See Through Their Minds: Learning Transferable Brain Decoding Models from Cross-Subject fMRI

Yulong Liu[1,2,3], **Yongqiang Ma**[1,2,3], **Guibo Zhu**[4,5*], **Haodong Jing**[1,2,3], **Nanning Zheng**[1,2,3*]

[1] National Key Laboratory of Human-Machine Hybrid Augmented Intelligence
[2] National Engineering Research Center of Visual Information and Applications
[3] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
[4]Institute of Automation, Chinese Academy of Sciences [5] Wuhan Al Research
{lylhubxy, jinghd}@stu.xjtu.edu.cn, {musayq, nnzheng}@mail.xjtu.edu.cn, gbzhu@nlpr.ia.ac.cn

Abstract

Deciphering visual content from fMRI sheds light on the human vision system, but data scarcity and noise limit brain decoding model performance. Traditional approaches rely on subject-specific models, which are sensitive to training sample size. In this paper, we address data scarcity by proposing shallow subject-specific adapters to map cross-subject fMRI data into unified representations. A shared deep decoding model then decodes these features into the target feature space. We use both visual and textual supervision for multimodal brain decoding and integrate high-level perception decoding with pixel-wise reconstruction guided by high-level perceptions. Our extensive experiments reveal several interesting insights: 1) Training with cross-subject fMRI benefits both high-level and low-level decoding models; 2) Merging high-level and low-level information improves reconstruction performance at both levels; 3) Transfer learning is effective for new subjects with limited training data by training new adapters; 4) Decoders trained on visually-elicited brain activity can generalize to decode imagery-induced activity, though with reduced performance.

Code — https://github.com/YulongBonjour/STTM

Introduction

Vision decoding from brain activity enables the inference of mental states and cognitive processes, aiding neuroscience development and potentially advancing brain-inspired artificial intelligence and brain-computer interfaces. To this end, functional Magnetic Resonance Imaging (fMRI) data are widely used as a non-invasive approach to capture brain activity. Recent years have witnessed great progress in fMRI-based vision decoding with the rapid development of Deep Learning and generative models(Ozcelik and VanRullen 2023; Liu et al. 2023; Scotti et al. 2023).

Despite significant progress, the subject-specific nature of most previous methods limits the models' generalization ability. Because collecting fMRI data is costly and prone to physiological noise, resulting in limited, low signal-to-noise samples per participant. Moreover, due to inter-individual variability in brain structure, models built from scratch for each new subject will lead to significant variability in results. While large datasets like the NSD dataset(Allen et al. 2022) have yielded exciting results, training models from scratch on smaller datasets remains vulnerable to overfitting.

Recent research shows that coarse-grained response topographies are highly similar across subjects, suggesting that individual idiosyncrasies are reflected in more nuanced response patterns(Chen et al. 2015; Güçlü and van Gerven 2017; Khosla et al. 2020). This indicates that decoding models can share representational spaces across subjects, reducing the challenges of limited per-subject data. Building on this insight, we propose a neural decoding model with shared decoding modules to capture common response patterns, and subject-specific adapters to accommodate individual biases. This approach allows for the integration of data from multiple subjects, viewing the same or different images, to learn general brain response patterns while capturing meaningful individual-level variations.

Our framework, called **STTM**, is shown in Figure 1. Inspired by bottom-up and top-down cognitive processes in the human brain (Katsuki and Constantinidis 2014; Miller 1999), STTM includes a high-level pipeline (**STTM-H**) that captures semantic perceptions and a semantic-guided low-level pipeline (**STTM-L**) focused on pixel-wise reconstruction, matching the original images' low-level features like color, texture, and spatial position. Both pipelines use subject-specific adapters to transform cross-subject inputs, with shared decoding modules processing the extracted features. The high-level pipeline aligns fMRI patterns with CLIP visual tokens using contrastive learning and a diffusion prior (Ramesh et al. 2022). It also aligns fMRI data with textual descriptions, allowing the model to handle multimodal brain decoding tasks. The low-level pipeline maps fMRI data onto the latent space of Stable Diffusion's VAE (Rombach et al. 2022) to generate initial blurry reconstructions. To improve pixel-wise reconstruction, we use the semantic features from the high-level pipeline to guide the low-level pipeline. Final reconstructions are produced by combining outputs from both pipelines in an img2img setting using Versatile Diffusion(Meng et al. 2022; Xu et al. 2023).

In this study, we conduct two types of experiments: 1) Model pre-training and testing using data from four subjects (1, 2, 5, 7) in the Natural Scenes Dataset (NSD); and 2)

