Uncovering Semantic Selectivity of Latent Groups in Higher Visual Cortex with Mutual Information-Guided Diffusion

Yule Wang, Joseph Yu, Chengrui Li, Weihan Li, Anqi Wu
Georgia Institute of Technology, Atlanta, GA 30332, USA
{yulewang, jyu425, cnlichengrui, weihanli, anqiwu}@gatech.edu

Abstract

Understanding how neural populations in higher visual areas encode object-centered visual information remains a central challenge in computational neuroscience. Prior works have investigated representational alignment between artificial neural networks and the visual cortex. Nevertheless, these findings are indirect and offer limited insights to the structure of neural populations themselves. Similarly, decoding-based methods have quantified semantic features from neural populations but have not uncovered their underlying organizations. This leaves open a scientific question: "*how feature-specific visual information is distributed across neural populations in higher visual areas, and whether it is organized into structured, semantically meaningful subspaces.*" To tackle this problem, we present MIG-Vis, a method that leverages the generative power of diffusion models to visualize and validate the visual-semantic attributes encoded in neural latent subspaces. Our method first uses a variational autoencoder to infer a group-wise disentangled neural latent subspace from neural populations. Subsequently, we propose a mutual information (MI)–guided diffusion synthesis procedure to visualize the specific visual-semantic features encoded by each latent group. We validate MIG-Vis on multi-session neural spiking datasets from the inferior temporal (IT) cortex of two macaques. The synthesized results demonstrate that our method identifies neural latent groups with clear semantic selectivity to diverse visual features, including object pose, inter-category transformations, and intra-class content. These findings provide direct, interpretable evidence of structured semantic representation in the higher visual cortex and advance our understanding of its encoding principles.

1 Introduction

Determining how populations of neurons in higher visual areas represent object-centred visual information remains a central question in computational neuroscience (DiCarlo et al., 2012; Yamins & DiCarlo, 2016). A major line of research (Lindsey & Issa, 2024; Xie et al., 2024) has explored representational alignment between deep neural networks (DNNs) and primate visual cortex, showing that DNNs trained for object recognition—especially those with disentangled representation subspaces—closely resemble inferior temporal (IT) cortex activity. Nevertheless, these findings are indirect, relying on artificial model architecture and specially designed representational similarity metrics. Meanwhile, previous works have focused on single-unit selectivity (Rust & DiCarlo, 2010) or decoding-based methods (Freiwald & Tsao, 2010; Chang & Tsao, 2017) that quantify semantic features like object category or viewpoint in higher visual cortex. Earlier studies have also used pre-trained diffusion models with fMRI data (Luo et al., 2023a;b; Cerdas et al., 2024) to verify and classify that various brain regions specialize in processing certain category's information.

However, existing approaches do not extract semantically interpretable neural representations from electrophysiological recordings in higher visual cortex. That said, no study has mapped the organization and structural patterns of higher visual area neural populations to distinct visual attributes. In practice, addressing this scientific problem is challenging due to neural single-unit activities in higher visual cortex exhibit mixed selectivity to multiple visual-semantic features (Chang & Tsao, 2017). We also verify this finding through an empirical study on the IT cortex of macaques in a passive object recognition task (Majaj et al., 2015) (the results are presented in Fig. 1).
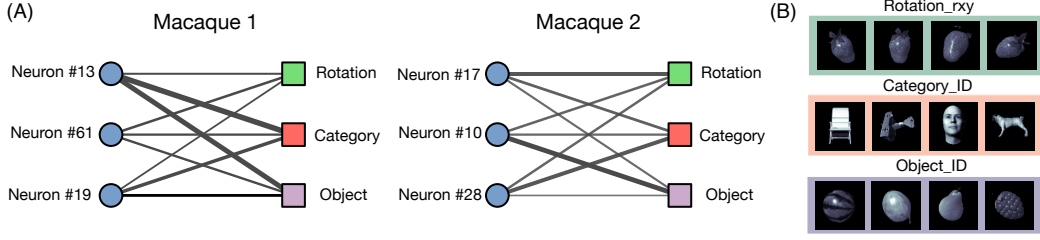
Figure 1: **Single-neuron linear decoding results in IT cortex of two macaques during a passive object recognition task.** **(A)** Decoding results on both two macaques reveal that IT neurons exhibit mixed selectivity. The connecting line thickness denotes the weight scale. These neurons contribute to both low-level pose attributes (e.g., rotation) and high-level semantic features (e.g., category id). **(B)** Images with variations along each visual-semantic feature.

