [17]	✗	-	-	-	-	78.2%	-	-	-
Takagi *et al.* [40]	✗	-	-	83.0%	83.0%	76.0%	77.0%	–	–
BrainDiffuser [25]	✗	.254	.356	94.2%	96.2%	87.2%	91.5%	.775	.423
MindEye1 [35]	✗	.309	.323	94.7%	97.8%	93.8%	94.1%	.645	.367
Gu *et al.* [8]	✗	.150	.325	-	-	-	-	.862	.465
MindVis [5]	✗	.080	.220	72.1%	83.2%	78.8%	76.2%	.854	.491
DREAM [43]	✗	.288	.338	95.0%	97.5%	94.8%	95.2%	.638	.413
MindBridge [42]	✓	.151	.263	87.7%	95.5%	92.4%	94.7%	.712	.418
MindEye2 [36]†	✓	.207	.350	91.6%	96.4%	89.4%	83.6%	.728	.423
NeuroPictor [10]	✓	.229	.375	96%	98.4%	94.5%	93.3%	.639	.350
UMBRAE [44]	✓	.283	.341	95.5%	97.0%	91.7%	93.5%	.700	.393
Psychometry [28]	✓	.297	.340	96.4%	98.6%	95.8%	96.8%	.628	.345
Ours	✓	**.318**	.356	**97.3%**	**98.8%**	96.7%	**97.5%**	.639	.345

† The result reported in original paper [36] is trained on 8 subjects, we re-train their model on 4 subjects using the official code for fair comparison.

color palettes, we follow the approach in T2I-Adapter [23], applying a 64× downsampling and subsequent upsampling to the original resolution.

Evaluation Metrics. Following existing works [25, 35, 36, 42], we use 8 evaluation metrics for the quantitative comparison from low and high levels. The low-level metrics include PixCorr, SSIM, AlexNet(2), and AlexNet(5), and the high-level metrics include Inception, CLIP, EffNet-B, and SwAV. For a detailed introduction to the metrics, please refer to the supplementary materials.

4.2. Implementation Details

We implement the proposed framework in PyTorch with Nvidia GeForce A100. The BAI framework is trained using the AdamW optimizer [16] with a learning rate of 1e-4. The batch size is set to 192 and we train for 1,000 epochs. The SRM and VCM require inaccurate representations for training, but the training set of NSD fits well with the trained BAI model, and their predicted representations contain fewer errors. To obtain the imprecise representations, we first generate 10,000 images using SD with random prompts generated using a Large Language Model (LLM), and then extract their representations as described in Sec. 4.1. We then reconstruct these representations with BAI's reconstruction pipelines, obtaining pseudo-imprecise representation-image pairs for training two modules. They are trained together with the AdamW optimizer [16] and a learning rate of 1e-4. We use Stable Diffusion V1.5 as our text-guided diffusion model, setting the inference steps in the DDIM sampler to 20. Two ControlNets are pre-trained on edge maps and color palettes respectively. For the structural details of the BAI, SRM, and VCM, please refer to the supplementary materials.

4.3. Comparison on Mind decoding

We present the quantitative comparisons with the state-of-the-art methods in Tab. 1. We can see that our method outperforms most of works on both low-level and high-level metrics. Especially, our method achieves higher values on PixCorr and SSIM metrics, which indicates that our decoded images are more similar to the stimulus images in structure and appearance. By predicting semantic knowledge through bidirectional mapping and refining it with SRM, our framework effectively decodes semantic information. As a result, our method receives the highest CLIP value among all methods. Moreover, our single cross-subject framework outperforms subject-specific frameworks on all metrics. By learning the bidirectional mapping with SBMM, our framework learns robust representation features that are agnostic to different subjects.

