SEEING THROUGH THE BRAIN: NEW INSIGHTS FROM DECODING VISUAL STIMULI WITH FMRI

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding how the brain encodes visual information is a central challenge in neuroscience and machine learning. A promising approach is to reconstruct visual stimuli—essentially images—from functional Magnetic Resonance Imaging (fMRI) signals. This involves two stages: transforming fMRI signals into a latent space and then using a pre-trained generative model to reconstruct images. The reconstruction quality depends on how similar the latent space is to the structure of neural activity and how well the generative model produces images from that space. Yet, it remains unclear which type of latent space best supports this transformation and how it should be organized to represent visual stimuli effectively.

We present two key findings. First, fMRI signals are more similar to the text space of a language model than to either a vision-based space or a joint text–image space. Second, text representations and the generative model should be adapted to capture the compositional nature of visual stimuli, including objects, their detailed attributes, and relationships. Building on these insights, we propose **PRISM**, a model that **P**rojects fM**RI** s**I**gnals into a **S**tructured text space as an inter**M**ediate representation for visual stimuli reconstruction. It includes an object-centric diffusion module that generates images by composing individual objects to reduce object detection errors, and an attribute–relationship search module that automatically identifies key attributes and relationships that best aligne with the neural activity. Extensive experiments on real-world datasets demonstrate that our framework outperforms existing methods, achieving up to an 8% reduction in perceptual loss. These results highlight the importance of using structured text as the intermediate space to bridge fMRI signals and image reconstruction.

1 INTRODUCTION

Decoding visual stimuli from brain activity provides a unique lens into human perception (Naselaris et al., 2011; Haufe et al., 2014). A central approach uses fMRI signals—which measure neural activity through blood-oxygen-level-dependent responses—to reconstruct the images perceived by subjects (Allen et al., 2022; Chang et al., 2019; Luo et al., 2023). Recent advances in deep generative models have significantly improved these reconstructions, deepening our understanding of visual representation in the brain (Chen et al., 2023) and enabling applications in brain-computer interfaces (Sitaram et al., 2008) and brain-driven content generation (Wang et al., 2024a; Qiu et al., 2025).

FMRI-to-Image reconstruction involves two stages: mapping fMRI signals into a latent space and then generating images from that space. Its success depends on the similarity between the latent space and neural activity (alignment) and how well the generative model produces high-quality images. While recent studies (Scotti et al., 2023; 2024; Mai et al., 2024) focus on enhancing image quality using advanced generative models (Podell et al., 2023; Xu et al., 2023), alignment remains underexplored. Prior work often assumes that the latent space should match the modality of the stimuli, i.e., using vision model representations to reconstruct visual stimuli (Scotti et al., 2023; Wang et al., 2024b; Xia et al., 2024). Some studies incorporate auxiliary semantic information from language models (LMs) (Lin et al., 2022; Quan et al., 2024), but still rely on vision-based representations as the core latent space. In contrast, we question whether matching the modality of visual stimuli is truly essential for reconstruction. In addition, prior work suffers from limited reconstruction quality due to a unified hidden representation that conflates objects and their attributes, often causing object detection errors, e.g., generating a tiger instead of a gray, tiger-striped cat (Appendix A). This

reflects a fundamental mismatch with human visual processing, which is object-centric and compositional rather than holistic (Marr, 1980; Bracci & Op de Beeck, 2023). Overcoming this limitation calls for generative models that explicitly capture the compositional structure of human perception.

To address these issues, we propose **PRISM**, a model that **P**rojects fM**RI** s**I**gnals into a **S**tructured text space as an inter**M**ediate representation for image reconstruction. To identify the most effective intermediate space, we compare fMRI signals with representations from pre-trained vision, language, and vision–language models using established metrics (Wang et al., 2020; Murphy et al., 2024; Keskar et al., 2016). Unlike prior work assuming vision-based representations are essential, our first finding (Section 3.1) shows that fMRI signals align more closely with the text space of LM, motivating the use of solely text as a bridge for reconstruction. Building on this, our second finding reveals that reconstruction quality improves when the text and the generative model are adapted to capture the compositional and relational nature of visual stimuli—encompassing objects, their attributes, and their relationships. Guided by these insights, we develop two core modules: an object-centric diffusion module that adapts the diffusion model to generate images by composing individual objects, and an attribute–relationship search module that uses a vision–language model (VLM) to automatically identify object attributes and relationships aligned with neural activity, providing structured guidance for reconstruction. Our contributions are summarized as follows:

- **Novel Findings:** To our knowledge, we are the first to show that accurate visual stimuli reconstruction can be achieved without image-based latent representations, with LM text space effectively bridging brain activity and generative models. Furthermore, we find that adapting this text space and the generative model to capture the compositional and relational nature of visual images further improves reconstruction quality.
- **Novel Framework:** Motivated by our empirical findings, we introduce a new fMRI-to-image reconstruction framework that adapts diffusion models for object-centric generation and leverages VLMs to automatically identify brain-aligned object attributes and relationships that can optimally guide the reconstruction.
- **Comprehensive Experiments:** Extensive evaluations on real-world fMRI datasets demonstrate that our method achieves up to an 8% reduction in perceptual loss compared to state-of-the-art models, highlighting the effectiveness of our framework.

2 PRELIMINARY

Notations. In our work, we denote the set of fMRI samples collected during image viewing as $\mathcal{X}$. Each sample is a preprocessed 1D vector $\mathbf{x}_i \in \mathbb{R}^v$, capturing neural activity across v voxels selected from brain regions (Scotti et al., 2023). The dataset is split into training and test subsets, with superscripts indicating the split. For example, we denote the training set with N samples as: $\mathcal{X}^{\text{train}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. The corresponding image stimulus for the i-th sample is denoted as: $\mathbf{Y}_i \in \mathbb{R}^{H\times W\times 3}$, which contains m objects.

Problem Setup. Our goal is to reconstruct the visual images that subjects viewed during fMRI recording. Formally, we seek to learn a reconstruction function $\mathcal{F} : \mathbb{R}^v \to \mathbb{R}^{H\times W\times 3}$ that maps each fMRI sample $\mathbf{x}_i$ to its corresponding image stimulus $\mathbf{Y}_i$.

Diffusion Model. Diffusion models (Rombach et al., 2022; Zhang et al., 2023) are a class of generative models that synthesize data by learning to reverse a multi-step noising process. Starting from Gaussian noise, they iteratively denoise a latent variable over a fixed number of timesteps using a denoising network—typically a U-Net—conditioned on the current timestep t. This process gradually produces samples resembling the training distribution. To incorporate external inputs such as text, the denoising network can take a conditioning input $\mathbf{C}$, usually text embeddings from a pre-trained encoder (Zhang et al., 2023). Conditioning is implemented via cross-attention mechanisms within the U-Net architecture (Williams et al., 2023), enabling the integration of textual information. In these layers, the latent representations $\mathbf{H}_t \in \mathbb{R}^{h\times w\times d}$ at time t serve as queries, and $\mathbf{C} \in \mathbb{R}^{d_t\times d'}$ serves as both keys and values (Williams et al., 2023; Yang et al., 2024):

$$\text{CrossAttention}(\mathbf{H}_t, \mathbf{C}) = \text{softmax}\left(\frac{\phi(\mathbf{H}_t)\cdot\mathbf{W}_Q\cdot(\varphi(\mathbf{C})\cdot\mathbf{W}_K)^\top}{\sqrt{d_k}}\right)\varphi(\mathbf{C})\cdot\mathbf{W}_V,$$

where $\mathbf{W}_Q \in \mathbb{R}^{d\times d_k}$, $\mathbf{W}_K \in \mathbb{R}^{d'\times d_k}$, $\mathbf{W}_V \in \mathbb{R}^{d'\times d}$ are projection matrices; and $\phi(\cdot)$ and $\varphi(\cdot)$ are learned transformations. Further details are available in Section B.

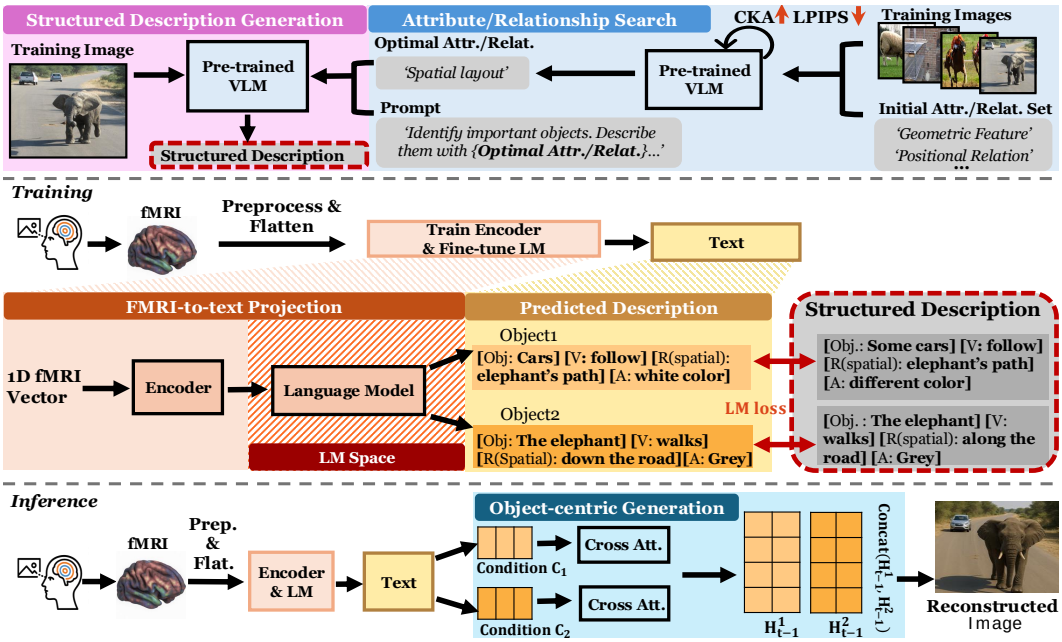

Figure 1: Framework Overview: **PRISM** generates structured text descriptions for each training image using a VLM to iteratively extract brain-aligned object attributes and relationships. These descriptions capture the image's compositional and relational content and serve as supervision to train an encoder and fine-tune a language model to map fMRI signals into the text space. During inference, the model predicts descriptions from fMRI signals, which then guide a pre-trained diffusion model for object-centric image reconstruction.