*Corresponding Author

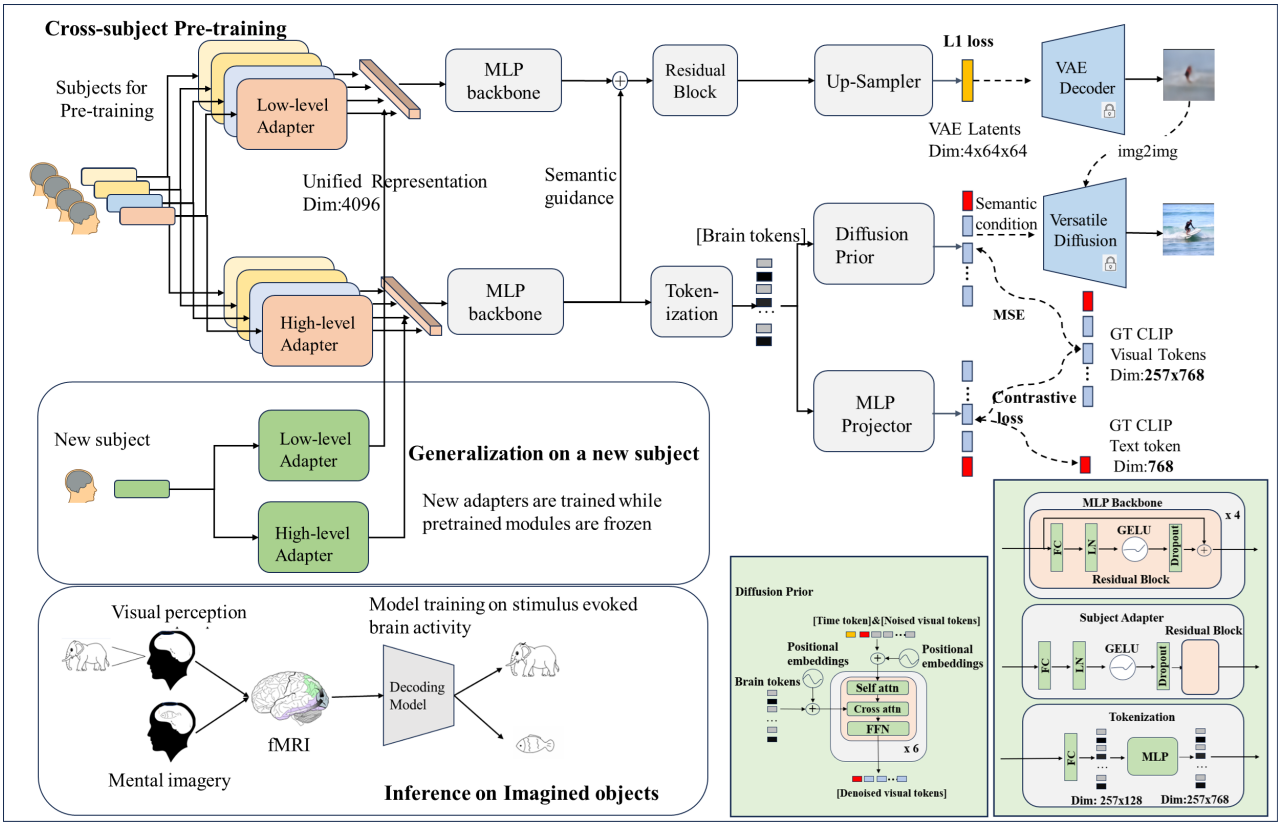

Figure 1: Overview of our STTM framework, which consists of a high-level perception decoding pipeline and a pixel-wise reconstruction pipeline(low-level pipeline). The two pipelines are trained sequentially. The high-level pipeline guides the pixel-wise reconstruction pipeline. The final reconstructions are generated in an img2img setting(Meng et al. 2022) using versatile diffusion model(Xu et al. 2023). Subject adapters transform cross-subject fMRI data into a unified feature space for the two pipelines. Transfer learning can be conducted by training new adapters for new subjects.

Transfer learning on the Generic Object Decoding (GOD) Dataset (Horikawa and Kamitani 2017), which has much fewer training samples per subject and is tested under a zero-shot setting. Our contributions are summarized as follows:

- To address the scarcity of fMRI data, we propose an adapter-based method to pre-train and fine-tune brain decoding models with cross-subject fMRI data. To our knowledge, this work is one of the earliest using this method, a concurrent work is MindEye2(Scotti et al. 2024).
- We identify the importance of the interaction between high-level and low-level perceptions for reconstruction performance and propose utilizing high-level perceptions to guide pixel-wise reconstruction.
- Our experiments demonstrate that the decoders trained on visual stimuli-evoked brain activity can generalize to decode imagery-induced brain activity but with reduced performance. This work is also the first one to generate high-resolution reconstructions of imagined objects.
- We open source a versatile brain decoding model with good transferability and high performance on a wide range of tasks, which may facilitate future multi-model brain decoding research.

Related Work

Visual Stimulus Decoding from FMRI. Early brain decoding studies primarily used linear models to map fMRI data onto an intermediate feature space for decoding basic visual attributes such as spatial position, orientation, and image categories(Thirion et al. 2006; Haynes and Rees 2005; Kamitani and Tong 2005; Cox and Savoy 2003; Haxby et al. 2001). With advances in deep learning and generative models, researchers began using CNNs, GANs, and diffusion models to reconstruct visual stimuli, resulting in more semantically accurate and faithful reconstructions(Shen et al. 2019; Beliy et al. 2019; Ren et al. 2021; Ozcelik et al. 2022; Lin, Sprague, and Singh 2022; Liu et al. 2023; Ozcelik and VanRullen 2023; Scotti et al. 2023; Ma et al. 2024). The emergence of models like CLIP(Radford et al. 2021) and related ones, such as Stable Diffusion(Rombach et al. 2022) and Versatile Diffusion(Xu et al. 2023), further advanced decoding approaches. Recent methods have integrated contrastive learning between fMRI data and CLIP model features, followed by reconstruction using diffusion models or GANs. For example, Mind-Reader(Lin, Sprague, and Singh 2022) and BrainCLIP(Liu et al. 2023) used the CLS token from CLIP's visual and textual encoder for global su-

pervision, while MindEye(Scotti et al. 2023) leveraged all 257 tokens from CLIP's visual encoder's last hidden layer, highlighting the benefits of fine-grained supervision for retrieval and reconstruction. In this paper, we propose combining global visual-linguistic contrastive learning with fine-grained visual contrastive learning to enhance multi-modal brain decoding. We also emphasize the importance of the interaction between bottom-up and top-down processes for stimulus reconstruction.

Mental Imagery and Visual Perception Mental imagery, the ability to create images in the mind without external stimuli, often produces weaker or less vivid images than those triggered by sensory input, yet both rely on the visual system(Dijkstra et al. 2018). Studies have found that around two-thirds to over 90% of brain regions involved in visual perception are also activated during mental imagery(Kosslyn, Thompson, and Alpert 1997; Kosslyn et al. 1999; Ganis, Thompson, and Kosslyn 2004; Lee, Kravitz, and Baker 2012). Thus, decoding the information in the visual cortex can help visualize the mental imagery process.