The qualitative comparison with other works can be seen in Fig. 1 and Fig. 6. We can observe that Takagi *et al.*'s method cannot reconstruct plausible images. By learning mappings to the versatile diffusion representation space [46], other methods reconstruct realistic images but fail to achieve semantic and visual consistency with the stimulus images. For example, none of them successfully reconstruct *"Pizza"* in the 1_{st} sample of Fig. 6, and they also fail to reconstruct *"Cat"* in the 2_{nd} sample. In contrast, our method captures the semantic information from fMRI voxels and reconstructs it correctly. Furthermore, our method also successfully captures the layouts of *"Bathroom"*, which is contributed by our edge representation prediction, capturing the structural information hidden in the fMRI voxels. Finally, our method successfully reconstructs the stimulus image color patterns, such as the color of the *"Bird"*, *"Train"*, and *"Horse"*, while none of the other methods recover the color of the *"Train"* successfully. This demonstrates the effectiveness of our color palette prediction. Unlike those versatile diffusion-based methods that map the fMRI voxels to CLIP visual embeddings, we disentangle the structure and appearance into edge and color palettes, together with SRM and VCM,

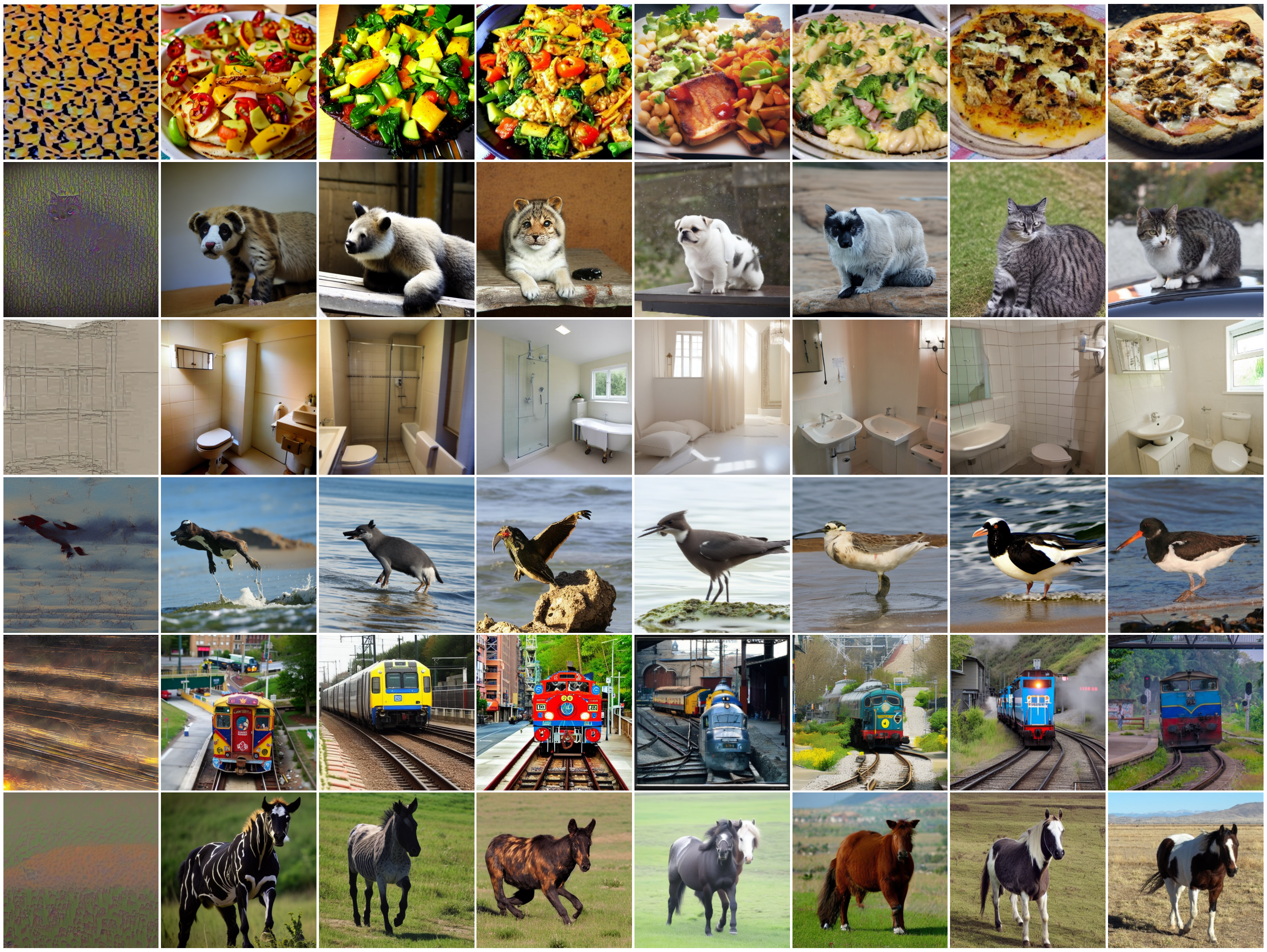


Figure 6. Qualitative comparison with competitors on mind decoding. Our reconstructed images are consistency with the stimulus images on semantic, structure, and appearance.

achieving a more faithful reconstruction.

4.4. Adaptation on New Subject

Our framework can be easily adapted to new subjects with few novel samples, which is valuable for real practice for reducing the number of fMRI-stimuli pairs. To simulate the scenarios of limited data, we follow MindBridge [42] that first trains the framework on three subjects (Subj01, 02, and 05), and adapts to the new subject (Subj07). Particularly, we only train the SBMM for the new subject while keeping the parameters in the other parameters fixed. We set 500, and 1,500 novel samples to adaptation. We also train a model from a sketch on a new subject with limited data. The adaptation results can be seen in Tab. 2. We can see that with the increasing