3 METHOD

In this section, we present our framework, **PRISM**, for fMRI-to-image reconstruction (Figure 1). We first show that fMRI signals align most strongly with the text space of LMs, compared to the hidden spaces of vision or vision–language models, under established metrics (Section 3.1). This finding motivates our choice of using pure text as the latent space. During training (Section 3.2), we annotate each training image with structured text descriptions that are object-centric, compositional, and relational. To generate these descriptions, we introduce an attribute–relationship search module (Section 3.2.1), which learns optimal prompts to guide the VLM in automatically identifying the key attributes and relationships most aligned with both the fMRI signals and images. These structured descriptions are then used to train an encoder and fine-tune the LM, mapping fMRI signals into the LM text space (Section 3.2.2). At inference time (Section 3.3), the predicted structured descriptions guide an adapted diffusion model to generate object-centric images directly from fMRI signals.

3.1 TEXT AS THE LATENT SPACE

We question whether using vision representations as the latent space is truly essential for reconstructing visual stimuli. In this section, we investigate the alignment between different model spaces and fMRI signals using various measures.

Measuring the alignment between model spaces and fMRI signals. We examine three representation spaces: (1) the text space of language models, (2) the joint text-image space of vision-language models, and (3) the latent space of vision models. For (2) and (3), image embeddings are extracted directly from the respective models. For (1), we use text embeddings from image captions to represent the stimuli. We extract embeddings by feeding either text or images into different models: T5 and LLaMA3 for text embeddings, LDM (Rombach et al., 2022) and ResNet50 (He et al., 2016) for image embeddings, and CLIP for both modalities.

Alignment is assessed using three metrics: Centered Kernel Alignment (CKA) (Murphy et al., 2024), Canonical Correlation Analysis (CCA) (Wang et al., 2020), and Generalization Gap (Keskar et al., 2016). CKA and CCA are widely used to quantify similarity between representation spaces

(Kriegeskorte et al., 2008; Wang et al., 2020). Generalization Gap reflects learnability by measuring the train-test loss difference when mapping fMRI signals to a target space using an MLP. Good alignment yields higher CKA and CCA values and a lower Generalization Gap.

Let $\mathbf{X} = \text{Concat}(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ denote concatenated fMRI samples and $\mathbf{K} = \text{Concat}(\mathbf{k}_1, \ldots, \mathbf{k}_N)$ the corresponding latent representations. With $\mathcal{K}(\cdot)$ as a kernel function, the empirical Hilbert-Schmidt Independence Criterion (HSIC) is: $\text{HSIC}(\mathbf{X}, \mathbf{K}) = \frac{1}{(N-1)^2} \operatorname{tr}\left(\mathcal{K}(\mathbf{X}) \cdot \mathcal{K}(\mathbf{K})\right)$, where $\operatorname{tr}(\cdot)$ denotes the trace operator. The CKA is the normalized form: $\text{CKA}(\mathbf{X}, \mathbf{K}) = \frac{\text{HSIC}(\mathbf{X},\mathbf{K})}{\sqrt{\text{HSIC}(\mathbf{X},\mathbf{X}) \cdot \text{HSIC}(\mathbf{K},\mathbf{K})}}$. We adopt a Gaussian radial basis function (RBF) kernel for $\mathcal{K}$ (Alvarez, 2022; Cortes et al., 2012). CCA identifies linear projections $\mathbf{u} = \mathbf{p}_1^\top \mathbf{X}$ and $\mathbf{v} = \mathbf{p}_2^\top \mathbf{K}$ that maximize their correlation. The first mode captures the dominant shared axis (Wang et al., 2020), with the canonical correlation coefficient: $\rho = \text{corr}(\mathbf{u}, \mathbf{v}) = \text{corr}(\mathbf{p}_1^\top \mathbf{X}, \mathbf{p}_2^\top \mathbf{K})$, reflecting the strongest linear alignment between brain activity and the model space.

FMRI aligns better with the embedding space of language model. Our results (Table 1) show that the text space of language models aligns best with fMRI data, outperforming both vision–language and vision-only models across all metrics. Surprisingly, vision–language models, despite integrating both modalities, underperform compared to pure language models. We hypothesize that this is because humans focus more on the meaning of an image rather than pixel-level details (Naselaris et al., 2009; Du et al., 2022). Unlike prior work (Scotti et al., 2023; Wang et al., 2024b; Xia et al., 2024; Lin et al., 2022) that primarily relies on vision representations, our findings motivate using pure text as the latent space.

Table 1: Alignment results between model representations and fMRI data, evaluated using CKA, Generalization Gap, and CCA. The best result is highlighted in red. ↑ denotes higher is better; ↓ denotes lower is better.

	CKA ↑	Generalization Gap ↓	CCA ↑
T5	0.5580	0.1132	0.8344
Llama3	0.5442	0.2216	0.8022
Clip text	0.5177	0.4532	0.7599
Clip img	0.3668	0.4860	0.7573
LDM	0.1957	1.2520	0.7215
Resnet50	0.1822	1.9800	0.6746

3.2 Training of **PRISM**

In this section, we describe the training process of **PRISM**, which consists of automatic structured description generation for training images and encoder training.

3.2.1 Automatic Description Generation

We design structured text descriptions as supervision for our framework. To capture the compositional and relational nature of human vision, these descriptions should explicitly distinguish between different objects and their relationships. Generating such descriptions with a VLM relies on carefully crafted prompts that specify the desired attributes and relations, since not all attributes and relationships are directly reflected in brain activity. To address this issue, we propose a VLM-assisted approach that automatically learns the most relevant attributes and relationships in an image based on the training data, ensuring they are both meaningful and brain-aligned.

We first show how structured descriptions can be generated from a VLM given a learned keyword a, and then present our approach for learning the optimal keyword. Given an image $\mathbf{Y}_i$ and a learned keyword a, we construct a prompt $\mathcal{P}(a)$ to guide the VLM in describing the most important objects in $\mathbf{Y}_i$ based on a. Formally, the VLM receives the image and the prompt as input and outputs a structured description D_i^a:

$$D_i^a = \text{VLM}(\mathbf{Y}_i, \mathcal{P}(a)). \tag{1}$$

The structured description is a list of m object-level tuples along with background information:

$$D_i^a = \left[(o_1 : d_1 : \text{loc}_1),\ (o_2 : d_2 : \text{loc}_2),\ \ldots,\ (o_m : d_m : \text{loc}_m), bg_i\right]. \tag{2}$$

Each o_j is an object in the image, d_j is its description containing attributes and relationships with other objects conditioned on keyword a, and loc_j denotes its location (selected from a predefined set). The term bg_i represents the background information of image $\mathbf{Y}_i$. To ensure meaningful generation, we further augment each relation description d_j with a structured header encoding its semantic roles, following the PropBank annotation format (Màrquez et al., 2008; He et al., 2017; Ross et al., 2021; Palmer et al., 2005).

The choice of keyword a strongly influences the attributes and relationships captured in the object descriptions, which in turn affects the quality of mapping fMRI signals to the text space. Ideally, these descriptions should capture the most important information shared between the fMRI signals and the stimulus images. To avoid manually selecting the keyword, we frame its discovery as a prompt optimization problem and introduce our attribute–relationship search module.

Concretely, given a set of training images $\mathcal{Y}^{\text{train}} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_N\}$ and the corresponding fMRI signals $\mathcal{X}^{\text{train}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, we define the following optimization problem to find the optimal a in the prompt ($\mathcal{P}(a)$) for the VLM:

$$\begin{aligned} \max_{a} \quad & \sum_{i=1}^{N} \mathcal{S}\left(\mathbf{Y}_i, \text{Diff}(\text{VLM}(\mathbf{Y}_i, \mathcal{P}(a)))\right) \\ \text{s.t.} \quad & \text{CKA}\left(\mathbf{X}, \mathbf{K}^a\right) > \beta; \\ & \mathbf{X} = \text{Concat}(\mathbf{x}_1, \ldots, \mathbf{x}_N);\ \mathbf{K}^a = \text{Concat}(\mathbf{k}_1^a, \ldots, \mathbf{k}_N^a); \\ & \mathbf{k}_i^a = \text{LM}_{\text{ENC}}(\text{VLM}(\mathbf{Y}_i, \mathcal{P}(a))) \text{ for } i = 1, \ldots, N; \end{aligned} \tag{3}$$

where $\mathcal{S}(\cdot, \cdot)$ denotes the similarity score between two images (e.g. negative perceptual loss); Diff is a pre-trained diffusion model that generates images from captions produced by the VLM; Concat indicates the concatenation operation across all training samples; and LM_{ENC} is a pre-trained language model to encode captions generated by the VLM. The constraint enforces that the CKA similarity between the fMRI data $\mathbf{X}$ and the caption embeddings $\mathbf{K}^a$ generated using keyword a exceeds a threshold β, ensuring strong alignment between the fMRI and text spaces. The objective ensures that descriptions derived from the optimal keyword support accurate reconstruction.