To bridge this gap, we propose **MIG-Vis** (short for **M**utual **I**nformation-**G**uided Diffusion for Uncovering Semantic Selectivity of Neural Latent Groups in Higher **Vis**ual Cortex), a method identifying interpretable neural latent subspaces that exhibit visual-semantic selectivity. We propose to employ a disentangled latent variable model (LVM) to infer low-dimensional neural latents for semantic interpretation. We assume visual-semantic features span a multi-dimensional subspace encoding different aspects of object structure and variation. To capture this, we use a **group-wise disentangled** variational autoencoder (Li et al., 2025), which enforces each multi-dimensional latent group to represent distinct semantic features.

Given an identified latent group in the neural subspace, our subsequent goal is to characterize the specific visual-semantic features it encodes. Here we leverage the expressivity of generative diffusion models (Wang et al., 2023) to synthesize images guided by target neural signals (Luo et al., 2023a; Cerdas et al., 2024). We perform semantic image editing (Meng et al., 2021) with deterministic DDIM, which eliminates stochastic sampling noise to ensure that editing is completely guided by neural signals. This procedure enables us to visualize and interpret the specific visual-semantic features encoded by each neural latent group.

For diffusion guidance, prior works typically optimize simple statistical moments of target neural dimensions, such as absolute activation values (Luo et al., 2023a) or variance (Wang et al., 2024). Such approaches are especially common in the fMRI literature, where the neural space is directly manipulated—zero indicates no activation, and larger positive values correspond to stronger activation. However, such guidance strategies are ill-suited for uncovering semantic features of specific latent axes in our setting, since both positive and negative coordinates in the latent space can carry distinct meanings that first- or second-order moment objectives cannot fully capture.

To deal with this issue, we propose a novel guidance objective: **maximizing mutual information (MI)** between the synthesized image and the target latent group. MI provides an information-theoretic measure of inherent statistical dependence (Hjelm et al., 2018) between the synthesized image and the target neural latent group, capturing both linear and higher-order relationships. Therefore, maximizing MI guides the synthesized image to faithfully represent the full distribution of visual-semantic information in the target latent group, achieving alignment beyond simple first- or second-order moments.

We experimentally evaluate the efficacy of MIG-Vis on multi-session neural spiking datasets (Majaj et al., 2015) from the IT cortex of two macaques during a passive object recognition task. Extensive diffusion-synthesized results demonstrate that MIG-Vis infers disentangled neural latent groups with visual-semantic selectivity. By synthesizing image with the guidance of maximizing the mutual information between the image output and each latent group, we characterize clear semantic specialization across the disentangled groups: inter-category variation, intra-category pose, and intra-category content details. Furthermore, we adopt this scientific interpretation method to within-group latent dimensions to uncover fine-grained semantic structure.