FMRI Foundation Models for Visual Decoding. While traditional brain decoding methods focused on subject-specific pipelines, recent efforts aim to develop fMRI foundation models by pretraining on large-scale data from diverse subjects to capture generalizable neural representations. Chen et al. (2023) used masked brain modeling (MBM) to transform fMRI series into embeddings. Malkiel et al. (2022) applied self-supervised techniques, including MBM, for audio-evoked fMRI data. Qian et al. (2023) converted fMRI signals into unified 2D representations using anatomical information, maintaining consistency across individuals while preserving distinct brain patterns. These pretrained models still require fine-tuning for individual subjects to account for biological nuances in visual stimuli generation. In our approach, instead of using a single network for cross-subject fMRI data, we train shallow, subject-specific adapters alongside a shared deep decoding network, enabling transfer learning by training new adapters for new subjects.

Method

Cross-Subject High-Level Perceptions Decoding

Our high-level pipeline captures semantic perceptions from fMRI data, aligning it with visual and textual modalities. This versatility supports tasks such as fMRI-to-image retrieval, fMRI-to-text retrieval, zero-shot classification, and fMRI-to-image generation.

High-Level Model Architecture The high-level model translates fMRI data into CLIP embedding space and includes several components: shallow subject-specific adapters, a shared MLP backbone with 4 residual blocks, a tokenization module, a diffusion prior module, and an MLP projector. Subject adapters use a linear projection and a residual block to align fMRI data across subjects into a unified feature space. The shared MLP backbone refines these features into a higher-level space. The tokenization module converts features into 257 fine-grained tokens. Tokens are processed in parallel by an MLP projector and a diffusion prior module. The model is trained end-to-end with MSE loss for the diffusion prior and bidirectional contrastive loss for the projector. Projector outputs support retrieval tasks, while diffusion prior outputs guide image generation using a pre-trained versatile diffusion model (Xu et al. 2023). Unless stated otherwise, we use the CLIP ViT/L-14 model.

Global Visual-Linguistic Contrastive Learning & **Fine-Grained Visual Contrastive Learning** Previous studies indicate that contrastive learning generates robust fMRI representations. Notably, BrainCLIP enhances reconstruction by using global embeddings from both CLIP's visual and textual encoders (Liu et al. 2023). MindEye utilizes all 257 CLIP visual tokens, demonstrating the advantages of fine-grained supervision for image retrieval and reconstruction (Scotti et al. 2023). To further improve brain decoding across visual and textual modalities, we propose combining global visual-linguistic contrastive learning (GVLC) with fine-grained visual contrastive learning (FVC). Specifically, FVC contrasts the 257 flattened and L2-normalized CLIP visual tokens (V_f) with the corresponding projected Brain tokens (B_f) from the MLP projector,

$$L_{FVC} = Contrast(V_f, B_f). \tag{1}$$

And the GVLC is applied to the CLS token(B_{CLS}) of the Brain tokens by combine the supervision from CLIP visual and textual CLS token(V_{CLS} and T_{CLS}), i.e.,

$$\begin{aligned} L_{GVLC} = \frac{1}{2}[&Contrast(V_{CLS}, B_{CLS}) \\ &+Contrast(T_{CLS}, B_{CLS})]. \end{aligned} \tag{2}$$

Two kinds of contrastive loss are used in this work. For the first 35% of total epochs, we use the BiMixCo loss(Scotti et al. 2023), which applies the Mixup technique to train models on synthetic fMRI data generated by combining two fMRI-stimulus pairs, addressing data scarcity for a single subject. For the remaining epochs, we switch to the Soft-CLIP loss(Scotti et al. 2023), which uses batch-wise visual CLIP embedding similarity instead of one-hot labels as the target. Both losses are bidirectional (see Appendix for details).

Efficient Diffusion Prior Learning Contrastive learning yields disjoint fMRI embeddings, known as the "Modality Gap"(Liang et al. 2022). To reconstruct images, we train a diffusion prior proposed in DALL-E2(Ramesh et al. 2022) to produce aligned CLIP embeddings from the outputs of the MLP backbone. These embeddings can serve as inputs to any pre-trained image generation model accepting CLIP image embeddings. MindEye applied a similar approach to map fMRI to CLIP image embeddings, predicting denoised CLIP tokens from brain and noise tokens. However, its encoder-only transformer architecture requires significant memory for long sequences. To reduce memory usage, we use a 6-layer transformer decoder for the diffusion prior.

The diffusion prior receives three token types: brain tokens, noisy CLIP visual tokens, and a time token. Brain tokens serve as keys and values for cross-attention modules

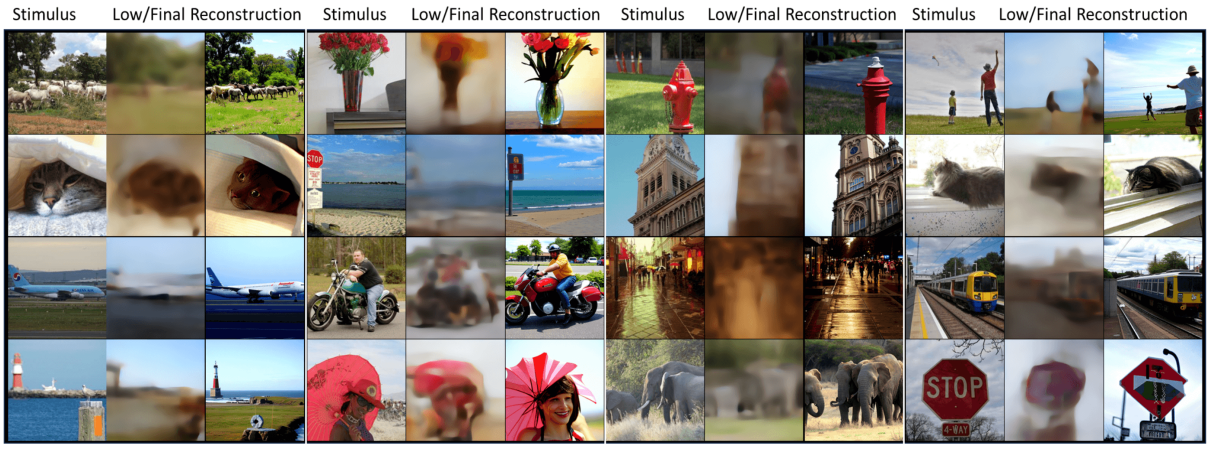

Figure 2: Reconstruction examples for subject 1 in the NSD dataset. The "Low reconstructions" are from the low-level pipeline and the "final reconstructions" are obtained in the img2img (Meng et al. 2022) setting.