number of data samples, our method boosts its performance both for the adapted version and training from the sketch version. Our adapted results outperform the model trained from sketch, which demonstrates the effectiveness of our framework on new subject adaptation. Additionally, our framework outperforms MindBridge in most metrics. The framework learns subject-irrelevant mappings between fMRI and representation spaces with SBMM, making it robust and flexible to new subjects.

4.5. Ablation Studies

In this section, we conduct ablation experiments to analyze our framework. We first analyze the effectiveness of our bidirectional mapping by proposing a Unidirectional Mapping (UM) from fMRI voxels to representations. In this variant, we remove the decoder of fMRI $\mathcal{E}_v$, the encoder of representations $\mathcal{D}_v$, and the MLP $\mathcal{MLP}_{R\Rightarrow V}(\cdot)$ that maps the representation features to the fMRI features. This variant is trained exclusively using translation losses $\mathcal{L}_V^{\text{Tr}}$, $\mathcal{L}_S^{\text{Tr}}$, and $\mathcal{L}_E^{\text{Tr}}$. To evaluate the effectiveness of SBMM on cross-subject decoding, we propose the variant w/o SBMM, by removing this module. Moreover, to demonstrate the influence of inaccurate representation predictions, we introduce the variant *Direct Addition*, by removing the SRM and VCM. In this

Table 2. Qualitative comparisons on new subject adaptation under various data limitation scenarios.

Method	# Samples	Adap-tation?	Low-Level				High-Level			
			PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
MindBridge	500	✗	.079	.171	73.5%	83.3%	74.4%	80.1%	.894	.587
MindBridge	500	✓	.112	.229	79.6%	85.0%	82.3%	86.7%	.840	.521
Ours	500	✗	.083	.213	74.2%	85.2%	76.2%	84.3%	.855	.572
Ours	500	✓	**.145**	**.234**	**82.1%**	**88.1%**	**85.4%**	**89.1%**	**.821**	**.503**
MindBridge	1,500	✗	.107	.206	79.4%	90.0%	82.4%	87.2%	.844	.523
MindBridge	1,500	✓	.140	.250	84.6%	**92.6%**	85.8%	91.0%	**.796**	.485
Ours	1,500	✗	.134	.242	82.1%	91.2%	83.4%	88.9%	.822	.513
Ours	1,500	✓	**.152**	**.267**	**85.2%**	92.3%	**86.1%**	**92.1%**	.814	**.491**

Table 3. Qualitative comparisons with variants on NSD dataset. All metrics are calculated as the average across 4 subjects.

