This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Modeling the Brain's Grammar: ROI-Guided fMRI Pretraining for Transferable and Interpretable Vision Decoding

Yulong Liu, Hua Xu, Yiyang Cai, Chunyang Jiang, Sirui Han*, Yike Guo
The Hong Kong University of Science and Technology
{yliuom,cjiangaq,ycaicj}@connect.ust.hk
{huaxu,siruihan,yikeguo}@ust.hk

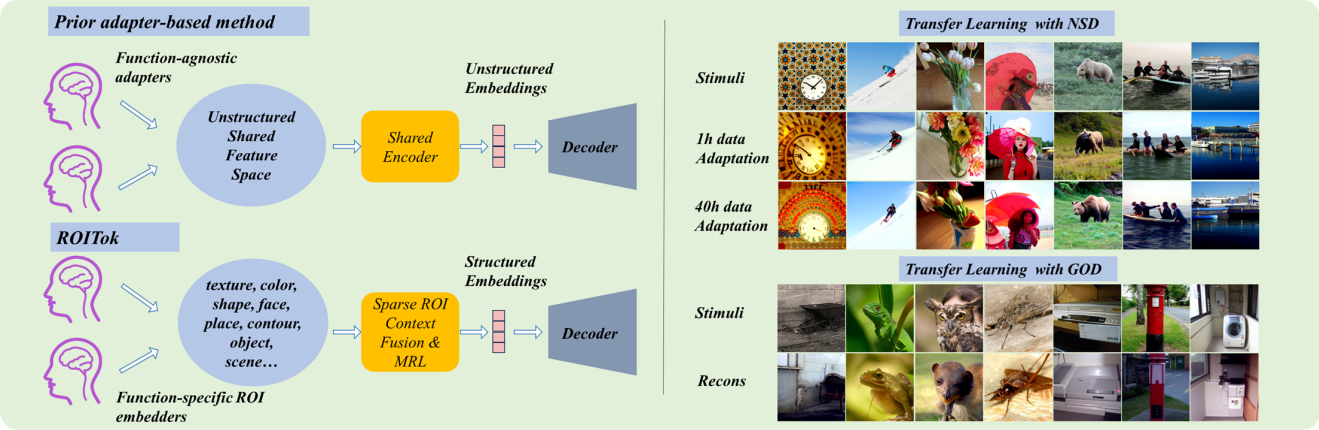


Figure 1. The left panel highlights how our method differs from prior adapter-based cross-subject alignment approaches: it integrates multi-ROI visual attribute processing and supports decoding from either a single or multiple ROIs. The right panel shows ROITok reconstruction examples, demonstrating robust performance even with limited training data.

Abstract

Recent advances in fMRI pretraining have significantly improved visual decoding accuracy by leveraging cross-subject neuroimaging datasets. A prevailing strategy aligns individual fMRI signals into a shared feature space using subject-specific adapters, followed by a shared decoder. However, this unstructured feature space overlooks the redundancy and functional correlations among voxels and fails to incorporate the brain's intrinsic functional architecture centered on regions of interest (ROIs). To address these limitations, we propose ***ROITok****, an ROI-guided fMRI pretraining framework. Our method introduces Sparse ROI Context Fusion to learn ROI-level visual representations and captures functional synergy between ROIs from cross-subject data. Inspired by Matryoshka Representation Learning (MRL), we design an embedding compression scheme that prioritizes the most informative visual components first, with later tokens adding progressively finer but still useful details. ROITok achieves strong transfer learning performance on the NSD and GOD datasets and shows strong resilience against high-level additive noises, while offering better interpretability. It allows for quantitative assessment of each brain region's contribution to decoding tasks. Our analysis shows that ROI-based pretraining can automatically learn the brain's visual hierarchy. Different ROIs can provide complementary contexts for decoding tasks; combining them improves decoding robustness.*

1. Introduction

Decoding human visual perception from functional magnetic resonance imaging (fMRI) signals has long been a central pursuit in cognitive neuroscience and brain–computer interface (BCI) research[4, 6–8, 11]. Recent advances in deep learning have demonstrated that pretraining shared decoding models on large-scale, cross-subject fMRI datasets can substantially improve accuracy[3, 5, 9, 20, 22]. These gains stem from the ability of data-driven models to discover common patterns across

*Corresponding Author

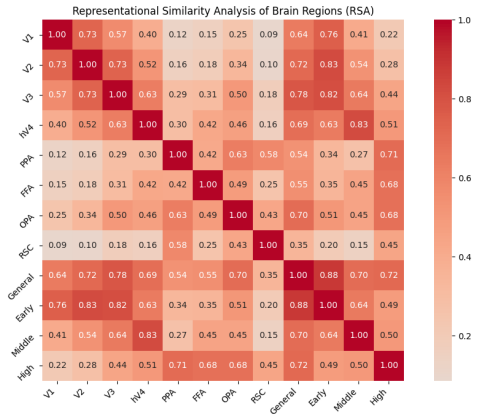


Figure 2. RSA between ROI Tokens of ROITok. Our method captures the hierarchical organization of visual information in the human brain.

individuals, thereby mitigating the well-known challenges of inter-subject anatomical and functional variability.

We summarize two dominant paradigms in cross-subject fMRI modeling here. The first treats fMRI volumes as 1D or 2D images and applies ViT-style spatial patchification to feed data into deep networks[3, 9, 26]. This strategy imposes arbitrary spatial discretizations that ignore the functional topology of the brain and are ill-suited to non-uniform resolution of BOLD signals. The second paradigm employs subject-specific adapters to project individual neural responses into a shared latent space before applying a common decoder[5, 20, 32]. This method overlooks the strong spatial redundancy and functional correlations that exist within and across brain regions. And its shared feature space lacks explicit structure constraints, which leads to suboptimal performance under few-shot scenarios.

Critically, both approaches overlook a core neuroscience insight: the human brain processes visual information through a network of functionally specialized, topographically organized regions, commonly referred to as Regions of Interest (ROIs). Neuroimaging studies consistently show that while voxel-level BOLD responses vary across individuals, ROI-level neural representations exhibit strong cross-subject alignment in both anatomy and function [10, 14, 15], suggesting that ROIs form a natural, biologically grounded basis for a shared representational space.