in the transformer decoder after adding positional embeddings. The time token and noisy CLIP visual tokens are concatenated as queries, with positional embeddings added to the latter. The diffusion prior module are trained to output denoised CLIP tokens, utilizing the same diffusion loss as Ramesh et al(Ramesh et al. 2022). Inspired by the success of masked autoencoder(He et al. 2022), we propose training the diffusion prior with only a small part(e.g., 35%) of the predicted Brain tokens. This further reduces memory usage, allowing training on accelerators with just 16GB memory.

Our total end-to-end loss for the high-level pipeline is defined as:

$$L_{high} = \alpha(L_{FVC} + \beta L_{GVLC}) + (1-\alpha)L_{prior}. \quad (3)$$

In our experiments, we set β to 0.4 and employ random weighting(Lin et al. 2022) between the contrastive losses and the diffusion prior loss.

Cross-Subject Pixel-Wise Reconstruction with High-Level Perception Guidance

Our pixel-wise reconstruction pipeline focuses on recovering low-level image details, such as texture and boundaries, that may be missing from CLIP visual and textual tokens. As shown in Figure 1, this pipeline combines bottom-up processing with top-down feedback mechanisms. It includes shallow, subject-specific adapters followed by a shared residual MLP backbone. The outputs from this backbone are enhanced with semantic information from the high-level pipeline, with the two sets of information added together. The merged features are processed by a residual block and up-sampled to the latent space of Stable-diffusion's VAE via a CNN. We use L1 loss during training. To maximize the use of both high-level and bottom-up information, we introduce a mechanism where, in 30% of training steps, the semantic feedback is replaced with a learnable embedding vector. Additionally, in 25% of steps, the low-level pipeline outputs are replaced with another learnable embedding vector. After training, the VAE decoder generates blurry reconstructions from the predicted latent embeddings, which serve as initial states for the versatile diffusion model.

Transfer Learning for New Subjects

Collecting large-scale training data for each new subject is challenging. To address this, we propose an adapter-based approach that leverages general knowledge from cross-subject fMRI patterns to improve decoding performance for new subjects with limited data. Our method involves freezing the pre-trained shared decoding modules and training a shallow adapter for each pipeline, aligning the new subject's fMRI data with the unified feature space. The training objective remains consistent with the pre-training stage. Once the subject adapter is trained, both the adapter and MLP backbone are frozen, and we fine-tune the remaining modules to better align the fMRI patterns with the target domain.

Experimental Results

Datasets and Setting

Datasets For cross-subject pre-training and evaluation we use the Natural Scenes Dataset (NSD) dataset(Allen et al. 2022), which is currently the largest neural imaging dataset for data-driven brain decoding. Following common practices, our research uses data from 4 subjects(1,2,5,7), who completed all the scan sessions. We used the NSDGeneral ROI mask at 1.8 mm resolution to derive ROIs for the 4 subjects, spanning visual areas from early to higher visual cortex. Corresponding captions were extracted from the COCO dataset. We then conduct transfer learning on the GOD dataset(Horikawa and Kamitani 2017), which has much fewer training samples for each subject and is under a zero-shot setting. Following previous works(Chen et al. 2023; Zeng et al. 2024), we mainly use the data of subject 3 in GOD for comparison. We use preprocessed regions of interest (ROIs)[1] that cover voxels from early to higher visual areas. The GOD dataset provides both stimulus-evoked and

[1]Preprocessed data and demo code available at http://brainliner.jp/data/brainliner/Generic_Object_Decoding

Methods	Low- Level				High-Level				Retrieval		
	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓	Image↑	Text@5↑	Brain↑
Mind Reader(Lin, Sprague, and Singh 2022)	-	-	-	-	78.2%	-	-	-	11.0%	-	49.0%
Takagi et al.(2022)	-	-	83.0%	83.0%	76.0%	77.0%	-	-	-	-	-
Gu et al.(2023)	.150	.325	-	-	-	-	.862	.465	-	-	-
Brain-diffuser(Ozcelik and VanRullen 2023)	.254	.356	94.2%	96.2%	87.2%	91.5%	.775	.423	21.1%	-	30.3%
BrainCLIP(Liu et al. 2023)	-	-	-	-	86.7%	94.8%	-	-	40.7%	31.1%	-
fMRI-PTE (MG)(Qian et al. 2023)	.131	.112	78.1%	88.6%	84.1%	82.2%	.837	.434	-	-	-
DREAM(Xia et al. 2024)	.288	.338	93.9%	96.7%	93.7%	94.1%	.645	.418	-	-	-
MindBridge(Wang et al. 2024)	.151	.263	87.7%	95.5%	92.4%	94.7%	.712	.413	-	-	-
NeuroPictor(Huo et al. 2025)	.229	.375	**96.5%**	98.4%	94.5%	93.3%	.639	.350	-	-	-
PSYCHOMETRY(Quan et al. 2024)	.297	.340	96.4%	**98.6%**	**95.8%**	**96.8%**	.628	.345	-	-	-
MindEye(Scotti et al. 2023)	.309	.323	94.7%	97.8%	93.8%	94.1%	.645	.367	93.6%	-	90.1%
MindEye-BOI(Kneeland et al. 2023)	.259	.329	93.9%	97.7%	93.9%	93.9%	.645	.367	-	-	-
MindEye2(Scotti et al. 2024)	.322	**.431**	96.1%	**98.6%**	95.4%	93.0%	.619	.344	**98.8%**	-	**98.3%**
STTM(Ours)	**.333**	.334	95.7%	98.5%	**95.8%**	95.7%	**.611**	**.338**	92.8%	**41.3%**	94.9%
MindEye(High-level)(Scotti et al. 2023)	.194	**.308**	**91.7%**	97.4%	93.6%	94.2%	.645	.369	**93.6%**	-	90.1%
STTM-H(Ours)	**.209**	.276	91.5%	**97.8%**	**95.4%**	**95.6%**	**.612**	**.344**	92.8%	**41.3%**	**94.9%**
MindEye(Low-level)(Scotti et al. 2023)	.360	.479	78.1%	74.8%	58.7%	59.2%	1.0	.663	-	-	-
MindEye2(Low-level)(Scotti et al. 2024)	**.399**	**.539**	70.5%	65.1%	52.9%	57.2%	.984	.673	-	-	-
STTM-L(Ours,with guidance)	.383	.488	**83.3%**	**86.0%**	**68.2%**	**67.1%**	**.968**	**.647**	-	-	-