2 PRELIMINARIES

Problem Formulation. For each single trial, data is composed of neural population and stimulus image pairs: $(\mathbf{x}, \mathbf{y})$. The neural population is denoted as $\mathbf{x} \in \mathbb{R}^N$, where N is the number of recorded neural units. The corresponding image is represented as $\mathbf{y} \in \mathbb{R}^{1 \times H \times W}$, where H and W denote the height and width of the gray-scale image. We develop a group-wise disentangled VAE
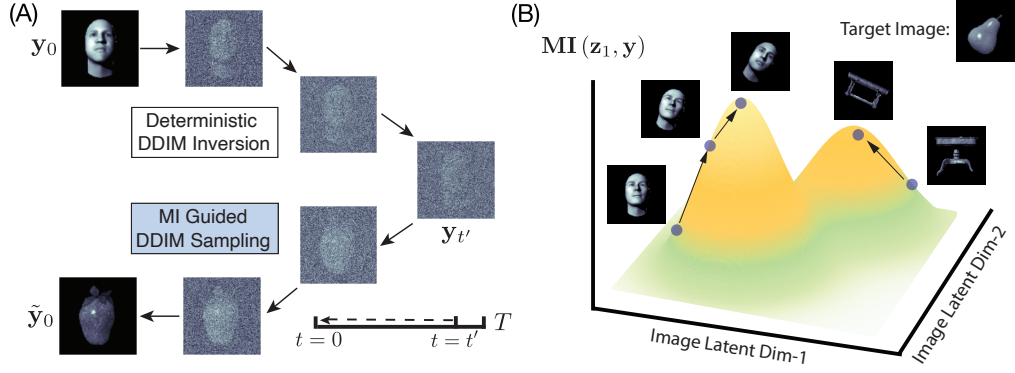
Figure 2: (A) **The semantic image editing procedure.** This consists of a deterministic forward process from $t = 0$ to an intermediate timestep $t = t'$, followed by a neural-guided deterministic synthesis process back to $t = 0$. (B) **A schematic of the Mutual Information (MI) Landscape.** This landscape is defined by our guidance objective, which maximizes the MI between a synthesized image and the neural latent group $\mathbf{z}_1$ of the target image (encoding object pose).

based on Li et al. (2025) to infer the neural latent groups, denoted as $\mathbf{z} = [\mathbf{z}_1, \ldots, \mathbf{z}_G]^\top \in \mathbb{R}^D$, in which D is the latent dimension number and G is the number of groups. Each latent group is set to have the same dimensionality D_g, thus $D = G \times D_g$. Our goal is to investigate the specific visual-semantic encoding of the g-th neural latent group $\mathbf{z}_g$.

Classifier-Guided Diffusion Models. Given a set of image data samples $\mathbf{y}$, a diffusion probabilistic model (Ho et al., 2020; Song et al., 2020b) estimates its density $p_{\text{data}}(\mathbf{y})$ by first perturbing the data points in a T-step forward process: $q(\mathbf{y}_t \mid \mathbf{y}_0) := \mathcal{N}(\mathbf{y}_t; \sqrt{\alpha_t}\mathbf{y}_0, (1 - \alpha_t)\mathbf{I})$, where $\{\alpha_t\}_{t=1}^T$ denotes a noise-schedule. In the reverse process, a neural network $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{y}_t, t)$ parameterized by $\boldsymbol{\theta}$ is trained to estimate the introduced noise at each timestep t. On the other hand, classifier-guided diffusion models (Dhariwal & Nichol, 2021) enable conditional generation $p(\mathbf{y} \mid \mathbf{z})$ given a class label $\mathbf{z}$. Formally, given the class label $\mathbf{z}$ and a guidance scale $\gamma > 0$, by Bayes' rule, the conditional distribution of sampled images $\mathbf{y}$ would be: $p^\gamma(\mathbf{y} \mid \mathbf{z}) \propto p(\mathbf{y})p(\mathbf{z} \mid \mathbf{y})^\gamma$. By taking the log-derivative on both sides, this gives:

$$\nabla_{\mathbf{y}_t} \log p^\gamma(\mathbf{y}_t \mid \mathbf{z}) = \underbrace{\nabla_{\mathbf{y}_t} \log p_{\boldsymbol{\theta}}(\mathbf{y}_t)}_{\text{est. by } \boldsymbol{\epsilon}_\theta(\mathbf{y}_t, t)} + \gamma \underbrace{\nabla_{\mathbf{y}_t} \log p_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{y}_t)}_{\text{est. by guidance}}. \tag{1}$$

The unconditional score term on the right-hand side (RHS) is a scaled form of the predicted noise, i.e., $\nabla_{\mathbf{y}_t} \log p(\mathbf{y}_t) = -\boldsymbol{\epsilon}_\theta(\mathbf{y}_t, t)/\sqrt{1 - \alpha_t}$. We propose a novel MI-based classifier parameterized by ϕ to estimate $p(\mathbf{z} \mid \mathbf{y}_t)$, and the second term on the RHS corresponds to its log-derivative.