Motivated by this evidence, we propose **ROITok**, a physiologically informed framework that reorients cross-subject fMRI pretraining around ROIs as computational units. Rather than treating the brain as a single entity, ROITok tokenizes neural activity at the region level, treating each ROI as a contextual token that encapsulates its distributed, multi-voxel response pattern. As shown in Fig. 2, this design aligns model architecture with the brain's modular organization, enabling the learning of cross-subject "representational grammars" that reflect how visual information is hierarchically processed and integrated across the cortical visual stream.

As shown in Fig.3, ROITok introduces two techniques. First, we propose Sparse ROI Context Fusion strategy, which learns ROI-level visual representations and inter-regional functional synergy by reconstructing visual information from masked ROIs. This leverages the inherent redundancy and functional dependencies among brain regions, enhancing robustness without compromising interpretability. Second, inspired by Matryoshka representation learning(MRL)[13], we design an encoder with a embedding compression strategy, which ensures that the sequence of tokens is ordered by information saliency[35]. With ROITok, we compare different finetuning strategies and two approaches to conditional image generation, namely, the IP-Adapter-based method[42] and the diffusion-prior-based method[28], highlighting their respective advantages and limitations.

We evaluate ROITok on the Natural Scenes Dataset (NSD)[1] and the Generic Object Decoding (GOD) dataset[8], achieving competitive transfer performance under both full-data and few-shot cross-subject settings. ROITok also shows strong resilience against high-level additive noises. Beyond accuracy, our method offers enhanced interpretability and unlocks more potential applications in neural decoding. ROITok provides a way to compare each ROI's contribution to decoding tasks and their functional correlation. Our experiments demonstrate that different ROIs can provide complementary contexts for decoding tasks, which leads to better transfer learning outcomes. With its ability of integrating information across ROIs, ROITok can even decode edited ROI sequences into coherent, blended visual reconstructions(see Appendix 7). This showcases the potential of ROI-based models for controllable and interpretable neural decoding.

2. Related work

Vision decoding from brain activity has been greatly advanced by deep learning. For example, fMRI feature extractors have evolved from simple linear models[4, 8, 23, 37, 38] to deep networks[2, 9, 20, 29, 32–34], and image reconstruction has progressed from CNNs[2, 33, 34], GANs[16, 23, 25, 29] to diffusion models[5, 18, 19, 24, 31, 32]. Early approaches relied on subject-specific training, which is prone to overfitting due to fMRI's low signal-to-noise ratio and limited per-subject data. To mitigate this, recent vision decoding approaches have adopted a pretrain-then-finetune paradigm using cross-subject fMRI data to learn shared, transferable neural representations. For instance, Chen et al.[3] split 1D fMRI signals into patches and pretrain a transformer encoder via Masked Language Modeling (MLM) in a self-supervised manner. fMRI-PTE[26] and NeuroPictor[9] map fMRI signals onto unified 2D grids

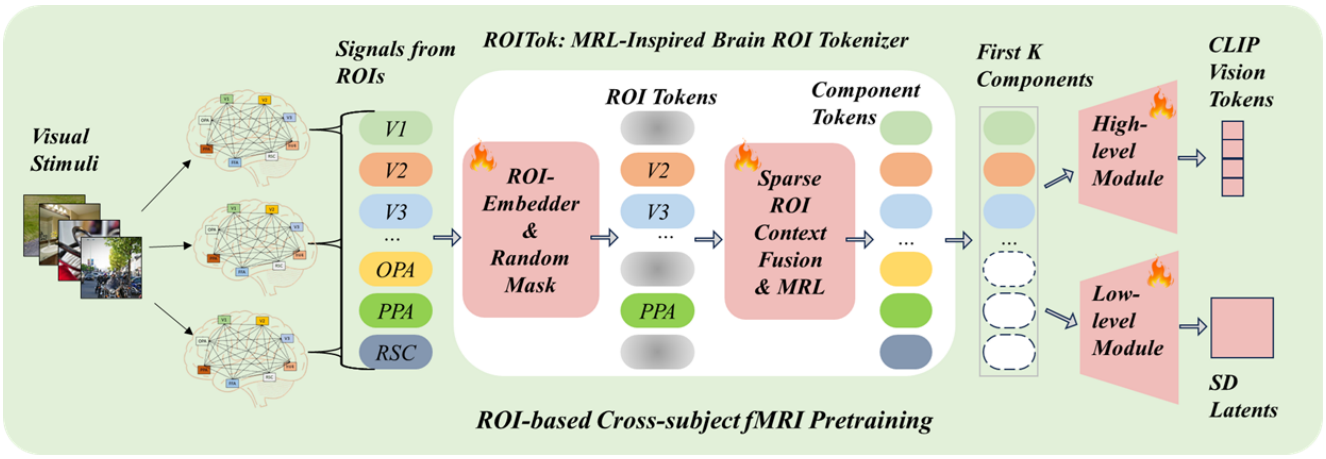


Figure 3. Cross-subject fMRI pretraining of ROITok. It captures inter-regional functional synergy from masked ROI sequences and employs a MRL-inspired[13] embedding compression strategy to prioritize the dominant modes of inter-subject visual representation.

using anatomical priors, enabling pretraining with ViT-like architectures. Meanwhile, MindEye2[32], STTM[20], and MindTuner[5] employ subject-specific adapters to project cross-subject fMRI into a shared feature space before applying a common decoder. These methods show the potential of foundation models for brain decoding but still underperform, especially with limited data. Patch-based approaches use arbitrary spatial grids that lack neuroscientific basis and ignore BOLD signal heterogeneity, while adapter-based methods align whole-brain activity without capturing fine-grained structure, missing voxel redundancy and functional correlations. Crucially, neither respects the brain's functional topology, limiting interpretability. In contrast, our ROI-based pretraining explicitly leverages the functional specificity of brain regions, achieving strong performance while maintaining neuroscientific interpretability.