Table 1: Quantitative comparison of reconstruction and retrieval performance. All results are averaged across the same 4 participants, except Lin et al.(Lin, Sprague, and Singh 2022) which only analyzed Subject 1. We use the same evaluation metrics(See Appendix for metric details) and the same image preprocessing as Brain-diffuser(Ozcelik and VanRullen 2023) and MindEye(Scotti et al. 2023). Bold indicates best performance within sections.

imagery-induced fMRI data. Corresponding captions can be obtained from the GOD-Cap dataset(Liu et al. 2023). More details can be found in the Appendix.

Implementation Details Our models are trained and tested on 8 Hygon DCUs with 16GB HBM2 memory. Using data from 4 NSD subjects, we pre-train the high-level pipeline for 280 epochs and the low-level pipeline for 540 epochs, both with a global batch size of 192. For the high-level pipeline on the GOD dataset, we first train a new subject adapter for 4,500 epochs with a global batch size of 880, while keeping the pre-trained parts frozen. Afterward, we freeze the adapter and MLP backbone, and fine-tune the remaining parts for 400 epochs with a batch size of 600. Similarly, for the low-level pipeline, we train a new subject adapter for 5,000 epochs with a batch size of 192, then freeze the adapter and MLP backbone, and fine-tune the rest for 800 epochs.We optimize using AdamW(Loshchilov and Hutter 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. We apply the OneCycle learning rate schedule(Smith and Topin 2019) with a maximum learning rate of 0.0005. For reconstruction evaluation metrics, we use MindEye's implementation. Further details are available in our code.

Decoding Performance on NSD

FMRI-to-Image Reconstruction on NSD For each test brain sample, we generate 16 CLIP image embeddings using the diffusion prior and then process them through the image variations pipeline of Versatile Diffusion. We start the denoising process with the blurry reconstruction from our low-level pipeline, using 20 timesteps and UniPCMultistep noise scheduling(Zhao et al. 2024), resulting in 16 reconstructions per sample. We then select the best reconstruction using our retrieval branch. Figure 2 shows reconstruction examples from the NSD dataset. The low-level reconstructions capture details like position and color distribution, while the final reconstructions improve the semantic recognizability of the initial blurry images. Table 1 quantitatively compares our method with recent works on the NSD dataset. Our approach competes favoritely against recent works across various metrics, with STTM-L notably surpassing MindEye (Low-level) and MindEye2 (Low-level) on most metrics. Note that MindEye2 is pre-trained with data from 7 subjects, whereas STTM-L is trained with data from only 4 subjects who completed all sessions.

The img2img strength used for the results in Table 1 is 0.3, and we have tested with other values, obtaining robust performances (see Appendix).

Brain-Image/Text Retrieval on NSD For image and text retrieval, we aim to match the correct image or text to a given fMRI pattern from multiple candidates. In image retrieval, we compute the cosine similarity between the flattened and normalized 257 tokens of a brain sample and each of 300 randomly selected images from the test set. This is repeated for each of the 982 brain samples in the test set, and the overall accuracy is averaged across 30 iterations to account for batch sampling variability. In text retrieval, each fMRI pattern is matched against all 982 image captions by calculating the cosine similarity between the CLS token of the fMRI pattern and the CLS tokens of the candidate captions. For brain retrieval, we use a similar procedure to image retrieval but swap images and brain samples to find the corresponding brain sample for a given image among 300 brain samples.

Table 1 compares our method with previous works across these retrieval tasks, reporting top-1 accuracy for image and brain retrieval, and top-5 accuracy for text retrieval due to similar captions for some test images. Our method shows strong generalization across all retrieval tasks.