3 METHODOLOGY

In the following, we first describe how MIG-Vis infers a neural latent subspace via a group-wise disentangled VAE. We then present our approach for visualizing and interpreting visual-semantic variations encoded by each latent group through mutual information-guided diffusion synthesis.

3.1 INFERRING GROUP-WISE DISENTANGLED NEURAL LATENT SUBSPACE

To uncover interpretable visual-semantic structures in neural populations, we use a group-wise disentangled VAE that infers latent subspaces encoding distinct factors of variation. Standard disentangled VAEs (Higgins et al., 2017; Chen et al., 2018) assume each generative factor can be captured by a single independent latent dimension. However, this setting is limited for high-level visual attributes (e.g., category identity, 3D rotation), which require multi-dimensional latent codes.

Hence, we relax the single-dimension constraint by using a group-wise disentangled VAE (Esmaeili et al., 2019; Li et al., 2025), which encourages statistical independency between multi-dimensional neural latent groups. As suggested by previous research works (Locatello et al., 2019), well-disentangled semantic latents seemingly cannot be inferred without (implicit) supervision. We incorporate certain image attributes (i.e., rotation angles and category identity) as priors to inform the latent subspace learning. The supervision labels are concatenated into a vector $\mathbf{u} \in \mathbb{R}^M$, where M

denotes its dimensionality. We decompose the neural latent vector $\mathbf{z}$ into supervised latent groups $\mathbf{z}^{(s)}$ and unsupervised latent groups $\mathbf{z}^{(u)}$, such that $\mathbf{z} = \left[\mathbf{z}^{(s)}, \mathbf{z}^{(u)} \right]^{\top}$. The latent groups within $\mathbf{z}^{(s)}$ are informed by labels, while the groups in $\mathbf{z}^{(u)}$ are inferred without supervision. We propose to optimize the following lower bound of evidence $p_{\boldsymbol{\xi}}(\mathbf{x}, \mathbf{u})$:

$$
\begin{aligned}
\log p_{\boldsymbol{\xi}}(\mathbf{x}, \mathbf{u}) \geq & \underbrace{\mathbb{E}_{q_{\boldsymbol{\psi}}(\mathbf{z}|\mathbf{x})} \left[\log p_{\boldsymbol{\xi}}(\mathbf{x} \mid \mathbf{z}) \right]}_{\text{Neural Reconstruction}} + \underbrace{\mathbb{E}_{q_{\boldsymbol{\psi}}(\mathbf{z}|\mathbf{x})} \left[\log p_{\boldsymbol{\xi}}(\mathbf{u} \mid \mathbf{z}^{(s)}) \right]}_{\text{Weak Label Supervision}} \\
& - \underbrace{\mathbb{D}_{\mathrm{KL}} \left(q_{\boldsymbol{\psi}}(\mathbf{z} \mid \mathbf{x}, \mathbf{u}) \,\|\, p(\mathbf{z}) \right)}_{\text{Prior Regularization}} - \beta \, \underbrace{\mathbb{D}_{\mathrm{KL}} \left(q_{\boldsymbol{\psi}}(\mathbf{z}) \,\middle\|\, \prod_{g=1}^{G} q_{\boldsymbol{\psi}}(\mathbf{z}_g) \right)}_{\text{Partial Correlation}},
\end{aligned}
\tag{2}
$$

where the probabilistic encoder $q_{\boldsymbol{\psi}}(\cdot)$ and decoder $p_{\boldsymbol{\xi}}(\cdot)$ are parameterized by $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$, respectively. $q_{\boldsymbol{\psi}}(\mathbf{z}) = \sum_{n=1}^{N} q_{\boldsymbol{\psi}} \left(\mathbf{z} \mid \mathbf{x}^{(n)} \right) q \left(\mathbf{x}^{(n)} \right)$ is the aggregated posterior over a sample set of size N, $g \in \{1, 2, \ldots, G\}$ denotes the latent group index, and hyperparameter β controls the penalty scale. We note that the group-wise factorized density in the partial correlation (PC) term is intractable, thus we use the importance sampling (IS) estimator (Li et al., 2025) to approximate the PC during training. Importantly, the use of weak-supervision labels and the PC penalty here does not degrade the neural reconstruction quality, as verified by the results in Section 4.