3. Method

3.1. ROITok: MRL-inspired Brain ROI tokenizer

Fig. 3 illustrates the overall architecture of our fMRI pretraining framework. Below, we detail each component.

ROI-based pretraining Given a set of subjects $\{S_i\}_{i=1}^{P}$, each fMRI scan is first parcellated into M anatomically or functionally defined regions of interest (ROIs) using a standard atlas or a functional localizer (fLoc) [1]. These ROIs correspond to well-established brain areas that specialize in processing distinct types of visual information. For instance, early visual areas such as V1 and V2 encode low-level features like edges and orientations, while higher-order regions like the fusiform face area (FFA) respond selectively to complex stimuli such as faces. For each subject S_i, corresponding fMRI training data is defined as $\{R_i^j\}_{j=1}^{M}$, $R_i^j \in \mathbb{R}^{N\times d_{i,j}}$, where R_i^j is the training data for j-th ROI of S_i, N is the number of samples, and $d_{i,j}$ is the voxel number of the ROI.

Each ROI is processed by a specific linear ROI-Embeder $E_{i,j}$ into a ROI token with a shared dimension D, where $i \in [1, 2, ..., P], j \in [1, 2, ..., M]$. Since most ROI only contain several hundreds of voxels and D is set to 400 in our experiments, this multi-embedder setting will not cause serious computational burden.

Brain representation learning with Sparse ROI Context Fusion ROI token sequences provide an over-complete representation of decodable visual information. They contain substantial redundancy and also offer complementary cues in the presence of noise. To fully exploit these correlations, we propose **Sparse ROI Context Fusion**, a method that learns robust, cross-subject representations from masked ROI tokens. During pretraining, we randomly mask 50% of the ROI tokens and replace them with a learnable [MASK] token. The resulting ROI tokens are concatenated and fed into a Residual MLP encoder $\mathcal{E}$, composed of eight residual MLP blocks. We adopt the Residual MLP architecture as the backbone due to its significantly lower GPU memory consumption compared to attention-based models, as well as its stable training dynamics enabled by residual connections. The encoder outputs contextualized embeddings $\mathbf{Z} \in \mathbb{R}^{M'\times D'}$, which consists of M' components $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_{M'})$, where $M' \times D' = M \times D$. We use $D' = D$ in our main experiments, ablation study about the value of D' is provided in the appendix 8.2.

As different ROIs encode distinct levels of visual information and functional attributes, Sparse ROI Context Fusion strategy encourages the model to mimic the hierarchical organization of the human visual system and to effectively integrate information from distributed ROI contexts. This design also fully leverages the brain's inherent redun-

dancy, thereby enhancing cross-subject generalization.

MRL-inspired structured fMRI embeddings The output of the residual MLP encoder is high-dimensional. Prior works such as MindEye2[32], STTM[20], and MindTuner[5] impose no explicit structural constraints on this output space beyond their decoding objectives. We observe that this design is suboptimal: the resulting embedding space lacks structure and is highly susceptible to overfitting. To improve interpretability and robustness, we introduce a structural regularization mechanism that encourages the learned fMRI embeddings $\mathbf{Z} \in \mathbb{R}^{M' \times D'}$ to prioritize dominant modes of inter-subject visual representation. Inspired by Matryoshka representations[13], we apply the following function to the fMRI embeddings $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_{M'})$ during training:

$$F(\mathbf{Z}, m) = (\mathbf{Z}_1, \ldots, \mathbf{Z}_m, \mathbf{Z}_\emptyset, \ldots, \mathbf{Z}_\emptyset), \tag{1}$$

where m is a random integer uniformly sampled from $\mathcal{U}\{1, \ldots, M'\}$, and $\mathbf{Z}_\emptyset$ denotes the learnable empty token that replace masked positions. This masking strategy encourages the earlier tokens $\mathbf{Z}_1, \ldots, \mathbf{Z}_m$ to prioritize capturing the most informative visual features. As a result, the model learns a hierarchy among the component tokens.

Training objective Together with ROITok, a lightweight high-level module $\mathcal{D}_\mathcal{H}$ is trained to align fMRI components with CLIP vision tokens Z_{clip}. This module comprises two branches: a retrieval branch optimized with SoftCLIP loss[31] and an auxiliary semantic reconstruction branch trained with MSE loss. The high-level module is implemented as a stack of MLP layers. In parallel, a low-level module $\mathcal{D}_\mathcal{L}$ that is realized as a CNN-based upsampler, is trained with L1 loss to map fMRI components to Stable Diffusion[30] VAE latents Z_{sd} of the visual stimuli, producing latent images at a resolution of $4 \times 64 \times 64$. Further implementation details are provided in our code. Our pretraining optimizes the fMRI components $\mathbf{Z}$ via the following objective:

$$\mathcal{L} = \mathbb{E}_{m \sim \mathcal{U}\{1,\ldots,M'\}} \begin{bmatrix} \|\mathcal{D}_\mathcal{H}(F(\mathbf{Z}, \mathbf{m})) - \mathbf{Z_{clip}}\|_2^2 \\ + \mathbf{SoftCLIP}(\mathcal{D}_\mathcal{H}(F(\mathbf{Z}, \mathbf{m})), \mathbf{Z_{clip}}) \\ + \|\mathcal{D}_\mathcal{L}(F(\mathbf{Z}, \mathbf{m})) - \mathbf{Z_{sd}}\| \end{bmatrix}. \tag{2}$$

3.2. fMRI-conditioned Image Reconstruction

Decoupling the conditional image generation module from the ROITok pretraining stage enhances flexibility in downstream applications of brain-derived features and significantly reduces computational cost. For instance, while MindEye2 requires 8xA100 80Gb GPUs for pretraining, our approach completes pretraining on a single H800 80Gb GPU. Once the Brain ROI tokenizer is obtained, a connector is needed to bridge fMRI embeddings with a pretrained image generator. We compare two controllable image generation strategies: the Diffusion-Prior-based [28] and IP-Adapter-based method [42]. As illustrated in Fig. 4, their key distinction lies in how conditioning is implemented.