To optimize the keyword a in equation 3, we guide the search along semantic links: keywords with similar meanings tend to yield comparable reconstructions, so generating new keywords based on the semantic relationships of top-performing candidates helps uncover more effective prompt expressions. In the search, we utilize an LLM as a keywords generator and iteratively search for improved relation keywords in a step-by-step manner. We begin by initializing a keywords set $\mathcal{A}$ with a collection of frequently-used relation keywords identified in prior works (Johnson et al., 2015; Lu et al., 2016; Krishna et al., 2017). We expand $\mathcal{A}$ through an ε-greedy search strategy: at each search step, the attribute generator proposes new candidate keywords based on either the top-performing keywords in $\mathcal{A}$ with probability $1-\varepsilon$, or randomly selected keywords from $\mathcal{A}$ with probability ε. Only candidates that exceed the similarity threshold are added to $\mathcal{A}$. This balances refinement of effective attributes and exploration of diverse novel attributes. See Appendix G for the detailed algorithm and search results.

3.2.2 Encoder Training

We design an encoder to map fMRI signals into the latent space of the language model, using structured and object-centric descriptions as supervision. Specifically, each object's information is independently encoded using an MLP. The resulting representations are concatenated and passed to the language model to generate estimated structured descriptions $\hat{D}_i^a$, which is given by:

$$\begin{aligned} \mathbf{f}_j &= \text{MLP}_j(\mathbf{x}_i),\ j = 1, \cdots, m \\ \hat{D}_i^a &= \text{LM}(\text{MLP}_g(\text{Concat}(\mathbf{f}_1, \ldots, \mathbf{f}_m))), \end{aligned} \tag{4}$$

The language model is fine-tuned using a loss over all m object descriptions (Chang et al., 2024; Gunel et al., 2020):

$$\mathcal{L}_{\text{LM}} = -\sum_{j=1}^{m} \sum_{t'=1}^{T} \log p(y_{t'} \mid y_{<t'}, \mathbf{f}_j), \tag{5}$$

where $y_{t'}$ denotes the t'-th token in the structured description. This training strategy enables fine-grained alignment between fMRI signals and structured textual descriptions. We first train the MLPs independently for a fixed number of epochs, then jointly fine-tune the language model and MLPs to maximize overall reconstruction performance.

3.3 **PRISM** Inference: Object-Centric Image Generation

The inference process of **PRISM** has two steps: (1) generate structured descriptions $\hat{D}_i^a$ from fMRI signals using the trained encoder and language model, and (2) reconstruct the image by composing objects conditioned on these descriptions with a pre-trained diffusion model: $\hat{\mathbf{Y}}_i = \text{Diff}(\hat{D}_i^a)$.

To reflect the brain's compositional understanding of visual scenes, during inference, we adapted a pre-trained diffusion model to perform compositional image generation, inspired by (Yang et al., 2024). Specifically, given an image $\mathbf{Y}_i$ and its predicted structured description $\hat{D}_i^a$ (from Equation (4)), we extract a set of predicted objects $\{o_j\}_{j=1}^m$ and a background description $\hat{bg}_i$. These are combined into a global context prompt $\hat{p}_0$ and embedded into a conditioning matrix $\mathbf{C}_0$. Similarly, each object description $\hat{d}_j$ is embedded into a corresponding conditioning matrix $\mathbf{C}_j$. These conditioning matrices guide the denoising process via cross-attention (Section 2) from time t to 0. At each step, the hidden representation of object j (or the global context when $j = 0$) at time $t - 1$ is computed as: $\mathbf{H}_{t-1}^j = \text{CrossAttention}(\mathbf{H}_t, \mathbf{C}_j)$, where $\mathbf{H}_t$ is the hidden representation of the full image at time t. We then resize and concatenate the object representations according to their predicted locations $\hat{\text{loc}}_j$:

$$\mathbf{H}_{t-1}^{\text{cat}} = \Psi(\{\mathbf{H}_{t-1}^j, \hat{\text{loc}}_j\}_{j=1}^m), \tag{6}$$

where $\Psi(\cdot)$ denotes the resizing and spatially-aware concatenation operation. Thus, the image generation model encodes each object independently from its description and then spatially concatenates their hidden representations according to the predicted locations.

To ensure smooth region boundaries and seamless fusion between objects and background, we compute a weighted sum of the global context latent and the object latents:

$$\mathbf{H}_{t-1} = \beta \cdot \mathbf{H}_{t-1}^{\text{cat}} + (1 - \beta) \cdot \mathbf{H}_{t-1}^0, \tag{7}$$

where β is a hyperparameter that controls the blending ratio. This process is repeated across denoising steps, enabling structured, object-aware generation aligned with the brain's visual understanding.

4 Experiments

In this section, we conduct extensive experiments to evaluate **PRISM**, guided by the following questions: (**RQ1**) How well does our framework **PRISM** perform on the image reconstruction task? (**RQ2**) How do different choices of latent space influence the reconstruction quality? (**RQ3**) What is the contribution of each component in our framework to the overall reconstruction performance?

4.1 Experimental Setup

We conduct experiments on three datasets: NSD (Allen et al., 2022), BOLD5000 (Chang et al., 2019), and GOD (Horikawa & Kamitani, 2017). Detailed descriptions of the datasets are provided in Appendix C. Each method is evaluated by comparing the reconstructed images to the ground truth using three metrics: Pixelwise Correlation (PixCorr), Structural Similarity Index (SSIM) (Wang et al., 2004), Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018), which reflects human perceptual similarity, CLIP two-way identification (CLIP) Scotti et al. (2024) and Inception V3 two-way identification (Inception V3) Scotti et al. (2024). We compare our method against the following baselines: Takagi & Nishimoto (Takagi & Nishimoto, 2023) (Takagi for short), Mindvis (Chen et al., 2023), Mindeye (Scotti et al., 2023), and Mindeye2 (Scotti et al., 2024). To ensure a fair comparison, we use the same generative model, Stable Diffusion 2.1 (Pernias et al., 2023; Rombach et al., 2022), for all methods. We additionally present results for **PRISM** and Mindeye2 (ranked second-best) with the newer SDXL backbone (Podell et al., 2023). More details about the baselines and training are provided in Appendix D and Appendix E.

4.2 Effectiveness of **PRISM**

We evaluate our fMRI-to-image reconstruction framework on test data and compare it with state-of-the-art methods. Table 2 summarizes the results, with all metrics and standard deviations averaged across subjects over five runs. Visualizations of reconstructed examples from the test set are shown in Figure 2. As shown in Table 2, **PRISM** outperforms state-of-the-art methods across all datasets and metrics, with up to a 17% improvement in LPIPS, indicating higher perceptual similarity to the original images. Unlike baselines such as Mindeye2 (Scotti et al., 2024) and Mindeye1 (Scotti et al., 2023), which often ignore key objects, **PRISM** successfully reconstructs all objects, yielding notable gains in PixCorr, SSIM, and LPIPS. These results demonstrate the effectiveness of our framework in translating brain activity into accurate, perceptually aligned image reconstructions.

Table 2: Comparison of our framework with state-of-the-art methods on three datasets. All methods use Stable Diffusion 2.1 as the backbone unless otherwise specified (+SDXL). Results are reported using PixCorr, SSIM, LPIPS, CLIP and Inception V3 metrics. The best result using the same backbone in each column is highlighted in red. ↑ indicates higher is better and ↓ indicates lower is better.