Method	Cross Subject?	Low-Level				High-Level			
		PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
UM	✓	.281	.301	92.0%	93.4%	93.0%	90.3%	.719	.436
w/o SBMM	✓	.266	.274	83.7%	91.9%	89.6%	87.2%	.751	.460
Direct Addition	✓	.228	.259	83.1%	89.8%	87.2%	89.8%	.765	.477
Direct Addition + SRM	✓	.252	.281	89.7%	94.2%	94.2%	95.7%	.671	.374
Direct Addition + VCM	✓	.298	.287	94.4%	96.0%	92.1%	91.7%	.714	.391
Our-SS	✗	**.321**	.341	97.5%	98.9%	96.5%	97.1%	**.637**	**.355**
Our	✓	.318	**.356**	**97.3%**	**98.8%**	**96.7%**	**97.5%**	.639	.345

variant, we feed the predicted representations to ControlNet directly and fuse the two conditional features by simple addition using Eq. 11. To demonstrate the effectiveness of each module, we also propose variants *Direct Addition* + SRM and *Direct Addition* + VCM, by adding a specific module on *Direct Addition*. Finally, we train subject-specific models of our framework (Ours-SS) for each subject.

The quantitative comparison of various variants can be seen in Tab. 3. We can see that variant UM performs worse than the final framework, as UM learns the unidirectional mapping from fMRI voxels to representations, leading to inaccurate predicted representations. Although our SRM and VCM modules further reduce the need for accurate representations, images reconstructed by UM still suffer from low fidelity. The variant w/o SBMM performs much worse than our final framework, which demonstrates its effectiveness on cross-subject mind decoding. Without this module, the framework cannot learn informative representations from multiple subjects due to individual differences. Variant *Direct Addition* gets the worst on all metrics, which evidences the necessity of tolerating inaccurate representations in the second decoding stage. As discussed in Sec. 3.2, it requires accurate representations for fidelity reconstruction, and any inaccurate representations will mislead the reconstruction process. In contrast, our two proposed modules, SRM and VCM, improve the quality of the reconstruction from semantic and visual perspectives. Adding SRM to *Direct Addition* increases semantic-related performances, such as the CLIP score. Compared to *Direct Addition*, variant *Direct Addition* + VCM improves the low-level metrics effectively. Finally, our cross-subject variant achieves similar performance to the Ours-SS, demonstrating the effectiveness of our framework in capturing subject-irrelevant representations. The qualitative comparison of various variants can be seen in the supplementary.

5. Conclusion

In this paper, we introduce a cross-subject mind decoding framework that reconstructs stimulus images from fMRI voxels. We leverage the bidirectional mappings between fMRI voxels and semantic/visual representations. Combined with the subject bias modulation module, our method effectively captures complex relationships between these two domains across different subjects. To further improve the fidelity of decoded images, we propose semantic refinement and visual coherence modules. These modules reduce the dependency on highly precise image representations in the decoding stage. Extensive evaluations of the NSD dataset demonstrate that our framework outperforms prior methods in mind decoding and can be easily adapted to new subjects with minimal additional data.

Acknowledgment: This project is supported by National Natural Science Foundation of China (Grant 62102381), and the Guangdong Natural Science Funds for Distinguished Young Scholars (Grant 2023B1515020097), the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No.: AISG3-GV-2023-011), the Singapore Ministry of Education AcRF Tier 1 Grant (Grant No.: MSS25C004), and the Lee Kong Chian Fellowships. Tingting Zhu was supported by the Royal Academy of Engineering under Research Fellowship scheme.