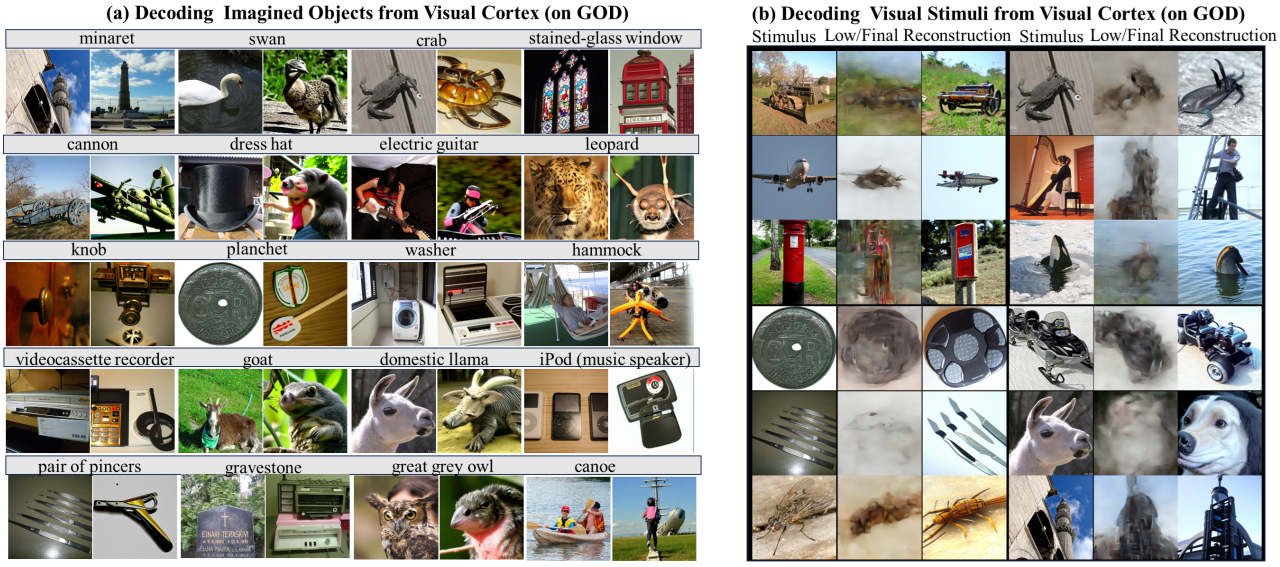

Figure 3: Reconstruction examples for mental imagery and visual stimuli. There is no ground truth for imagery data, we provide a reference image(left) for each imagery reconstruction(right) in sub-figure(a). For visual stimulus reconstruction, we display a pixel-wise reconstruction and a final reconstruction for each sample.

Methods	Low- Level				High-Level			
	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
IC-GAN(Sub3) (Ozcelik et al. 2022)	.195	.386	88.1%	95.3%	**85.8%**	84.0%	.855	.486
MinD-Vis(Sub3)(Chen et al. 2023)	.119	.390	82.8%	93.8%	81.0%	82.6%	.833	.491
CMVDM(Sub3)(Zeng et al. 2024)	.279	.454	88.4%	93.9%	81.6%	82.0%	**.810**	**.485**
STTM-H(Ours,Sub3)	.133	.328	87.9%	93.3%	79.7%	86.4%	.851	.521
STTM-L(Ours,Sub3)	**.322**	**.501**	90.3%	92.7%	66.8%	58.2%	.954	.704
STTM(Ours,Sub3)	.253	.367	**92.0%**	**95.6%**	81.3%	**87.0%**	.870	.505

Table 2: Evaluation of stimulus reconstruction on GOD. Our results are obtained by transfer learning. The results for Mind-Vis and CMVDM are calculated based on their reported reconstructions, while the results for IC-GAN are obtained by rerunning its model with provided weights.

Transfer Learning on GOD

We assess the transfer learning outcomes through two tasks: the zero-shot classification task and the image reconstruction task. We test with both stimulus-evoked and imagery-induced brain activities to assess whether our decoders, originally trained on brain activity induced by visual stimuli, possess the capability to generalize to decode imagery-induced brain activity.

Imagery & Stimulus Reconstruction on GOD In Figure 3, we present visualizations for mental imagery(a) and stimulus reconstructions(b). Note that the imagery data were gathered while subjects were freely imagining objects cued by text, with their eyes closed. Thus, there are no ground-truth images. We provide each category name with a reference image(on the left) and a generated image(on the right). As we can see, the reconstructed imagery and the reference images show a strong correlation in certain attributes, such as the decoded results accurately indicating whether the cor-

Methods	Prompt	top-1	top-5
CADA-VAE (Schonfeld et al. 2019)	-	17.7	53.3
MVAE (Wu and Goodman 2018)	-	17.1	52.5
MMVAE (Shi et al. 2019)	-	22.1	56.3
MoPoE-VAE (Sutter, Daunhawer, and Vogt 2021)	-	22.7	61.8
BraVL (Du et al. 2023)	-	24.0	62.1
BrainCLIP-VAE(Liu et al. 2023)	Text	20.0	58.0
BrainCLIP-VAE (Liu et al. 2023)	CoOp	21.33	64.7
STTM-H(Ours)	Text	**25.6**±1.96	**68.0**±1.79
STTM-H(imagery decoding)	Text	12.0±1.79	36.4±1.50

Table 3: Zero-shot classification results for the visual stimuli decoding(top section) and imagined objects decoding(bottom section). There are 50 classes in total, thus the chance levels for top-1 and top-5 accuracy are 2.0% and 10.0%, respectively. These results are for subject 3 of GOD.

Methods	Low- Level				High-Level			
	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
STTM-L(Sub1, I)	.417	.487	87.9%	87.8%	63.2%	62.8%	.980	.664
STTM-L(Sub1, II)	.422	.490	**89.1%**	**89.0%**	64.8%	63.6%	.974	.664
STTM-L(Sub1, III, w/o guidance)	.440	.498	84.5%	83.3%	64.6%	63.3%	.981	.639
STTM-L(Sub1, III)	**.445**	**.498**	87.4%	88.6%	**67.8%**	**68.2%**	**.964**	**.644**
STTM-H(Sub1, I)	.150	.266	83.6%	92.7%	91.4%	91.2%	.710	.392
STTM-H(sub1,III, w/o GVLC)	.214	.276	92.4%	98.3%	95.6%	95.6%	.606	.342
STTM-H(Sub1, III)	**.221**	**.278**	**92.6%**	**98.3%**	**96.0%**	**96.3%**	**.599**	**.332**

Table 4: Ablation studies were conducted on subject 1 of NSD. Condition I means both STTM-L and STTM-H are trained with single-subject data. Condition II means STTM-L is trained with single-subject data, while STTM-H uses cross-subject data. Condition III represents both models are trained with cross-subject data.

responding target category is animate or inanimate. Notably, our work is the first to generate high-resolution reconstructions of mental imagery.