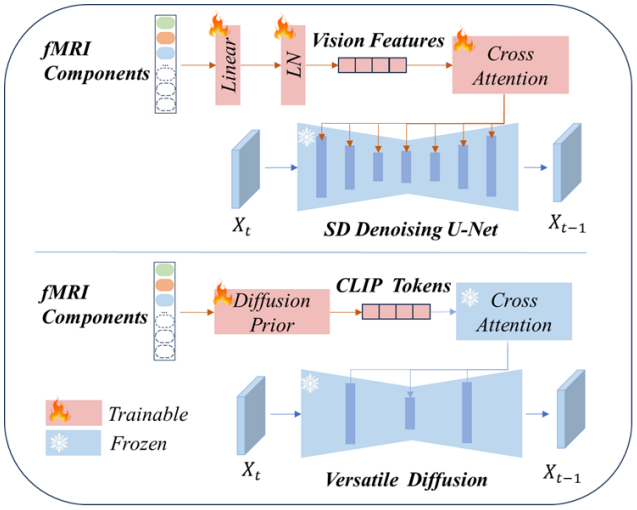


Figure 4. Two approaches to fMRI-to-image generation. The upper pipeline adapts the IP-Adapter framework by injecting fMRI-derived semantic tokens into the SD U-Net through trainable cross-attention layers. The lower pipeline uses a diffusion prior to predict CLIP vision tokens from fMRI, which subsequently condition a pretrained diffusion model for image synthesis.

In the Diffusion-Prior-based framework, following MindEye2 [32], a transformer-based diffusion model is trained to map fMRI features to aligned CLIP vision tokens. These tokens then serve as conditioning inputs to a frozen, pretrained generative model (e.g., Versatile Diffusion[41]) to synthesize images. In contrast, the IP-Adapter-based method employs a lightweight linear projection to convert fMRI components into semantic tokens, which are injected into the U-Net of Stable Diffusion through newly trained cross-attention layers to guide image generation. We will demonstrate that these two paradigms offer complementary advantages in terms of fidelity.

During inference, we adopt an image-to-image setting , where the predictions from the low-level module are used as the initial blur reconstructions , which can provide abundant structure and pixel-wise information. Then the blur reconstructions are refined with the pretrained diffusion model under the guidance of fMRI components. The Image-to-image strength is set to 0.75 in our experiments.

3.3. Transfer learning with limited samples

When data from a new subject are available, we first freeze all pretrained parameters and train only new ROI-embedders to extract ROI representations. Remarkably, this

minimal adaptation already yields strong decoding accuracy. Further fine-tuning all relevant parameters for a few steps consistently improves results. Our ROI-based modeling enables more stable cross-subject transfer by modeling ROI functional correlation as a universal "brain grammar."

4. Experiments

4.1. Data and Settings

NSD & GOD We employ the Natural Scenes Dataset (NSD) [1] for pretraining. Following the standard protocol used in prior studies [25, 32, 37], we split the data such that each subject has about 9,000 training images and 1,000 held-out test images. If not specially indicated, decoding performances are averaged across the four subjects (1, 2, 5, and 7) who completed all 40 scanning sessions. The dataset provides region-of-interest (ROI) masks; we utilize the widely used vision-related ROIs: V1, V2, V3, hV4, FFA, PPA, OPA, RSC, Early (early visual areas), Middle (middle ventral stream), High (high-level ventral stream), and General (all task-relevant voxels). To evaluate cross-dataset transferability, we adopt the Generic Object Decoding (GOD) dataset [8]. Following established practice [3, 20, 43], we compare with previous works on subject 3 using the preprocessed ROIs from Kamitani et al. [11]: V1, V2, V3, hV4, FFA, PPA, Early, High, and General.

Pretraining & Finetuning We adopt the pretraining protocol of MindEye2. Specifically, we pretrain on data from seven subjects (excluding the held-out test subject) for 80,000 optimization steps with a total batch size of 630. Subsequently, using extracted fMRI component tokens, we pretrain a diffusion prior module for 80,000 steps and an IP-Adapter for 200k steps. Finetuning on NSD follows a two-stage strategy: (1) new ROI-Embedders are trained for 5,000 steps with all other modules frozen, followed by (2) joint finetuning of all parameters for 5,000 steps. For transfer to GOD, we first pretrain a decoding model on all 8 NSD subjects and then apply a similar two-stage finetuning procedure. All training is conducted on a single NVIDIA H800 80Gb GPU, requiring approximately 30 hours for pretraining. Optimization uses AdamW [21] with weight decay of 0.1 for the ROI-Embedder and 0.01 for shared parameters. Learning rates follow a OneCycle schedule [36], peaking at 1×10^{-4} during pretraining and 5×10^{-5} during finetuning. Evaluation metrics (reconstruction and retrieval) are implemented following MindEye2.

4.2. Decoding Results

In Fig. 5, we present a qualitative comparison of our method against several recent approaches under the 1-hour data adaptation setting. Quantitative results for image reconstruction and 300-way retrieval on the NSD dataset(under both 40-hour and 1-hour training regimes), and cross-dataset transfer to GOD are summarized in Tab. 1.

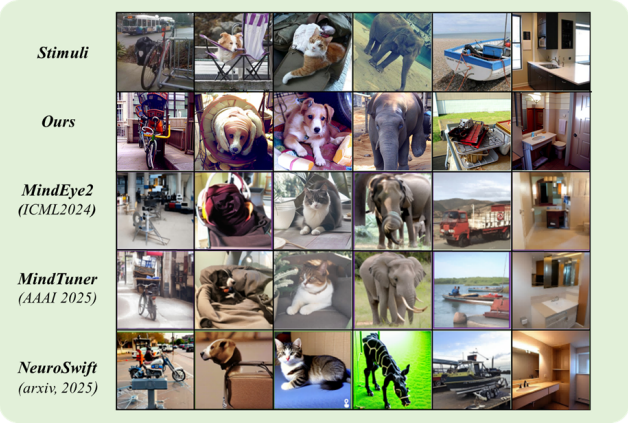


Figure 5. Visual comparison with recent approaches under the 1-hour data transfer learning setting.