NSD	PixCorr ↑	SSIM ↑	LPIPS ↓	CLIP ↑	Inception V3 ↑
PRISM	$0.3404_{\pm 0.05}$	$0.4640_{\pm 0.02}$	$0.5963_{\pm 0.02}$	$0.9467_{\pm 0.03}$	$0.9516_{\pm 0.03}$
Takagi	$0.2100_{\pm 0.01}$	$0.388_{\pm 0.04}$	$0.7665_{\pm 0.04}$	$0.8811_{\pm 0.06}$	$0.9086_{\pm 0.07}$
Mindvis	$0.2736_{\pm 0.06}$	$0.3868_{\pm 0.06}$	$0.6789_{\pm 0.02}$	$0.9000_{\pm 0.05}$	$0.9135_{\pm 0.05}$
Mindeye1	$0.3114_{\pm 0.05}$	$0.3868_{\pm 0.06}$	$0.6501_{\pm 0.03}$	$0.9121_{\pm 0.04}$	$0.9198_{\pm 0.03}$
Mindeye2	$0.3160_{\pm 0.04}$	$0.4447_{\pm 0.02}$	$0.6338_{\pm 0.04}$	$0.9201_{\pm 0.03}$	$0.9308_{\pm 0.03}$
PRISM+SDXL	$0.3645_{\pm 0.02}$	$0.4983_{\pm 0.04}$	$0.5563_{\pm 0.02}$	$0.9600_{\pm 0.01}$	$0.9765_{\pm 0.01}$
Mde2+SDXL	$0.3471_{\pm 0.04}$	$0.4425_{\pm 0.04}$	$0.6002_{\pm 0.01}$	$0.9599_{\pm 0.02}$	$0.9602_{\pm 0.01}$
BOLD5000					
PRISM	$0.2315_{\pm 0.01}$	$0.5341_{\pm 0.02}$	$0.6198_{\pm 0.02}$	$0.7720_{\pm 0.03}$	$0.6601_{\pm 0.07}$
Takagi	$0.1815_{\pm 0.03}$	$0.4418_{\pm 0.06}$	$0.7558_{\pm 0.06}$	$0.6990_{\pm 0.04}$	$0.5667_{\pm 0.01}$
Mindvis	$0.2122_{\pm 0.05}$	$0.4944_{\pm 0.04}$	$0.6463_{\pm 0.05}$	$0.7720_{\pm 0.04}$	$0.5701_{\pm 0.08}$
Mindeye1	$0.1942_{\pm 0.01}$	$0.4838_{\pm 0.03}$	$0.6913_{\pm 0.04}$	$0.7288_{\pm 0.03}$	$0.6222_{\pm 0.07}$
Mindeye2	$0.2265_{\pm 0.02}$	$0.5164_{\pm 0.02}$	$0.6416_{\pm 0.03}$	$0.7600_{\pm 0.04}$	$0.6428_{\pm 0.03}$
PRISM+SDXL	$0.2442_{\pm 0.02}$	$0.5600_{\pm 0.03}$	$0.5909_{\pm 0.04}$	$0.7881_{\pm 0.04}$	$0.6881_{\pm 0.02}$
Mde2+SDXL	$0.2310_{\pm 0.04}$	$0.5185_{\pm 0.02}$	$0.6186_{\pm 0.02}$	$0.7503_{\pm 0.04}$	$0.6556_{\pm 0.02}$
GOD					
PRISM	$0.2571_{\pm 0.01}$	$0.5200_{\pm 0.02}$	$0.6213_{\pm 0.01}$	$0.8567_{\pm 0.05}$	$0.8428_{\pm 0.06}$
Takagi	$0.2322_{\pm 0.05}$	$0.4944_{\pm 0.04}$	$0.6463_{\pm 0.05}$	$0.7232_{\pm 0.01}$	$0.7556_{\pm 0.02}$
Mindvis	$0.1921_{\pm 0.02}$	$0.4304_{\pm 0.03}$	$0.697_{\pm 0.04}$	$0.7162_{\pm 0.03}$	$0.6119_{\pm 0.03}$
Mindeye1	$0.2286_{\pm 0.03}$	$0.4766_{\pm 0.02}$	$0.6807_{\pm 0.05}$	$0.8093_{\pm 0.01}$	$0.8002_{\pm 0.04}$
Mindeye2	$0.2442_{\pm 0.03}$	$0.4952_{\pm 0.01}$	$0.6586_{\pm 0.02}$	$0.8322_{\pm 0.02}$	$0.8280_{\pm 0.04}$
PRISM+SDXL	$0.2669_{\pm 0.01}$	$0.5537_{\pm 0.01}$	$0.5989_{\pm 0.03}$	$0.8727_{\pm 0.03}$	$0.8820_{\pm 0.04}$
Mde2+SDXL	$0.2500_{\pm 0.04}$	$0.5511_{\pm 0.04}$	$0.6224_{\pm 0.01}$	$0.8678_{\pm 0.02}$	$0.8556_{\pm 0.01}$

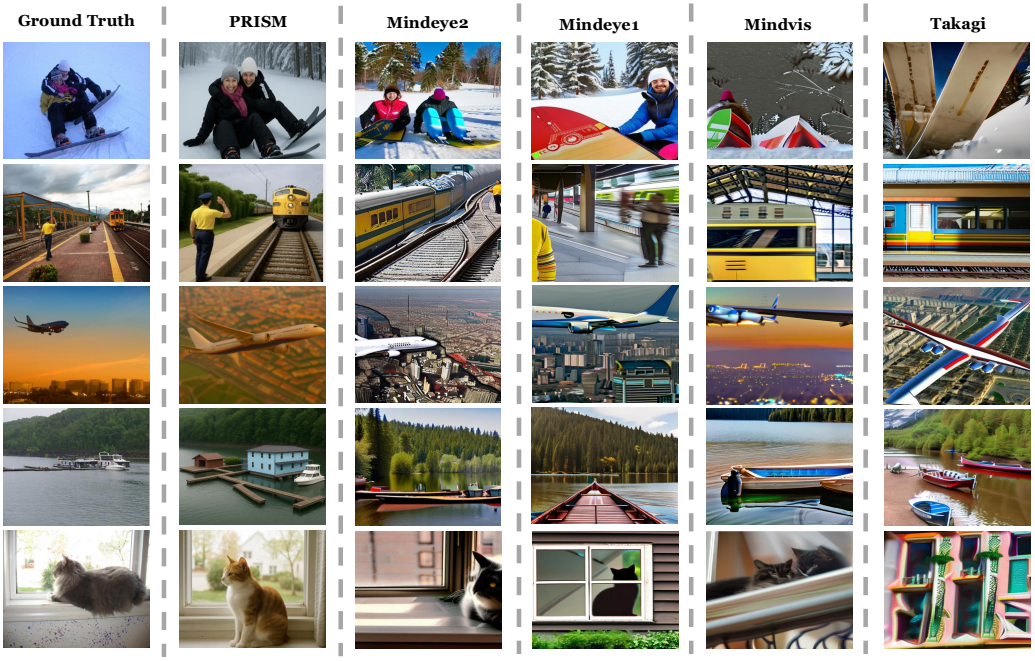

Figure 2: Reconstructed images from different methods. The first column shows the original viewed images. The rest of the columns show the reconstructed images from different methods.

We further evaluate our method on a question answering (QA) task using reconstructed test images from the NSD dataset. For each image, we retrieve a corresponding question–answer pair from the COCO dataset (Lin et al., 2014) and use Qwen2.5 (Bai et al., 2025) to answer the question based on the generated image. QA accuracy is reported in Table 3. Our method achieves an accuracy of 60.54%, significantly outperforming state-of-the-art methods. This demonstrates that our reconstructions are not only visually faithful but also semantically meaningful.

4.3 Ablation Study

In this subsection, we present ablation studies to justify our choice of text as the latent space and to evaluate the effectiveness of the object-centric diffusion and attribute–relationship search modules. Experiments are conducted on the NSD dataset, though the trend generalizes to other datasets.

Table 3: Image QA results on reconstructed NSD images. **Mdvs**, **Mde1** and **Mde2** refer to **Mindvis**, **Mindeye1** and **Mindeye2**, respectively. Results are reported as accuracy, with the best highlighted in red.

	PRISM	Takagi	Mdvs	Mde1	Mde2
Acc.↑	0.6054	0.4011	0.5037	0.5516	0.5765

We first compare reconstruction performance across three latent spaces: (1) language model embeddings (ours), (2) CLIP text embeddings (CLIP-Text), and (3) the image latent space of a diffusion model (LDM). As shown in Table 4, aligning fMRI signals to the language model text space consistently outperforms the other two spaces across all metrics. This demonstrates that textual representations alone can capture multiple levels of visual information, making text space a more brain-aligned and effective intermediate representation for fMRI-to-image reconstruction. Results for the remaining two datasets are provided in the Appendix (Table 7).

Table 4: Reconstruction performance across three latent spaces. The best result in each column is highlighted in red. ↑ indicates higher is better and ↓ indicates lower is better.

NSD	PixCorr ↑	SSIM ↑	LPIPS ↓	CLIP ↑	Inception V3 ↑
PRISM	$0.3404_{\pm 0.05}$	$0.4640_{\pm 0.02}$	$0.5963_{\pm 0.02}$	$0.9467_{\pm 0.03}$	$0.9516_{\pm 0.03}$
Clip text	$0.3208_{\pm 0.04}$	$0.3725_{\pm 0.06}$	$0.6611_{\pm 0.05}$	$0.9197_{\pm 0.02}$	$0.9011_{\pm 0.04}$
LDM	$0.2090_{\pm 0.07}$	$0.3727_{\pm 0.07}$	$0.7502_{\pm 0.04}$	$0.8602_{\pm 0.06}$	$0.8925_{\pm 0.05}$

Next, we evaluate the effectiveness of the two proposed modules. Results are in Table 5. To evaluate the Object-centric Diffusion module, we compare against a variant (**w/o ObjC.**) that replaces object-level cross-attention with standard U-Net cross-attention. To assess the attribute–relationship search module, we test two variants that skip the search process and rely only on the initial keyword set: **w/o AttOpt.+Bst**, which fixes the prompt to the highest-scoring (best) keyword, and **w/o AttOpt.+Wst**, which fixes it to the lowest-scoring (worst) keyword.

Overall, removing or replacing the two modules consistently degrades performance across all metrics. Specifically, eliminating object cross-attention leads to notable declines that cannot be recovered through prompt optimization, highlighting its essential role in reconstructing perceptually accurate images. Likewise, bypassing prompt optimization and using the best or worst initial attribute also reduces performance, indicating that the initial attributes alone are insufficient and underscoring the importance of prompt optimization in our model. The ablation study on the number of objects in our framework is shown in Section F.2.