References

[1] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022. 2, 5

[2] Shariq Farooq Bhat, Niloy Mitra, and Peter Wonka. Loosecontrol: Lifting controlnet for generalized depth conditioning. In *SIGGRAPH*, pages 1–11, 2024. 3

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019. 1, 2

[4] Bowen Pan Alex Andonian Emilie Josephs Alex Lascelles Camilo Fosco, Benjamin Lahner and Aude Oliva. Brain netflix: Scaling data to reconstruct videos from brain signals. In *ECCV*, 2024. 2

[5] Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *CVPR*, pages 22710–22720, 2023. 1, 2, 6

[6] David D Cox and Robert L Savoy. 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. 2

[7] Bing Du, Xiaomu Cheng, Yiping Duan, and Huansheng Ning. fmri brain decoding and its applications in brain–computer interface: A survey. *Brain Sciences*, 12(2):228, 2022. 1

[8] Zijin Gu, Keith Jamison, Amy Kuceyeski, and Mert Sabuncu. Decoding natural image stimuli from fmri data with a surface-based convolutional network. In *PMLR*, 2024. 5, 6

[9] Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):15037, 2017. 1, 2

[10] Jingyang Huo, Yikai Wang, Yun Wang, Xuelin Qian, Chong Li, Yanwei Fu, and Jianfeng Feng. Neuropictor: Refining fmri-to-image reconstruction via multi-individual pretraining and multi-level modulation. In *ECCV*, pages 56–73, 2024. 6, 7

[11] Nidhi Jain, Aria Wang, Margaret M Henderson, Ruogu Lin, Jacob S Prince, Michael J Tarr, and Leila Wehbe. Selectivity for food in human ventral visual cortex. *Communications Biology*, 6(1):175, 2023. 1

[12] Yukiyasu Kamitani and Frank Tong. Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5):679–685, 2005. 2

[13] Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17 (11):4302–4311, 2002. 1

[14] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020. 1, 2

[15] Benjamin Lahner, Kshitij Dwivedi, Polina Iamshchinina, Monika Graumann, Alex Lascelles, Gemma Roig, Alessandro Thomas Gifford, Bowen Pan, SouYoung Jin, N Apurva Ratan Murty, Oliva Aude Kay Kendrick, and Cichy Radoslaw. Modeling short visual events through the bold moments video fmri dataset and metadata. *Nature communications*, 15(1): 6241, 2024. 2

[16] Loshchilov LIlya and Hutter Frank. Decoupled weight decay regularization. In *ICLR*, 2017. 6

[17] Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind reader: Reconstructing complex images from brain activities. *NeurIPS*, 35:29624–29636, 2022. 2, 6

[18] 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*, pages 740–755, 2014. 5

[19] Haofeng Liu, Chenshu Xu, Yifei Yang, Lihua Zeng, and Shengfeng He. Drag your noise: Interactive point-based editing via diffusion semantic propagation. In *CVPR*, pages 6743–6752, 2024. 3

[20] Xiaoyu Liu, Yuxiang Wei, Ming Liu, Xianhui Lin, Peiran Ren, Xuansong Xie, and Wangmeng Zuo. Smartcontrol: Enhancing controlnet for handling rough visual conditions. In *ECCV*, 2024. 2, 3, 5

[21] Yizhuo Lu, Changde Du, Qiongyi Zhou, Dianpeng Wang, and Huiguang He. Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion. In *ACM Multimedia*, pages 5899–5908, 2023. 1, 2

[22] Huafeng Kuang Jie Wu Zhaoning Wang Xuefeng Xiao Chen Chen Ming Li, Taojiannan Yang. Controlnet++: Improving conditional controls with efficient consistency feedback. In *ECCV*, 2024. 3

[23] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, pages 4296–4304, 2024. 2, 3, 5, 6

[24] Milad Mozafari, Leila Reddy, and Rufin VanRullen. Reconstructing natural scenes from fmri patterns using bigbigan. In *IJCNN*, pages 1–8, 2020. 2

[25] Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023. 1, 3, 6, 7

[26] 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 *IJCNN*, pages 1–8, 2022. 2

[27] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. In *NeurIPS*, 2023. 3

[28] Ruijie Quan, Wenguan Wang, Zhibo Tian, Fan Ma, and Yi Yang. Psychometry: An omnifit model for image reconstruction from human brain activity. In *CVPR*, pages 233–243, 2024. 2, 5, 6

[29] 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 *ICLR*, pages 8748–8763, 2021. 2, 5

[30] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3, 2022. 2

[31] Jingjing Ren, Wenbo Li, Haoyu Chen, Renjing Pei, Bin Shao, Yong Guo, Long Peng, Fenglong Song, and Lei Zhu. Ultrapixel: Advancing ultra high-resolution image synthesis to new peaks. In *NeurIPS*, pages 111131–111171, 2024.