Our method achieves competitive performance on both NSD and GOD, not only in data-rich scenarios but also under severely limited training data, demonstrating remarkable robustness. Notably, our ROI-based approach attains the highest low-level reconstruction fidelity across both datasets while maintaining high-level semantic performance on par with state-of-the-art methods. This suggests that our model effectively captures fine-grained visual features embedded in fMRI patterns, a capability we attribute to the functional synergy learned through our sparse ROI context fusion strategy.

Our results further reveal that IP-adapter–based methods inherently possess a higher upper bound for preserving low-level visual details compared to diffusion-prior–based approaches. However, due to the limited availability of fMRI–image paired data, current IP-adapter instantiations struggle to achieve high-level semantic fidelity, falling short of diffusion-prior methods in this regard. These results may provide references for future research about how to choose appropriate conditional generation methods. Moreover, we analyze the contribution of individual ROIs and find that integrating information across multiple regions improves decoding accuracy(See Section 4.3).

4.3. Ablation Study and Interpretability

The influence of ROI mask ratio We pretrain and finetune decoding models with different ROI mask ratios. The results are displayed in Fig 6, which shows that sparse ROIs context fusion can fully leverage redundancy and correlation between ROIs to improve decoding accuracy.

The MRL-like fMRI components In Tab. 2, we evaluate the impact of the proposed MRL-inspired embedding structure. The results show that this structured representation significantly enhances pixel-wise reconstruction and

Table 1. Quantitative decoding results for image reconstruction and retrieval on the NSD and GOD datasets are reported. Results are averaged across participants, except for Mind Reader[16], which was evaluated only on Subject 1. Results from related works are either taken directly from their respective publications or reproduced using the authors' released model weights. The "**Upper Bound**" denotes performance when both training and testing are conducted on the test set, providing an empirical estimate of the theoretical decoding ceiling for each method. Within each section, the best performance is highlighted in red, and the second-best in blue.

Methods	Low- Level				High-Level				Retrieval	
	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓	Image↑	Brain↑
Results on NSD										
Brain-diffuser[24]	.254	.356	94.2%	96.2%	87.2%	91.5%	.775	.423	21.1%	30.3%
BrainCLIP[19]	-	-	-	-	86.7%	94.8%	-	-	40.7%	-
DREAM[40]	.288	.338	93.9%	96.7%	93.7%	94.1%	.645	.418	-	-
MindBridge[39]	.151	.263	87.7%	95.5%	92.4%	94.7%	.712	.413	-	-
PSYCHOMETRY[27]	.297	.340	96.4%	98.6%	95.8%	96.8%	.628	.345	-	-
MindEye[31]	.309	.323	94.7%	97.8%	93.8%	94.1%	.645	.367	93.6%	90.1%
NeuroPictor[9]	.229	.375	96.5%	98.4%	94.5%	93.3%	.639	.350	-	-
STTM[20]	.333	.334	95.7%	98.5%	95.8%	95.7%	.611	.338	92.8%	94.9%
MindEye2[32]	.285	.389	96.3%	98.7%	95.4%	93.0%	.619	.343	91.7%	97.4%
MindTuner[5]	.322	.421	95.8%	98.8%	95.6%	93.8%	.612	.340	98.9%	98.3%
NeuroSwift[44]	.335	.437	96.5%	98.1%	95.4%	97.1%	.615	.354	-	-
ROITok(IP-Adapter)	.475	.426	97.6%	98.7%	94.7%	92.9%	.640	.388	98.7%	98.7%
ROITok(Diffusion Prior)	.470	.351	98.0%	99.0%	95.7%	95.2%	.605	.336	98.7%	98.7%
Upper Bound(IP-Adapter)	.751	.468	100.0%	100.0%	99.8%	99.5%	.410	.237	100.0%	100.0%
Upper Bound(Diffusion Prior)	.743	.401	100.0%	100.0%	99.6%	99.6%	.440	.234	100.0%	100.0%
Upper Bound(Low-level)	.867	.592	99.6%	98.3%	72.4%	73.2%	.940	.658	-	-
MindEye(Low-level)[31]	.360	.479	78.1%	74.8%	58.7%	59.2%	1.0	.663	-	-
MindEye2(Low-level)[32]	.399	.539	70.5%	65.1%	52.9%	57.2%	.984	.673	-	-
STTM-L[20]	.383	.488	83.3%	86.0%	68.2%	67.1%	.968	.647	-	-
ROITok(Low-level)	.549	.514	89.6%	83.0%	60.6%	61.9%	.990	.647	-	-
MindEye2(1h)[32]	.195	.419	84.2%	90.6%	81.2%	79.2%	.810	.468	79.0%	57.4%
MindTuner(1h)[5]	.224	.420	87.8%	93.6%	84.8%	83.5%	.780	.440	83.1%	76.0%
NeuroSwift(1h)[44]	.253	.431	90.7%	94.1%	88.6%	85.9%	.791	.453	-	-
ROITok(1h, IP-Adapter)	.341	.421	91.2%	94.7%	87.4%	84.0%	.779	.484	86.9%	85.5%
ROITok(1h, Diffusion Prior)	.303	.318	92.0%	95.0%	88.1%	87.6%	.738	.424	86.9%	85.5%
Results on GOD										
IC-GAN [25]	.195	.386	88.1%	95.3%	85.8%	84.0%	.855	.486	-	-
MinD-Vis[3]	.119	.390	82.8%	93.8%	81.0%	82.6%	.833	.491	-	-
CMVDM[43]	.279	.454	88.4%	93.9%	81.6%	82.0%	.810	.485	-	-
STTM[20]	.253	.367	92.0%	95.6%	81.3%	87.0%	.870	.505	-	-
ROITok(IP-Adapter)	.321	.422	90.5%	92.9%	76.0%	82.6%	.867	.534	-	-
ROITok(Diffusion Prior)	.308	.381	93.6%	96.8%	82.7%	83.4%	.848	.507	-	-

also yields consistent improvements in retrieval tasks. Furthermore, Fig. 7 illustrates the contribution of each component tokens, which shows that earlier tokens encode substantially more visually informative content than later ones.