Table 5: Effectiveness of the object-centric diffusion module and attribute–relationship search module on NSD data. The best result is highlighted in red.

	PixCorrorr ↑	SSIM ↑	LPIPS ↓
PRISM	$0.3404_{\pm 0.05}$	$0.4640_{\pm 0.02}$	$0.5963_{\pm 0.05}$
w/o ObjC.	$0.3291_{\pm 0.06}$	$0.4299_{\pm 0.06}$	$0.6111_{\pm 0.05}$
w/o AttOpt.+Bst	$0.3311_{\pm 0.04}$	$0.4421_{\pm 0.01}$	$0.6005_{\pm 0.02}$
w/o AttOpt.+Wst	$0.3068_{\pm 0.05}$	$0.4167_{\pm 0.02}$	$0.6398_{\pm 0.05}$

4.4 Case Study in Keyword Search

To better understand the keywords selected by our attribute–relationship search module, Table 6 presents the top-scoring keywords across different rounds of the ε-greedy search. The results show that, despite extensive exploration, the top-scoring keywords consistently converges toward spatially oriented relationships such as `Spatial Layout` and `Relative Position`. This suggests that: (1) descriptions emphasizing spatial information are most effective for guiding the diffusion model to accurately reconstruct images, as indicated by their highest LPIPS scores; and (2) these attributes also align well with fMRI data, as their CKA scores are no lower than those of the initial keywords, in accordance with the search constraints. This result is consistent with prior neuro-scientific findings showing that neural representations in the brain are sensitive to spatial arrangements and relative positions of objects (Zopf et al., 2018; Graumann et al., 2022). Therefore, we use `Spatial Layout` as the optimal keyword a to generate structured descriptions for model training. Further details are provided in section G.

Table 6: Top-5 relationship keywords scored by $1 - \text{LPIPS}$ before searching and after 10, 20, 30 search steps. The top-5 results remain unchanged after 30 search rounds. The search results indicate a clear preference for keywords related to spatial and positional relations, with most of the top-performing keywords in the final results containing the term '*spatial*'.

Rank	Initial	Round 10	Round 20	Round 30+
#1	Spatial Configuration	Spatial Arrangement	Spatial Organization	Spatial Layout
#2	Positional Relation	Spatial Configuration	Spatial Structure	Spatial Patterns
#3	Location Relation	Spatial Interaction	Spatial Arrangement	Spatial Organization
#4	Descriptive Attribute	Positional Relation	Spatial Configuration	Relative Position
#5	Inclusion Dependency	Feature Relation	Spatial Interaction	Spatial Relationships

5 Related Works

fMRI-Image Reconstruction. Early approaches leveraged linear models to decode fMRI signals into visual features (Kay et al., 2008; Takagi & Nishimoto, 2023). More recent work employs deep learning to map fMRI signals to the latent space of GANs (Lin et al., 2022; Ozcelik et al., 2022; Goodfellow et al., 2020) for image reconstruction. With advances in vision-language models (Radford et al., 2021; Liang et al., 2024), several studies have mapped fMRI signals to CLIP's image embedding space (Scotti et al., 2024; 2023) and used diffusion models for reconstruction (Rombach et al., 2022; Xu et al., 2023; Podell et al., 2023). Unlike prior work that directly maps fMRI signals to joint text–image spaces (Wang et al., 2024b; Quan et al., 2024), we compare multiple representation spaces and find that text embeddings from language models (Raffel et al., 2020) exhibit the strongest alignment with fMRI signals. This insight motivates our approach of reconstructing images via the embedding space of language models.

Diffusion models. Diffusion models have become foundational in generative tasks like image creation and editing (Wijmans & Baker, 1995; Gal et al., 2022; Song et al., 2020), as well as text-to-image synthesis (Ruiz et al., 2023). To enhance control over generated content, ControlNet (Zhang et al., 2023) introduces high-level image features for controlling and GLIGEN (Li et al., 2023; Zhang et al., 2025) incorporates position-aware adapters for spatial grounded generation. Meanwhile, there are also training-free methods that adjust latent or attention maps during inference to guide outputs without additional training (Chen et al., 2024; Yang et al., 2024). In our work, we guide the diffusion process by modifying cross-attention layers during inference to integrate object-level descriptions derived from fMRI data for image reconstruction.

Prompt Optimization. Prompt optimization aims to discover effective textual prompts for LLMs without model fine-tuning. Gradient-based methods (Shin et al., 2020; Shi et al., 2022; Wen et al., 2023) update prompts using gradients or differentiable embeddings. Gradient-free approaches treat LLMs as black boxes, using heuristic search (Prasad et al., 2022; Pryzant et al., 2023), reinforcement learning (Deng et al., 2022; Zhang et al., 2022), or evolutionary strategies (Zhou et al., 2022; Yang et al., 2023; Guo et al., 2025). We designed our gradient-free prompt optimization based on beam search to optimize attribute keywords for black-box vision-language models.

6 Conclusion

In this work, we addressed the challenge of reconstructing visual stimuli from fMRI signals. Our analysis revealed that fMRI signals align more closely with the text space of language models than with vision-based or joint text–image representations, identifying text as a brain-aligned intermediate space. Building on this insight, we showed that explicitly modeling the compositional structure of visual perception—capturing objects along with their attributes and relationships—further improves reconstruction quality. Guided by these findings, we developed **PRISM**, a framework that maps fMRI signals into a structured text space and incorporates two specialized modules: an object-centric diffusion module that generates images by composing individual objects, and an attribute–relationship search module that automatically discovers attributes and relationships aligned with neural activity. Experiments on real-world fMRI datasets demonstrate that PRISM reduces perceptual loss by up to 8% compared to prior methods, underscoring the power of structured text as a bridge between brain activity and image generation.

7 ETHICS STATEMENT

Our work does not involve human or animal subjects, personally identifiable data, or sensitive information. The datasets used are publicly available, and we follow their respective licenses. The methods and findings presented do not pose foreseeable risks of misuse, discrimination, or harm. We therefore believe our work raises no specific ethical concerns under the ICLR Code of Ethics.

8 REPRODUCIBILITY STATEMENT

Section 3 details the proposed framework and its design. Section 4 describes the datasets, baseline methods, and evaluation protocols used for comparison. Additional implementation details, including training procedures and hyperparameter settings, are provided in the Appendix. Upon acceptance of this paper, we will release our code on GitHub.

REFERENCES

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.

George A Alvarez and Steven L Franconeri. How many objects can you track?: Evidence for a resource-limited attentive tracking mechanism. *Journal of vision*, 7(13):14–14, 2007.

Sergio A Alvarez. Gaussian rbf centered kernel alignment (cka) in the large-bandwidth limit. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6587–6593, 2022.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Stefania Bracci and Hans P Op de Beeck. Understanding human object vision: a picture is worth a thousand representations. *Annual review of psychology*, 74(1):113–135, 2023.

Nadine Chang, John A Pyles, Austin Marcus, Abhinav Gupta, Michael J Tarr, and Elissa M Aminoff. Bold5000, a public fmri dataset while viewing 5000 visual images. *Scientific data*, 6(1):49, 2019.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024.

Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 5343–5353, 2024.

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*, pp. 22710–22720, 2023.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13(1):795–828, 2012.

Nelson Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, 24(1):87–114, 2001.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P Xing, and Zhiting Hu. Rlprompt: Optimizing discrete text prompts with reinforcement learning. *arXiv preprint arXiv:2205.12548*, 2022.

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.

Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 1778–1785. IEEE, 2009.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Monika Graumann, Caterina Ciuffi, Kshitij Dwivedi, Gemma Roig, and Radoslaw M Cichy. The spatiotemporal neural dynamics of object location representations in the human brain. *Nature human behaviour*, 6(6):796–811, 2022.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*, 2020.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. Evoprompt: Connecting llms with evolutionary algorithms yields powerful prompt optimizers, 2025. URL https://arxiv.org/abs/2309.08532.

Mandar Haldekar, Ashwinkumar Ganesan, and Tim Oates. Identifying spatial relations in images using convolutional neural networks. In *2017 international joint conference on neural networks (ijcnn)*, pp. 3593–3600. IEEE, 2017.

Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Bießmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, 87:96–110, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 473–483, 2017.

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

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3668–3678, 2015.

Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

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

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22511–22521, 2023.

Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pp. 405–409, 2024.

Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind reader: Reconstructing complex images from brain activities. *Advances in Neural Information Processing Systems*, 35:29624–29636, 2022.

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 *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 852–869. Springer, 2016.

Andrew F Luo, Margaret M Henderson, Michael J Tarr, and Leila Wehbe. Brainscuba: Fine-grained natural language captions of visual cortex selectivity. *arXiv preprint arXiv:2310.04420*, 2023.

Weijian Mai, Jian Zhang, Pengfei Fang, and Zhijun Zhang. Brain-conditional multimodal synthesis: A survey and taxonomy. *IEEE Transactions on Artificial Intelligence*, 2024.

Lluís Màrquez, Xavier Carreras, Kenneth C Litkowski, and Suzanne Stevenson. Semantic role labeling: an introduction to the special issue, 2008.

David Marr. Visual information processing: The structure and creation of visual representations. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 290(1038): 199–218, 1980.

Alex Murphy, Joel Zylberberg, and Alona Fyshe. Correcting biased centered kernel alignment measures in biological and artificial neural networks. *arXiv preprint arXiv:2405.01012*, 2024.