[32] Jingjing Ren, Wenbo Li, Zhongdao Wang, Haoze Sun, Bangzhen Liu, Haoyu Chen, Jiaqi Xu, Aoxue Li, Shifeng Zhang, Bin Shao, et al. Turbo2k: Towards ultra-efficient and high-quality 2k video synthesis. *arXiv preprint arXiv:2504.14470*, 2025.

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 2

[34] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, pages 36479–36494, 2022. 2

[35] Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. In *NeurIPS*, 2023. 1, 2, 3, 6, 7

[36] Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. In *ICML*, 2024. 1, 2, 6, 7

[37] Justine Sergent, Shinsuke Ohta, and Brennan Macdonald. Functional neuroanatomy of face and object processing: a positron emission tomography study. *Brain*, 115(1):15–36, 1992. 1

[38] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 5

[39] Zhuo Su, Jiehua Zhang, Longguang Wang, Hua Zhang, Zhen Liu, Matti Pietikäinen, and Li Liu. Lightweight pixel difference networks for efficient visual representation learning. *IEEE TPAMI*, 45(12):14956–14974, 2023. 5

[40] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *CVPR*, pages 14453–14463, 2023. 1, 2, 5, 6, 7

[41] Bertrand Thirion, Edouard Duchesnay, Edward Hubbard, Jessica Dubois, Jean-Baptiste Poline, Denis Lebihan, and Stanislas Dehaene. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage*, 33(4): 1104–1116, 2006. 2

[42] Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Mindbridge: A cross-subject brain decoding framework. In *CVPR*, pages 11333–11342, 2024. 1, 2, 3, 4, 5, 6, 7

[43] Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. Dream: Visual decoding from reversing human visual system. In *WACV*, pages 8226–8235, 2024. 2, 3, 5, 6

[44] Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. Umbrae: Unified multimodal brain decoding. In *ECCV*, pages 242–259, 2024. 2, 6

[45] Chenshu Xu, Yangyang Xu, Huaidong Zhang, Xuemiao Xu, and Shengfeng He. Dreamanime: Learning style-identity textual disentanglement for anime and beyond. 2024. 1, 3

[46] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *ICCV*, pages 7754–7765, 2023. 1, 2, 6

[47] Yangyang Xu, Xuemiao Xu, Jianbo Jiao, Keke Li, Cheng Xu, and Shengfeng He. Multi-view face synthesis via progressive face flow. *IEEE TIP*, 30:6024–6035, 2021. 1

[48] Yangyang Xu, Shengfeng He, Kwan-Yee K Wong, and Ping Luo. Rigid: Recurrent gan inversion and editing of real face videos. In *CVPR*, pages 13691–13701, 2023.

[49] Yangyang Xu, Shengfeng He, Kwan-Yee K Wong, and Ping Luo. Rigid: Recurrent gan inversion and editing of real face videos and beyond. *IJCV*, 133(6):3437–3455, 2025. 1

[50] Yuyang Yu, Bangzhen Liu, Chenxi Zheng, Xuemiao Xu, Huaidong Zhang, and Shengfeng He. Beyond textual constraints: Learning novel diffusion conditions with fewer examples. In *CVPR*, pages 7109–7118, 2024. 3

[51] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 2, 3, 5

[52] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Unicontrolnet: All-in-one control to text-to-image diffusion models. In *NeurIPS*, 2023. 3

[53] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *CVPR*, pages 17907–17917, 2022. 1, 2

[54] Haiming Zhu, Yangyang Xu, Chenshu Xu, Tingrui Shen, Wenxi Liu, Yong Du, Jun Yu, and Shengfeng He. Stable score distillation. In *ICCV*, 2025. 3

[55] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017. 3