We also quantitatively evaluate visual stimulus reconstruction performance in Table 2. Our low-level pipeline excels in PixCorr and SSIM, two pixel-wise metrics. Our final reconstruction performs well compared to previous state-of-the-art methods. Notably, Mind-Vis(Chen et al. 2023), which is also a pretraining-based decoding method, is outperformed by our method on most metrics. This suggests that our approach is a promising alternative for pre-training fMRI models.

Zero-Shot Classification on GOD by Prompting STTM leverages CLIP's well-aligned embedding space for zero-shot visual content classification by comparing the embeddings of the CLS token from Brain tokens with the classification weights generated by CLIP's text encoder. The text encoder uses textual prompts like "a photo of a [CLASS]," where [CLASS] is replaced by the specific class name. We use the same text prompts as BrainCLIP.

Table 3 presents results for visual stimulus and imagery classification. Decoders trained on visual stimulus-induced brain activity demonstrate the capacity to generalize to imagery-induced brain activity but with reduced accuracy. This indicates some transferability between the two types of brain activity but also highlights the unique challenges of decoding imagery-induced brain activity.

Ablations

In Table 4, we present ablation studies conducted on Subject 1's data from NSD, focusing on three types of experiments regarding the source of training data. In Type I, both STTM-L and STTM-H are trained with single-subject data. In Type II, STTM-L is trained with single-subject data, while STTM-H uses cross-subject data. In Type III, both models are trained with cross-subject data. We also assess the impact of high-level perception guidance on STTM-L and the effect of global vision-linguistic contrastive learning on STTM-H. These results lead to the following insights:

Cross-subject fMRI training enhances both STTM-L and STTM-H. A potential explanation is that training with cross-subject fMRI data improves model generalizability by exposing the shared decoding model to a more diverse data distribution, allowing it to extract essential fMRI features more effectively.

Interaction between high-level and low-level perception matters, and better high-level guidance leads to better low-level reconstructions. Integrating high-level perception guidance boosts the performance of our low-level pipeline across nearly all metrics. Better high-level guidance leads to a better STTM-L model. Additionally, combining both pipelines improves final reconstruction performance compared to using only the high-level pipeline, which effectively mimics the interaction between bottom-up and top-down processes in the human brain.

Global visual-linguistic contrastive learning slightly influences the reconstruction performance but provides multimodal decoding ability. While it has a modest impact on image reconstruction, global visual-linguistic contrastive learning equips the model with the ability to handle text-related tasks, offering a path for future multimodal brain decoding.

Conclusion

In this work, we focus on developing a robust brain-decoding model and its transferability to new subjects with limited training data. We propose an adapter-based cross-subject fMRI pretraining and transfer learning framework, highlighting the importance of high-level and low-level feature interaction for visual stimulus reconstruction. Our method shows promising results on both the NSD and GOD datasets. We also assess the transferability from stimulus-evoked to imagery-induced brain activity, generating high-resolution visualizations of mental imagery for the first time. Although with exciting results, our approach has its limitations, e.g., during pre-training, each subject requires a unique adapter, which restricts the number of subjects due to accelerator memory constraints, making it preferable to use subjects with more training samples.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NO. 62088102, NO.62076235), STI2030-Major Projects (NO. 2022ZD0208801), and China National Postdoctoral Program for Innovative Talents from China Postdoctoral Science Foundation (NO. BX2021239).

References

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. 2022. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1): 116–126.

Beliy, R.; Gaziv, G.; Hoogi, A.; Strappini, F.; Golan, T.; and Irani, M. 2019. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fMRI. *Advances in Neural Information Processing Systems*, 32.

Chen, P.-H. C.; Chen, J.; Yeshurun, Y.; Hasson, U.; Haxby, J.; and Ramadge, P. J. 2015. A reduced-dimension fMRI shared response model. *Advances in neural information processing systems*, 28.

Chen, Z.; Qing, J.; Xiang, T.; Yue, W. L.; and Zhou, J. H. 2023. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22710–22720.

Cox, D. D.; and Savoy, R. L. 2003. Functional magnetic resonance imaging (fMRI)"brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage*, 19(2): 261–270.

Dijkstra, N.; Mostert, P.; Lange, F. P. d.; Bosch, S.; and van Gerven, M. A. 2018. Differential temporal dynamics during visual imagery and perception. *Elife*, 7: e33904.

Du, C.; Fu, K.; Li, J.; and He, H. 2023. Decoding Visual Neural Representations by Multimodal Learning of Brain-Visual-Linguistic Features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ganis, G.; Thompson, W. L.; and Kosslyn, S. M. 2004. Brain areas underlying visual mental imagery and visual perception: an fMRI study. *Cognitive Brain Research*, 20(2): 226–241.

Gu, Z.; Jamison, K.; Kuceyeski, A.; and Sabuncu, M. 2023. Decoding natural image stimuli from fMRI data with a surface-based convolutional network. arXiv:2212.02409.

Güçlü, U.; and van Gerven, M. A. 2017. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, 145: 329–336.

Haxby, J. V.; Gobbini, M. I.; Furey, M. L.; Ishai, A.; Schouten, J. L.; and Pietrini, P. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539): 2425–2430.

Haynes, J.-D.; and Rees, G. 2005. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature neuroscience*, 8(5): 686–691.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Horikawa, T.; and Kamitani, Y. 2017. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1): 1–15.

Huo, J.; Wang, Y.; Wang, Y.; Qian, X.; Li, C.; Fu, Y.; and Feng, J. 2025. Neuropictor: Refining fmri-to-image reconstruction via multi-individual pretraining and multi-level modulation. In *European Conference on Computer Vision*, 56–73. Springer.

Kamitani, Y.; and Tong, F. 2005. Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5): 679–685.