Thomas Naselaris, Ryan J Prenger, Kendrick N Kay, Michael Oliver, and Jack L Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915, 2009.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

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)*, pp. 1–8. IEEE, 2022.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106, 2005.

Pablo Pernias, Dominic Rampas, Mats L Richter, Christopher J Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. *arXiv preprint arXiv:2306.00637*, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. Grips: Gradient-free, edit-based instruction search for prompting large language models. *arXiv preprint arXiv:2203.07281*, 2022.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. Automatic prompt optimization with" gradient descent" and beam search. *arXiv preprint arXiv:2305.03495*, 2023.

Weikang Qiu, Zheng Huang, Haoyu Hu, Aosong Feng, Yujun Yan, and Rex Ying. Mindllm: A subject-agnostic and versatile model for fmri-to-text decoding. In *Forty-second International Conference on Machine Learning*, 2025.

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*, pp. 233–243, 2024.

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 *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

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*, pp. 10684–10695, 2022.

Alexis Ross, Tongshuang Wu, Hao Peng, Matthew E Peters, and Matt Gardner. Tailor: Generating and perturbing text with semantic controls. *arXiv preprint arXiv:2107.07150*, 2021.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023.

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. *Advances in Neural Information Processing Systems*, 36:24705–24728, 2023.

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. *arXiv preprint arXiv:2403.11207*, 2024.

Weijia Shi, Xiaochuang Han, Hila Gonen, Ari Holtzman, Yulia Tsvetkov, and Luke Zettlemoyer. Toward human readable prompt tuning: Kubrick's the shining is a good movie, and a good prompt too? *arXiv preprint arXiv:2212.10539*, 2022.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020.

Ranganatha Sitaram, Nikolaus Weiskopf, Andrea Caria, Ralf Veit, Michael Erb, and Niels Birbaumer. fmri brain-computer interfaces. *IEEE Signal processing magazine*, 25(1):95–106, 2008.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Jingyuan Sun, Mingxiao Li, Zijiao Chen, Yunhao Zhang, Shaonan Wang, and Marie-Francine Moens. Contrast, attend and diffuse to decode high-resolution images from brain activities. *Advances in Neural Information Processing Systems*, 36:12332–12348, 2023.

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

J Jay Todd and René Marois. Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature*, 428(6984):751–754, 2004.

Hao-Ting Wang, Jonathan Smallwood, Janaina Mourao-Miranda, Cedric Huchuan Xia, Theodore D Satterthwaite, Danielle S Bassett, and Danilo Bzdok. Finding the needle in a high-dimensional haystack: Canonical correlation analysis for neuroscientists. *NeuroImage*, 216:116745, 2020.

Ling Wang, Chen Wu, and Lin Wang. Braindreamer: Reasoning-coherent and controllable image generation from eeg brain signals via language guidance. *arXiv preprint arXiv:2409.14021*, 2024a.

Yanchen Wang, Adam Turnbull, Tiange Xiang, Yunlong Xu, Sa Zhou, Adnan Masoud, Shekoofeh Azizi, Feng Vankee Lin, and Ehsan Adeli. Decoding visual experience and mapping semantics through whole-brain analysis using fmri foundation models. *arXiv preprint arXiv:2411.07121*, 2024b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems*, 36:51008–51025, 2023.

Johannes G Wijmans and Richard W Baker. The solution-diffusion model: a review. *Journal of membrane science*, 107(1-2):1–21, 1995.

Christopher Williams, Fabian Falck, George Deligiannidis, Chris C Holmes, Arnaud Doucet, and Saifuddin Syed. A unified framework for u-net design and analysis. *Advances in Neural Information Processing Systems*, 36:27745–27782, 2023.

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*, pp. 8226–8235, 2024.

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*, pp. 7754–7765, 2023.

Yaoda Xu and Marvin M Chun. Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature*, 440(7080):91–95, 2006.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*, 2023.

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *Forty-first International Conference on Machine Learning*, 2024.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Tianjun Zhang, Xuezhi Wang, Denny Zhou, Dale Schuurmans, and Joseph E Gonzalez. Tempera: Test-time prompting via reinforcement learning. *arXiv preprint arXiv:2211.11890*, 2022.

Yuyao Zhang, Jinghao Li, and Yu-Wing Tai. Layercraft: Enhancing text-to-image generation with cot reasoning and layered object integration, 2025. URL https://arxiv.org/abs/2504.00010.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*, 2022.

Yixin Zhu, Yibiao Zhao, and Song Chun Zhu. Understanding tools: Task-oriented object modeling, learning and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2855–2864, 2015.

Regine Zopf, Marina Butko, Alexandra Woolgar, Mark A Williams, and Anina N Rich. Representing the location of manipulable objects in shape-selective occipitotemporal cortex: Beyond retinotopic reference frames? *Cortex*, 106:132–150, 2018.

A A COMMON ERROR IN GENERATIVE MODELS

Original Image **Generated Image**

Figure 3: A Common Error in Generative Models. While the original image shows "a gray tiger-striped cat," the model incorrectly generates "a grey tiger," illustrating semantic distortion.

In this section, we present a common failure, attribute binding, encountered in generative models (especially for the diffusion model), where generative models misattribute visual properties to objects. Figure 3 compares original (left) and generated (right) images. The original depicts a gray tiger-striped cat on a wooden bench, while the generated version incorrectly shows a gray tiger instead of a cat. This issue arises because diffusion models usually rely on text encoders such as CLIP, which are known to lack the ability to capture complex linguistic structures (Yuksekgonul et al., 2022). Consequently, the diffusion process loses awareness of the bindings between objects and their attributes, leading to mismatched visual properties. This impairs fMRI-to-image reconstruction. To address this, we introduce a neuroscience-inspired, object-centric generation approach that improves reconstruction quality.

B PRELIMINARY: DIFFUSION MODEL

Diffusion models are a class of generative models that synthesize data by reversing a gradual noising process. Given a data point $\mathbf{H}_0$ (e.g., an image), the forward process perturbs it into Gaussian noise over T time steps. The model then learns the reverse process to reconstruct samples from noise. The forward process is a Markov chain defined by:

$$q(\mathbf{H}_t \mid \mathbf{H}_{t-1}) = \mathcal{N}(\mathbf{H}_t; \sqrt{1-\beta_t}\,\mathbf{H}_{t-1}, \beta_t \mathbf{I}), \quad t = 1, \ldots, T,$$

where $\{\beta_t\}_{t=1}^T$ is a predefined variance schedule. The model is trained to predict the noise $\boldsymbol{\epsilon}$ added to the input, using a neural network $\boldsymbol{\epsilon}_\theta$, by minimizing:

$$\mathcal{L} = \mathbb{E}_{\mathbf{H}_0, \boldsymbol{\epsilon}, t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{H}_t, t)\|^2\right].$$

Here, $\mathbf{H}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{H}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}$, with $\bar{\alpha}_t = \prod_{s=1}^t (1-\beta_s)$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$. To guide the generation process with external information $\mathbf{C}$ (e.g., a text prompt), the denoising network is extended as:

$$\boldsymbol{\epsilon}_\theta(\mathbf{H}_t, t, \mathbf{C}).$$

Then, the training objective becomes:

$$\mathcal{L}_{\text{cond}} = \mathbb{E}_{\mathbf{H}_0, \boldsymbol{\epsilon}, t, \mathbf{C}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{H}_t, t, \mathbf{C})\|^2\right].$$

This formulation is widely used in text-to-image diffusion models, where $\mathbf{C}$ is the embedding of a textual description obtained from a pre-trained text encoder (e.g., CLIP). In practice, the condition $\mathbf{C}$ is incorporated into the U-Net via cross-attention modules.

In our work, we adopt the latent diffusion framework (Rombach et al., 2022), where the diffusion process is applied in the latent space of a pre-trained VAE, rather than directly in pixel space. Specifically, an input image $\mathbf{Y} \in \mathbb{R}^{H\times W\times 3}$ is first encoded by a VAE encoder into a compact latent representation $\mathbf{Z} \in \mathbb{R}^{h\times w\times c}$:

$$\mathbf{Z} = \text{Encoder}(\mathbf{Y}).$$

The diffusion process is applied on $\mathbf{Z}$, where the perturbed latent representation $\mathbf{Z}_T$ is obtained after T steps. The reversed denoising steps then generate a denoised latent $\hat{\mathbf{Z}}$ over T steps. The final image is reconstructed by the denoised latent:

$$\hat{\mathbf{Y}} = \text{Decoder}(\hat{\mathbf{Z}}).$$

This formulation greatly reduces computational cost while maintaining high-quality image generation and is particularly well-suited for conditioning on high-level semantic representations such as text or fMRI-derived embeddings.

C Dataset

In this subsection, we provide information about the three pre-processed datasets used for the fMRI-to-image reconstruction task: NSD (Allen et al., 2022), BOLD5000 (Chang et al., 2019), and GOD (Horikawa & Kamitani, 2017).