The influence of our two-stage finetuning strategy In Tab. 2, we compare different fine-tuning strategies using 1-hour NSD data. As the results show, if we directly fintune all parameters, the performances are significantly worse. The optimal approach is to first train the ROI embedders while keeping the pretrained modules frozen, followed by full fine-tuning with all parameters unfrozen.

Table 2. Ablation studies on the MRL-like embedding architecture and fine-tuning strategies, with results averaged across subjects.

Methods	Low- Level Recons				Retrieval	
	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Image↑	Brain↑
ROITok(40h,w/o MRL-like)	.426	.493	85.8%	81.8%	97.3%	98.2%
ROITok(40h,MRL-like)	.549	.514	89.6%	83.0%	98.7%	98.7%
ROITok(1h, single stage)	.239	.478	63.9%	62.4%	66.1%	71.9%
ROITok(1h,stage 1)	.337	.488	71.3%	66.9%	82.1%	81.4%
ROITok(1h, stage 2)	.365	.489	75.8%	71.1%	86.9%	85.5%

How robust fMRI decoding is against noises In Fig. 8, we evaluate the 1000-way retrieval accuracy of our method

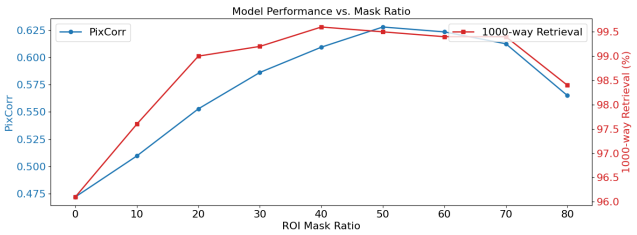


Figure 6. Ablation study on the ROI masking ratio during training using 40-hour data from subject 1 of NSD. 1000-way image retrieval and low-level reconstruction performances are provided.

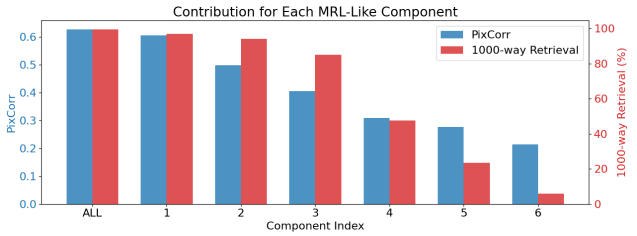


Figure 7. Contribution of each MRL-like fMRI components for vision decoding on NSD(40h data, subject 1). We only show the first 6 components.

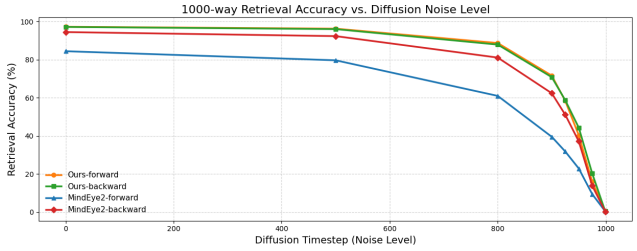


Figure 8. The average accuracy of 1000-way retrieval under different noise levels. The diffusion process undergoes a cosine schedule with $\beta_{max} = 0.5$. Note that when $T = 900, SNR = 0.182$ (against 48488 for $T = 0$), even though the fMRI patterns are badly noised, they still contain a lot of visual information.

and MindEye2 under increasing levels of synthetic noise. Following a cosine noise schedule (with maximum $\beta_t = 0.5$), we progressively corrupt test fMRI patterns with additive Gaussian noise. Remarkably, both models remain robust: even at $T = 900$ (signal-to-noise ratio ≈ 0.182), they recover meaningful information. This suggests that multivoxel fMRI representations encode information redundantly across large-scale spatial configurations, making them resilient to severe signal degradation. We note that the added noise is *synthetic* and does not model physiological noise sources (e.g., cardiac or motion artifacts). Nevertheless, the observed robustness implies that decoding systems may tolerate realistic fMRI noise to some extent, particularly when it manifests as reduced SNR. This insight might inform future fMRI data augmentation strategies.

How different ROIs contribute to vision decoding In Fig. 9, we quantitatively assess the contribution of different ROIs to decoding performance. The results show that early visual areas alone already provide substantial information. Notably, for 1000-way retrieval, combining "V1–hV4' and "Early" tokens outperforms using either "Early" alone or "V1–hV4" tokens (87.9% vs. 83.5% and 86.4%, in 40h setting). Similarly, using all ROIs yields higher accuracy than relying solely on the "General" token (97.3% vs. 96.2%). Accuracy consistently increases with the number of ROIs included, similar trend appears in both reconstruction and retrieval performance, suggesting that separating and reintegrating ROI-specific information enhances decoding robustness. We further test ROITok on synthetic fMRI patterns: for each target ROI, we generate an artificial activation map by sampling from $\mathcal{N}(2, 0.5^2)$ and masking all other regions. As shown in Fig. 9(c), reconstructions align semantically with each ROI's known function: FFA produces faces; OPA evokes local scenes (e.g., kitchen corners); PPA generates large-scale environments (e.g., hotel lobbies); and RSC produces navigable corridors. Reconstructions from early visual areas (V1–hV4) exhibit a hierarchical progression, from edges and textures to complex semantic content. These results confirm ROITok's biologically plausible decoding behavior and highlight its potential for probing neural representations.

Representational similarity between ROIs Our method provides a framework for interpreting the brain's visual organization. Using ROI tokens extracted from NSD test set, we perform representational similarity analysis(RSA) [12], shown in Fig. 2. The similarity matrix reveals a clear hierarchy: V1–V3 are highly similar to each other but show low similarity with higher-order regions like PPA, FFA, RSC. Notably, similarity with "High" increases along the ventral stream, from 0.22 in V1 to 0.51 in hV4, reflecting a progressive abstraction of visual representations, consistent with the established view of increasing invariance and semantic specificity along the visual hierarchy.