Given an inferred neural latent $\mathbf{z}$, the subsequent target is to characterize the specific visual-semantic factors encoded by a target latent group $\mathbf{z}_g$ within it. The efficacy of traditional latent traversal with neural decoders is limited in practice, as latent dimensions within a group may differ in scale, and simple perturbations to them often result in low-fidelity or uninterpretable images. We present empirical evaluations of these methods in Section 4.

3.2 MUTUAL INFORMATION MAXIMIZATION GUIDANCE

To tackle this challenge, we use a classifier-guided diffusion (Eq. 1). We adopt it due to its explicit conditional gradient term provides scientific interpretability. Our goal is to approximate the two terms on the RHS of Eq. 1. The unconditional score term is estimated by a denoiser $\epsilon_{\theta}(\mathbf{y}_t, t)$. To faithfully capture the semantic features encoded in latent group $\mathbf{z}_g$, we propose an MI-based guidance objective to construct the classifier. From it, we derive the conditional score $\nabla_{\mathbf{y}_t} \log p \left(\mathbf{z} \mid \mathbf{y}_t \right)$.

Mutual Information-Maximization Guidance Objective. Our goal is to characterize the visual-semantic features encoded in each neural latent group, capturing the entire statistical information rather than just first- or second-order statistics. To this end, we propose the following workflow: given an original image $\mathbf{y}$ and a target latent group $\mathbf{z}_g$, we first perturb the original image and then denoise it with the diffusion model by guiding the output to maximize its group-wise mutual information with $\mathbf{z}_g$. The MI between $\mathbf{z}_g$ and $\mathbf{y}$ is given by:

$$
\mathbf{MI}\left(\mathbf{z}_g, \mathbf{y}\right) = \mathbb{E}_{p(\mathbf{z}_g, \mathbf{y})} \left[\log \frac{p\left(\mathbf{z}_g, \mathbf{y}\right)}{p(\mathbf{z}_g) p\left(\mathbf{y}\right)} \right] = \mathbb{E}_{p(\mathbf{z}_g, \mathbf{y})} \left[\log \frac{p\left(\mathbf{y} \mid \mathbf{z}_g\right)}{p\left(\mathbf{y}\right)} \right].
\tag{3}
$$

Note that the MI captures the entire statistical dependence between $\mathbf{z}_g$ and the synthesized image $\mathbf{y}$. Given the fixed $\mathbf{z}_g$, the guidance objective of maximizing the MI enforces the synthesized $\mathbf{y}$ to faithfully contain the semantic information encoded in $\mathbf{z}_g$. Based on Eq. 1, our classifier-guided conditional score is:

$$
\nabla_{\mathbf{y}_t} \log p^{\gamma}\left(\mathbf{y}_t \mid \mathbf{z}\right) = \nabla_{\mathbf{y}_t} \log p_{\boldsymbol{\theta}}(\mathbf{y}_t) + \gamma \nabla_{\mathbf{y}_t} \mathbf{MI}\left(\mathbf{z}_g, \mathbf{y}_t\right).
\tag{4}
$$

Estimation of the Group Mutual Information. However, the mutual information term in Eq. 4 is intractable and notoriously difficult to compute in practice (Hjelm et al., 2018). According to InfoNCE (Oord et al., 2018), we approximate the density ratio portion $\frac{p(\mathbf{y}|\mathbf{z}_g)}{p(\mathbf{y})}$ in Eq. 3 using a neural network $s_{\boldsymbol{\phi}}\left(\mathbf{z}_g, \mathbf{y}\right)$. To construct the InfoNCE loss, we need to get positive and negative samples for $\mathbf{z}_g$. For an image $\mathbf{y}$, a positive sample $\mathbf{z}_g$ comes from $q_{\boldsymbol{\phi}}(\mathbf{z}_g \mid \mathbf{x})$ where $\mathbf{x}$ is the corresponding neural signal for $\mathbf{y}$, while a negative sample comes from $q_{\boldsymbol{\phi}}(\mathbf{z}_g \mid \hat{\mathbf{x}})$ with $\hat{\mathbf{x}}$ unrelated to $\mathbf{y}$. During training, for each $\mathbf{y}$ and batch size B, we collect $\mathcal{Z}_g = \{\mathbf{z}_g^{(1)}, \ldots, \mathbf{z}_g^{(B)}\}$, where $\mathbf{z}_g^{(1)}$ is the positive