- **NSD** (Allen et al., 2022): The Natural Scenes Dataset (NSD) is a large-scale public fMRI dataset capturing brain responses of human participants viewing naturalistic stimuli from COCO images (Lin et al., 2014). The dataset includes scans for 30–40 hours across 30–40 separate sessions. During each session, participants viewed 750 images for 3 seconds each. Each image was presented three times across sessions, with most images unique to each subject, except for 1,000 shared images seen by all subjects. Following prior NSD reconstruction studies (Scotti et al., 2023; Takagi & Nishimoto, 2023), we adopt the standardized train/test split, where the shared images serve the test set. Consequently, the training set for each subject contains 8,859 image stimuli and 24,980 fMRI trials, while the test set includes 982 image stimuli and 2,770 fMRI trials.
- **BOLD5000** (Chang et al., 2019): The BOLD5000 dataset is a publicly available fMRI dataset capturing brain activity as subjects viewed a series of images. It contains 4,916 unique images, including 2,000 from the COCO dataset and 1,916 from ImageNet (Deng et al., 2009). Each image was presented as a visual stimulus in individual trials. Of these, 4,803 images were shown once, while 113 images were repeated three or four times across trials, resulting in a total of 5,254 stimulus trials. We follow the standardized train/test split used in prior BOLD5000 reconstruction studies (Chen et al., 2023; Wang et al., 2024b). Specifically, the training set includes trials with non-repeated image stimuli, comprising 4,803 samples, while the test set consolidate repeated image stimulus trials into 113 samples.
- **GOD** (Horikawa & Kamitani, 2017): The Generic Object Decoding (GOD) is a public dataset developed for fMRI based decoding. It aggregates fMRI data gathered through the presentation of images from 200 representative object categories, originating from ImageNet. We follow the standardized train/test set split employed in existing GOD image reconstruction studies (Sun et al., 2023) and get 1200 training samples and 50 test samples.The Generic Object Decoding (GOD) dataset is a publicly available fMRI dataset designed for decoding object representations. It includes fMRI data collected during the presentation of images from 200 representative object categories sourced from ImageNet. Following the standardized train/test split used in prior GOD reconstruction studies (Sun et al., 2023), we use 1,200 training samples and 50 test samples.

D BASELINES

In this section, we provide details about the baselines used in our experiments.

- Takagi & Nishimoto (Takagi & Nishimoto, 2023): This baseline maps fMRI signals to the latent space of a pre-trained VAE within a diffusion model using linear regression, enabling image reconstruction. The method combines image latent representations with text embeddings extracted from a CLIP text encoder, both mapped from fMRI signals in higher (ventral) visual cortex regions, to improve reconstruction quality. For a fair comparison, we adapt this approach to the Diffusion 2.1 pipeline by retraining the linear regression to map fMRI signals to both the VAE latent space and the textual conditioning used in Diffusion 2.1.
- Mindvis (Chen et al., 2023): This baseline uses a self-supervised representation of fMRI data using masked modeling within a high-dimensional latent space in an encoder–decoder framework. The learned representation is then projected into the conditioning space of LDM by fine-tuning the model. For fair comparison, we adapt the image generator of this approach to Diffusion 2.1 by fine-tuning it with the learned projection module, following the strategy outlined by the original authors.
- Mindeye (Scotti et al., 2023): This model proposed two modules to map the fMRI signal to the CLIP image space. Specifically, the model first uses contrastive learning to align fMRI signals with image embeddings. Second, the paper trains a diffusion prior to reconstructing images from these embeddings via mapping brain activity into CLIP image space, enabling the generation of images that closely resemble the original stimuli. To adapt the method for fair comparison, we replace the Versatile Diffusion with Diffusion 2.1.
- Mindeye2 (Scotti et al., 2024): This method trains multiple MLPs to project fMRI signals from all subjects into a shared representation space, followed by training a diffusion prior to map these representations into the CLIP image embedding space. The final image is then reconstructed using a pre-trained SDXL (Podell et al., 2023). To adapt this method to our setting, we replace the generative backbone with Diffusion 2.1.

E TRAINING DETAILS

In this section, we provide the training details of our model. Our model is implemented with Pytorch and trained on two NVIDIA-L40 GPUs with 48GB of memory. We use T5 as the language model to generate the object-level descriptions. For NSD data, we train the model for 80 epochs: 60 epochs for MLP training (E_{MLP}) with a learning rate $\text{lr}_1 = 1 \times 10^{-5}$, followed by 20 epochs of joint training, where we continue training the MLP and fine-tune the T5 model (E_{T5}) using a learning rate $\text{lr}_2 = 5 \times 10^{-7}$. For BOLD5000, we set $E_{\text{mlp}} = 50$, $\text{lr}_1 = 1e-5$, $E_{\text{T5}} = 5$, and $\text{lr}_2 = 1e^{10-8}$. For GOD, we set $E_{\text{mlp}} = 40$, $\text{lr}_1 = 1e-5$, $E_{T5} = 5$, and $\text{lr}_2 = 5e^{-9}$. For image reconstruction at inference time, we set the blending ratio $\beta = 0.5$ and the denoising step as 40 for all the datasets. We use `GPT 4o-mini` to generate the object-centric descriptions with the prompt shown in H. For images that do not have a caption, we first use `GPT-4o-mini` to generate a short caption and then use our prompt to generate the object-centric description.

F ADDITIONAL EXPERIMENTS

F.1 RECONSTRUCTION PERFORMANCE ACROSS THREE LATENT SPACES.

We report the ablation studies to justify our choice of text as the latent space on BOLD5000 and GOD. As shown in Table 7, aligning fMRI signals to the language model text space (our method) consistently outperforms alignment to the other two spaces across all metrics, including CLIP and Inception V3. This supports our core contribution: textual representations alone are sufficient to capture both high-level semantic and low-level visual information, and text space provides a more brain-aligned and effective intermediate representation for fMRI-to-image reconstruction.

Table 7: Reconstruction performance across three latent spaces. The best result in each column is highlighted in red. ↑ indicates higher is better and ↓ indicates lower is better.

	PixCorr ↑	SSIM ↑	LPIPS ↓	CLIP ↑	Inception V3 ↑
BOLD5000					
Ours	$0.2315_{\pm 0.01}$	$0.5341_{\pm 0.02}$	$0.6198_{\pm 0.02}$	$0.7720_{\pm 0.03}$	$0.6601_{\pm 0.07}$
Clip text	$0.2000_{\pm 0.06}$	$0.4885_{\pm 0.06}$	$0.6520_{\pm 0.04}$	$0.7265_{\pm 0.04}$	$0.6199_{\pm 0.04}$
LDM	$0.1622_{\pm 0.03}$	$0.4300_{\pm 0.02}$	$0.7894_{\pm 0.08}$	$0.7025_{\pm 0.08}$	$0.5590_{\pm 0.05}$
GOD					
Ours	$0.2571_{\pm 0.01}$	$0.5200_{\pm 0.02}$	$0.6213_{\pm 0.01}$	$0.8567_{\pm 0.05}$	$0.8428_{\pm 0.06}$
Clip text	$0.2200_{\pm 0.05}$	$0.4682_{\pm 0.04}$	$0.6827_{\pm 0.04}$	$0.8120_{\pm 0.05}$	$0.7602_{\pm 0.04}$
LDM	$0.1900_{\pm 0.02}$	$0.3999_{\pm 0.07}$	$0.7100_{\pm 0.02}$	$0.7099_{\pm 0.01}$	$0.7484_{\pm 0.03}$

Table 8: Comparison of the number of objects in our framework. Results are reported using PixCorr, SSIM, LPIPS, CLIP, and Inception V3 metrics. The best result in each column is highlighted in red. ↑ indicates higher is better and ↓ indicates lower is better.

	PixCorr ↑	SSIM ↑	LPIPS ↓	CLIP ↑	Inception V3 ↑
Ours (Two Objs)	$0.3404_{\pm 0.05}$	$0.464_{\pm 0.02}$	$0.5943_{\pm 0.02}$	$0.9467_{\pm 0.03}$	$0.9516_{\pm 0.03}$
One Obj	$0.3355_{\pm 0.05}$	$0.4532_{\pm 0.04}$	$0.6014_{\pm 0.04}$	$0.9344_{\pm 0.02}$	$0.9342_{\pm 0.01}$
Four Objs	$0.3202_{\pm 0.03}$	$0.4469_{\pm 0.04}$	$0.6284_{\pm 0.02}$	$0.9400_{\pm 0.05}$	$0.9322_{\pm 0.05}$

F.2 Analysis on the number of objects in our framework

In our framework, we fix the number of objects per image to two and assign a separate MLP to each. We learn to assign each object a location label from a predefined set of spatial positions (e.g., left/right or top/bottom). This fixed assignment of each object to a dedicated MLP, along with the predefined spatial labeling scheme, is used consistently during both training and inference. At inference time, each MLP independently encodes fMRI signals for one object, and the language model generates a structured description for each. These descriptions are then passed to the object-centric diffusion model, which generates object images independently and places them into their corresponding spatial positions to form the final image.

We conduct an experiment to determine the optimal number of objects (MLPs) in our framework; the results on the NSD dataset are shown in Table 8. The results reveal that setting the number of objects per image to two yields the best performance. This choice is also supported by neuroscientific findings suggesting that, although the number of objects in an image is inherently uncertain, human attention and memory are limited to only a few of the objects. Both empirical experiments Alvarez & Franconeri (2007) and neural evidence Cowan (2001); Todd & Marois (2004) show that humans can attend to only 3–4 simple objects (e.g., a circle on a white background) at a time. For complex objects (e.g., those with intricate color patterns), this capacity drops to around 2 due to the increased cognitive load per item Xu & Chun (2006). This cognitive bottleneck limits the amount of information that can be decoded from fMRI signals. As a result, increasing the number of m does not necessarily enhance the level of detail in the reconstructed image and may even lead to less reliable reconstructions—for example, by hallucinating non-existent objects. Therefore, we choose to use $m = 2$, as it empirically yields the best performance.