Probing Cross-Subject Alignability of ROIs We assessed the cross-subject alignability of ROIs via RSA on ROI tokens from the NSD test set, under two conditions (within the same decoding model and across different models) with RSA values averaged over four test subjects' pairwise combinations. As shown in Fig. 10, the "General" token has the highest within-model alignability, while the high-level ventral stream shows the strongest cross-model consistency. Notably, higher-order category-selective regions (PPA, FFA, OPA) consistently show greater cross-subject alignability than early visual areas (V1–V3) under both conditions, in line with their level of semantic abstraction. In addition, ROI size (number of voxels) can

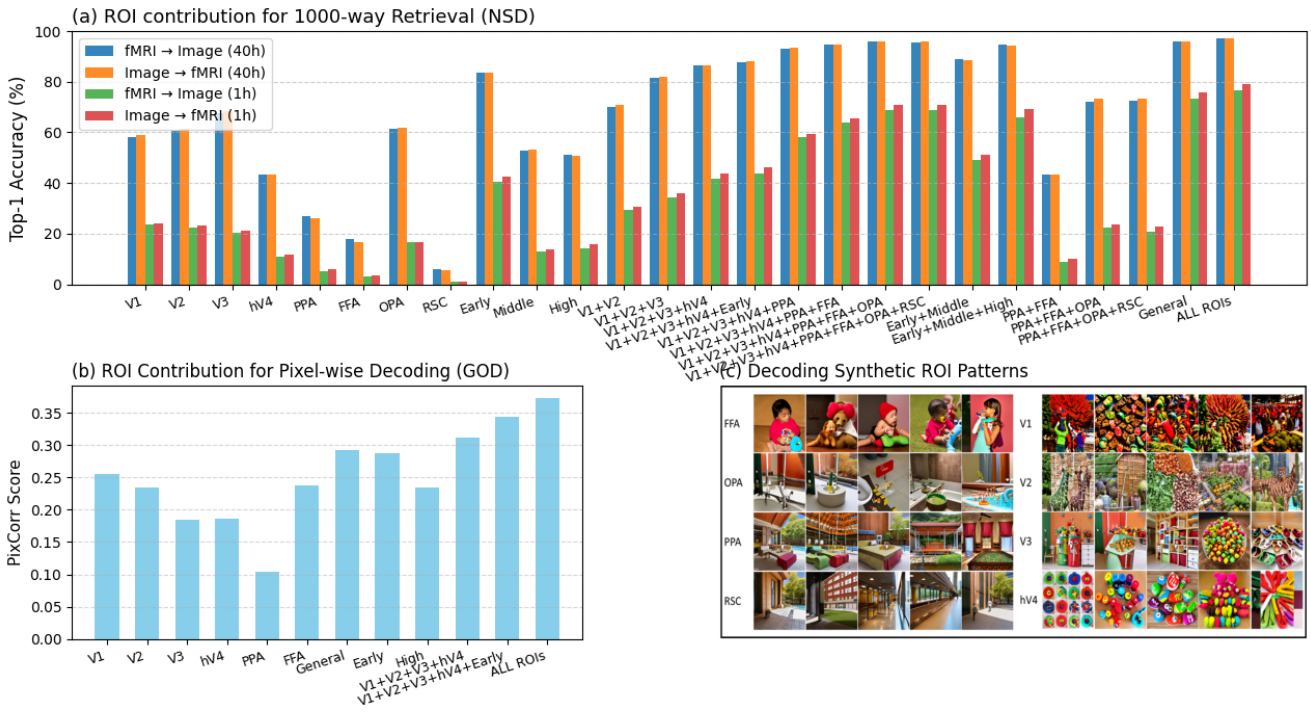


Figure 9. ROI-based decoding performance analysis. When a single ROI or a group of ROIs is selected, all other regions are masked. (a) evaluate the contribution of each ROI and some combination of ROIs on 1000-way retrieval task. Results are averaged across subjects. (b) evaluate the contribution of each ROI on low-level reconstruction. c) Examples for decoding with synthetic fMRI patterns. For each output image, only the target ROI is activated and sampled from $\mathcal{N}(2, 0.5^2)$. These results reveal that ROITok successfully reflect the functional specificity of each brain region and that different ROIs provide complementary contexts for decoding tasks.

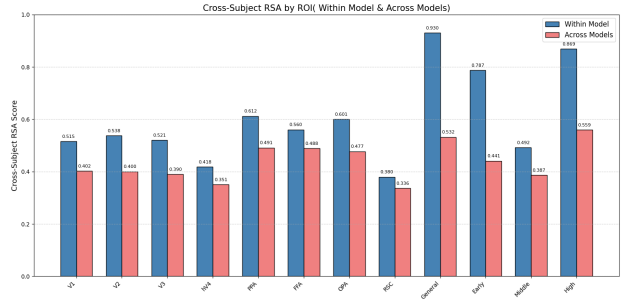


Figure 10. We employ representational similarity analysis(RSA) to quantify the cross-subject alignability of ROIs.

also modulate alignability: the "Early" ROI (combining V1-V3) shows greater cross-subject alignability than each individual area in both settings, suggesting that aggregating responses across a large, functionally coherent voxel population stabilizes representational geometry by mitigating idiosyncratic noise or anatomical variability. However, greater alignability does not always correlate with better decoding performance: PPA, FFA, and "High" ROI tokens, despite strong cross-subject alignment, have lower image retrieval accuracy than V1-V3. This dissociation indicates higher-order regions may lack fine-grained visual detail for high-fidelity image decoding. Finally, this RSA only quantifies alignability under linear transformations; non-linear correspondences between subjects' neural representations may exist but remain undetected.