Katsuki, F.; and Constantinidis, C. 2014. Bottom-up and top-down attention: different processes and overlapping neural systems. *The Neuroscientist*, 20(5): 509–521.

Khosla, M.; Ngo, G. H.; Jamison, K.; Kuceyeski, A.; and Sabuncu, M. R. 2020. A shared neural encoding model for the prediction of subject-specific fMRI response. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VII 23*, 539–548. Springer.

Kneeland, R.; Ojeda, J.; St-Yves, G.; and Naselaris, T. 2023. Brain-optimized inference improves reconstructions of fMRI brain activity. arXiv:2312.07705.

Kosslyn, S. M.; Pascual-Leone, A.; Felician, O.; Camposano, S.; Keenan, J. P.; Ganis, G.; Sukel, K.; and Alpert, N. 1999. The role of area 17 in visual imagery: convergent evidence from PET and rTMS. *Science*, 284(5411): 167–170.

Kosslyn, S. M.; Thompson, W. L.; and Alpert, N. M. 1997. Neural systems shared by visual imagery and visual perception: A positron emission tomography study. *Neuroimage*, 6(4): 320–334.

Lee, S.-H.; Kravitz, D. J.; and Baker, C. I. 2012. Disentangling visual imagery and perception of real-world objects. *Neuroimage*, 59(4): 4064–4073.

Liang, V. W.; Zhang, Y.; Kwon, Y.; Yeung, S.; and Zou, J. Y. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35: 17612–17625.

Lin, B.; Ye, F.; Zhang, Y.; and Tsang, I. W. 2022. Reasonable Effectiveness of Random Weighting: A Litmus Test for Multi-Task Learning. arXiv:2111.10603.

Lin, S.; Sprague, T. C.; and Singh, A. 2022. Mind Reader: Reconstructing complex images from brain activities. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Liu, Y.; Ma, Y.; Zhou, W.; Zhu, G.; and Zheng, N. 2023. BrainCLIP: Bridging Brain and Visual-Linguistic Representation Via CLIP for Generic Natural Visual Stimulus Decoding. arXiv:2302.12971.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Ma, Y.; He, Y.; Wang, H.; Wang, A.; Qi, C.; Cai, C.; Li, X.; Li, Z.; Shum, H.-Y.; Liu, W.; and Chen, Q. 2024. Follow-Your-Click: Open-domain Regional Image Animation via Short Prompts. arXiv:2403.08268.

Malkiel, I.; Rosenman, G.; Wolf, L.; and Hendler, T. 2022. Self-supervised transformers for fmri representation. In *International Conference on Medical Imaging with Deep Learning*, 895–913. PMLR.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. arXiv:2108.01073.

Miller, E. K. 1999. Straight from the top. *Nature*, 401(6754): 650–651.

Ozcelik, F.; Choksi, B.; Mozafari, M.; Reddy, L.; and VanRullen, R. 2022. Reconstruction of perceived images from fmri patterns and semantic brain exploration using instance-conditioned gans. In *2022 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Ozcelik, F.; and VanRullen, R. 2023. Natural scene reconstruction from fMRI signals using generative latent diffusion. arXiv:2303.05334.

Qian, X.; Wang, Y.; Huo, J.; Feng, J.; and Fu, Y. 2023. fMRI-PTE: A Large-scale fMRI Pretrained Transformer Encoder for Multi-Subject Brain Activity Decoding. arXiv:2311.00342.

Quan, R.; Wang, W.; Tian, Z.; Ma, F.; and Yang, Y. 2024. Psychometry: An omnifit model for image reconstruction from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 233–243.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125.

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

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Schonfeld, E.; Ebrahimi, S.; Sinha, S.; Darrell, T.; and Akata, Z. 2019. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8247–8255.

Scotti, P. S.; Banerjee, A.; Goode, J.; Shabalin, S.; Nguyen, A.; Cohen, E.; Dempster, A. J.; Verlinde, N.; Yundler, E.; Weisberg, D.; Norman, K. A.; and Abraham, T. M. 2023. Reconstructing the Mind's Eye: fMRI-to-Image with Contrastive Learning and Diffusion Priors. arXiv:2305.18274.

Scotti, P. S.; Tripathy, M.; Villanueva, C. K. T.; Kneeland, R.; Chen, T.; Narang, A.; Santhirasegaran, C.; Xu, J.; Naselaris, T.; Norman, K. A.; and Abraham, T. M. 2024. MindEye2: Shared-Subject Models Enable fMRI-To-Image With 1 Hour of Data. arXiv:2403.11207.

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

Shi, Y.; Paige, B.; Torr, P.; et al. 2019. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in Neural Information Processing Systems*, 32.

Smith, L. N.; and Topin, N. 2019. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, 369–386. SPIE.

Sutter, T. M.; Daunhawer, I.; and Vogt, J. E. 2021. Generalized Multimodal ELBO. arXiv:2105.02470.

Takagi, Y.; and Nishimoto, S. 2022. High-resolution image reconstruction with latent diffusion models from human brain activity. *bioRxiv*, 2022–11.

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

Wang, S.; Liu, S.; Tan, Z.; and Wang, X. 2024. Mindbridge: A cross-subject brain decoding framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11333–11342.

Wu, M.; and Goodman, N. 2018. Multimodal generative models for scalable weakly-supervised learning. *Advances in Neural Information Processing Systems*, 31.

Xia, W.; de Charette, R.; Oztireli, C.; and Xue, J.-H. 2024. Dream: Visual decoding from reversing human visual system. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 8226–8235.

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

Zeng, B.; Li, S.; Liu, X.; Gao, S.; Jiang, X.; Tang, X.; Hu, Y.; Liu, J.; and Zhang, B. 2024. Controllable mind visual diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6935–6943.

Zhao, W.; Bai, L.; Rao, Y.; Zhou, J.; and Lu, J. 2024. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36.