G Relation Optimization Details

The detailed algorithm used to solve the prompt optimization problem in equation 3 is provided in Algorithm 1. In our experiments, we adopt the scoring function $\mathcal{S}(\mathbf{Y}_1, \mathbf{Y}_2) = 1 - \text{LPIPS}(\mathbf{Y}_1, \mathbf{Y}_2)$. The CKA threshold β is initialized to the minimum CKA score among the initial candidates. We set $\varepsilon = 0.5, k_1 = 8, k_2 = 2$ and search for $T = 40$ rounds. We randomly sampled 667 images from the training set of **NSD** (Allen et al., 2022) for prompt optimization. While this subset was used for efficiency, our method is applicable to the full training set and generalizes to other settings.

we use `GPT-4o-mini` as the VLM and the LLM that generates new keywords. We use CLIP-text (Radford et al., 2021) as LM_{ENC}, and Stable Diffusion 2.1 (Pernias et al., 2023; Rombach et al., 2022) as the diffusion model.

Algorithm 1 ε-Greedy Prompt Optimization

Input: Training set $\mathcal{X}^{\text{train}}, \mathcal{Y}^{\text{train}}$; Initial keyword set $\mathcal{A}$; Search rounds T; Parameters ε, k_1, k_2
Initialize: Threshold $\beta \leftarrow \min_{a \in \mathcal{A}} \text{CKA}\left(\mathbf{X}, \mathbf{K}^a\right)$
for $t = 1$ to T **do**
  Filter: $\mathcal{A} \leftarrow \{a \in \mathcal{A} \mid \text{CKA}\left(\mathbf{X}, \mathbf{K}^a\right) > \beta\}$
  Sort: Rank $a \in \mathcal{A}$ in descending order by $\sum_{i=1}^{N} \mathcal{S}\left(\mathbf{Y}_i, \text{Diff}(\text{VLM}(\mathbf{Y}_i, \mathcal{P}(a)))\right)$
  if `random()` $< \varepsilon$ **then**
    Sample: $\mathcal{S} \leftarrow \text{RandomSample}(\mathcal{A}, k_1)$ % Randomly sample k_1 keywords from $\mathcal{A}$
  else
    Select: $\mathcal{S} \leftarrow \text{Top}(\mathcal{A}, k_1)$ % Select top-k_1 keywords from $\mathcal{A}$
  end if
  Generate: Use LLM to synthesize k_2 new keywords $\mathcal{A}_{\text{new}}$ based on $\mathcal{S}$
  Update: $\mathcal{A} \leftarrow \mathcal{A} \cup \mathcal{A}_{\text{new}}$
end for
Output: $\arg\max_{a \in \mathcal{A}} \sum_{i=1}^{N} \mathcal{S}\left(\mathbf{Y}_i, \text{Diff}(\text{VLM}(\mathbf{Y}_i, \mathcal{P}(a)))\right)$

The search is initialized with six widely-used keywords describing object relationships: Semantic Relation (Johnson et al., 2015), Positional Relation (Lu et al., 2016; Haldekar et al., 2017), Functional Relation (Zhu et al., 2015), Action Relation (Lu et al., 2016), Visual Attributive Relation (Farhadi et al., 2009), and Part–Whole Relation(Lu et al., 2016). For each type, we use `GPT-4o` to generate four synonymous keywords, resulting in an initial pool of 24 candidate attributes. Figure 4 reports the LPIPS scores of all initial and subsequently discovered relation keywords.

H THE USE OF LARGE LANGUAGE MODELS (LLMS)

We declare that Large Language Models (LLMs) were confined to peripheral tasks and had no influence on the methodology, results interpretation, or theoretical insights of this work. Specifically, they were used for (i) generating training datasets required for our experiments and (ii) grammar correction and minor word-level refinements. All language edits were carefully reviewed by the authors to ensure that no hallucinations were introduced and that the text faithfully reflects the original intent. The technical development, experimental design, analysis, and conclusions presented here are entirely the work of the authors.

The prompt used by the `GPT-4o-mini` is shown below:

Prompt in generating structured descriptions

Given the image and caption, first describe the background color style of the image with 3-5 words. Second, detect the TWO most important objects in the image. Then, describe each of the objects and their relationship using: **{keyword}** with TWO sentences. For each sentence, use 5-10 words and as easy as possible.

Then, detect the absolute position of the two objects in the image, and select from [right, left, top, bottom]. "left" and "right" should appear together for horizontal objects, and "top" and "bottom" should appear together for vertical objects. DO NOT mix.

Example:

Background color style: Grayscale urban.

The Man [left]
1. The man is standing near the sidewalk edge. The Man is close to the building wall.

The Suitcase [right]
1. The suitcase is beside the man's foot. The Suitcase is placed on the street's curved edge.

Now, given the image I uploaded and the caption "**{caption}**", detect the two most important objects with absolute position, describe them using **{keyword}** with EXACTLY the example format:

Prompt in attribute–relationship optimization

System prompt: You are a helpful brainstormer. Given a list of keywords, generate **{gen_num}** related or similar keywords. Respond with a comma-separated list of keywords.

User prompt: keyword: **{keywords}**

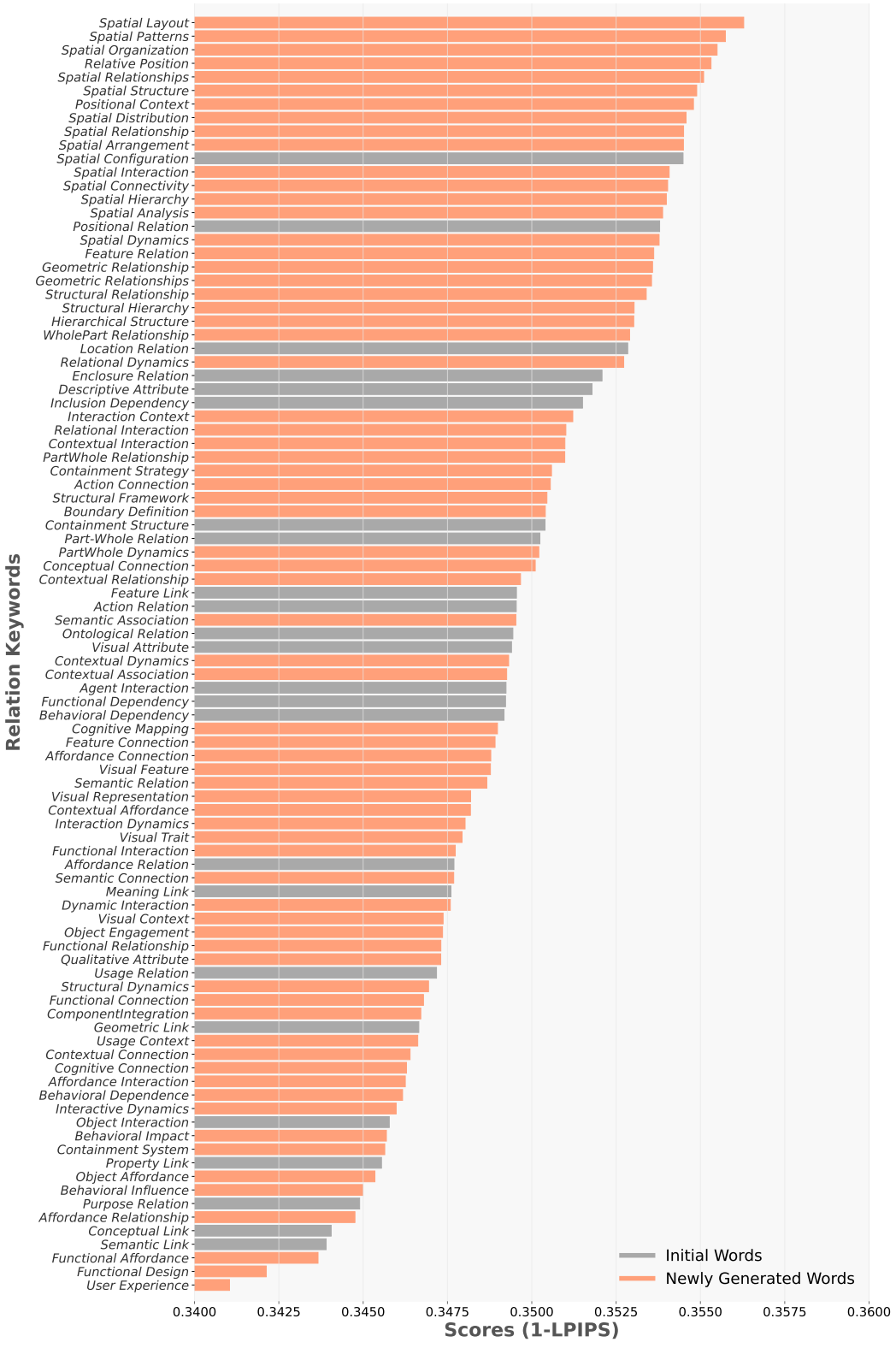

Figure 4: Scores of all tested relation keywords during the prompt search. Among the top-10 relation keywords, the most frequent keyword is 'spatial'.