5. Conclusion

We propose ROITok, a neuroscientifically grounded framework that constructs a shared feature space structured by the functional specificity of brain ROIs. By integrating sparse ROI context fusion and an MRL-inspired embedding compression strategy, ROITok captures inter-regional functional correlations and hierarchically organizes representations according to information saliency. Evaluated on the NSD and GOD benchmarks, ROITok achieves state-of-the-art cross-subject transfer performance in both data-rich and few-shot settings, while offering high interpretability. This alignment of AI representations with the brain's functional organization paves the way for more explainable and generalizable brain–computer interfaces, as well as structured, region-level neural analysis. Further, our work and prior studies suggest the possible existence of an optimal shared feature space that can be aligned across subjects using simple transformations. ROITok provides a promising direction, but its performances are not perfect yet. In future work, we will continue to enhance this shared space.

Acknowledgement

This work was supported in part by the HKUST Start-up Fund (R9911), the Theme-based Research Scheme (T45-205/21-N), the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government, and the research grant from the HKUST-DXM AI for Finance Joint Laboratory (DXM25EG01).

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, 3, 5

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

[3] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22710–22720, 2023. 1, 2, 5, 6

[4] 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. 1, 2

[5] Zixuan Gong, Qi Zhang, Guangyin Bao, Lei Zhu, Rongtao Xu, Ke Liu, Liang Hu, and Duoqian Miao. Mindtuner: Cross-subject visual decoding with visual fingerprint and semantic correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14247–14255, 2025. 1, 2, 3, 4, 6

[6] James V Haxby, M Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001. 1

[7] John-Dylan Haynes and Geraint Rees. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature neuroscience*, 8(5):686–691, 2005.

[8] Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):1–15, 2017. 1, 2, 5, 14

[9] 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 *European Conference on Computer Vision*, pages 56–73. Springer, 2025. 1, 2, 6

[10] Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016. 2

[11] Yukiyasu Kamitani and Frank Tong. Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5):679–685, 2005. 1, 5

[12] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008. 7

[13] Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249, 2022. 2, 3, 4, 11

[14] Juha M Lahnakoski, Enrico Glerean, Iiro P Jääskeläinen, Jukka Hyönä, Riitta Hari, Mikko Sams, and Lauri Nummenmaa. Synchronous brain activity across individuals underlies shared psychological perspectives. *NeuroImage*, 100:316–324, 2014. 2

[15] Daniel D Leeds, Darren A Seibert, John A Pyles, and Michael J Tarr. Comparing visual representations across human fmri and computational vision. *Journal of vision*, 13 (13):25–25, 2013. 2

[16] Sikun Lin, Thomas Christopher Sprague, and Ambuj Singh. Mind reader: Reconstructing complex images from brain activities. In *Advances in Neural Information Processing Systems*, 2022. 2, 6

[17] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 13

[18] Aoyang Liu, Haodong Jing, Yulong Liu, Yongqiang Ma, and Nanning Zheng. Hidden states in llms improve eeg representation learning and visual decoding. In *ECAI*, pages 2130–7, 2024. 2

[19] Yulong Liu, Yongqiang Ma, Wei Zhou, Guibo Zhu, and Nanning Zheng. Brainclip: Bridging brain and visual-linguistic representation via clip for generic natural visual stimulus decoding, 2023. 2, 6

[20] Yulong Liu, Yongqiang Ma, Guibo Zhu, Haodong Jing, and Nanning Zheng. See through their minds: Learning transferable brain decoding models from cross-subject fmri. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5730–5738, 2025. 1, 2, 3, 4, 5, 6

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 5

[22] Itzik Malkiel, Gony Rosenman, Lior Wolf, and Talma Hendler. Self-supervised transformers for fmri representation. In *International Conference on Medical Imaging with Deep Learning*, pages 895–913. PMLR, 2022. 1

[23] Milad Mozafari, Leila Reddy, and Rufin VanRullen. Reconstructing natural scenes from fmri patterns using bigbigan. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2020. 2

[24] Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion, 2023. 2, 6

[25] Furkan Ozcelik, Bhavin Choksi, Milad Mozafari, Leila Reddy, and Rufin VanRullen. Reconstruction of perceived images from fmri patterns and semantic brain exploration using instance-conditioned gans. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022. 2, 5, 6

[26] Xuelin Qian, Yun Wang, Jingyang Huo, Jianfeng Feng, and Yanwei Fu. fmri-pte: A large-scale fmri pretrained transformer encoder for multi-subject brain activity decoding, 2023. 2

[27] Ruijie Quan, Wenguan Wang, Zhibo Tian, Fan Ma, and Yi Yang. Psychometry: An omnifit model for image reconstruction from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 233–243, 2024. 6

[28] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 2, 4

[29] Ziqi Ren, Jie Li, Xuetong Xue, Xin Li, Fan Yang, Zhicheng Jiao, and Xinbo Gao. Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning. *NeuroImage*, 228:117602, 2021. 2

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 4

[31] Paul S. Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J. Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth A. Norman, and Tanishq Mathew Abraham. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors, 2023. 2, 4, 6, 13

[32] Paul S. Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq Mathew Abraham. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data, 2024. 2, 3, 4, 5, 6

[33] Guohua Shen, Kshitij Dwivedi, Kei Majima, Tomoyasu Horikawa, and Yukiyasu Kamitani. End-to-end deep image reconstruction from human brain activity. *Frontiers in Computational Neuroscience*, page 21, 2019. 2

[34] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1):e1006633, 2019. 2

[35] Jonathon Shlens. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*, 2014. 2

[36] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, pages 369–386. SPIE, 2019. 5

[37] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. *bioRxiv*, pages 2022–11, 2022. 2, 5

[38] 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

[39] Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Mindbridge: A cross-subject brain decoding framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11333–11342, 2024. 6

[40] Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. Dream: Visual decoding from reversing human visual system. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8226–8235, 2024. 6

[41] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7754–7765, 2023. 4

[42] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 4

[43] Bohan Zeng, Shanglin Li, Xuhui Liu, Sicheng Gao, Xiaolong Jiang, Xu Tang, Yao Hu, Jianzhuang Liu, and Baochang Zhang. Controllable mind visual diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6935–6943, 2024. 5, 6

[44] Shiyi Zhang, Dong Liang, and Yihang Zhou. Neuroswift: A lightweight cross-subject framework for fmri visual reconstruction of complex scenes. *arXiv preprint arXiv:2510.02266*, 2025. 6

[45] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 13