sample and $\mathbf{z}_g^{(i)}$ ($i \in \{2, \ldots, B\}$) are negative samples. The noise-contrastive loss $\mathcal{L}_{\mathrm{N}}$ is optimized as follows:

$$\mathcal{L}_{\mathrm{N}}\left(\boldsymbol{\phi}\right) = -\mathbb{E}_{p(\mathbf{z}_g, \mathbf{y})} \left[\log \frac{\exp\left(s_{\boldsymbol{\phi}}\left(\mathbf{z}_g^{(1)}, \mathbf{y}\right)\right)}{\sum_{\mathbf{z}_g^{(i)} \in \mathcal{Z}_g} \exp\left(s_{\boldsymbol{\phi}}\left(\mathbf{z}_g^{(i)}, \mathbf{y}\right)\right)} \right]. \tag{5}$$

Note that $-\mathcal{L}_{\mathrm{N}}$ provides a lower bound (Oord et al., 2018) on the group-wise mutual information, i.e., $\mathbf{MI}(\mathbf{z}_g, \mathbf{y}) \geq \log(B) - \mathcal{L}_{\mathrm{N}}$. $s_{\boldsymbol{\phi}}(\mathbf{z}_g, \mathbf{y})$ specifically approximates the density ratio component of the mutual information $\mathbf{MI}(\mathbf{z}_g, \mathbf{y})$ for the g-th latent group. Since the RHS in Eq. 5 provides a normalized probability function, we have:

$$\nabla_{\mathbf{y}_t} \log p_{\boldsymbol{\phi}}\left(\mathbf{z}_g \mid \mathbf{y}_t\right) = -\nabla_{\mathbf{y}_t} \mathcal{L}_{\mathrm{N}}\left(\boldsymbol{\phi}\right) \approx \nabla_{\mathbf{y}_t} \mathbf{MI}\left(\mathbf{z}_g, \mathbf{y}_t\right). \tag{6}$$

During MI-guided image synthesis, at each step t, we first estimate its noise-free approximation by $\hat{\mathbf{y}}_0 = \left(\mathbf{y}_t - \sqrt{1 - \alpha_t}\,\boldsymbol{\epsilon}_\theta(\mathbf{y}_t, t)\right)/\sqrt{\alpha_t}$, which is then used as input to the $s_{\boldsymbol{\phi}}(\cdot)$ network. Figure 2(B) illustrates our MI guidance objective, which guides image synthesis toward the first latent group $\mathbf{z}_1$ of the target image, with its strength controlled by the guidance scale γ.

3.3 Semantic Image Editing with Deterministic DDIM

In diffusion models, early-step noise perturbations primarily distort semantic attributes (e.g., object and category identity) of the clean image, while largely preserving its structural information (e.g., layout, contours, and color composition) (Choi et al., 2022; Wu et al., 2023). As our goal is to uncover the conceptual semantic features encoded within specific neural latent groups, we employ the image editing approach (Meng et al., 2021). This method stops the noise-perturbation process at an intermediate timestep $t = t' \in (0, T)$ and then initiates the backward synthesis process from that step. This approach is employed to retain the basic structure of the reference image. Otherwise, these foundational components will also be generated from pure noise, making it difficult to interpret the effect of neural semantic features built upon them. Furthermore, to ensure that our interpretation is not compromised by the stochastic sampled noise from the standard reverse process, we use the deterministic DDIM sampler.

Formally, we employ a two-stage deterministic image synthesis process. First, we take an original image $\mathbf{y}_0$ and apply deterministic DDIM Inversion (Song et al., 2020a; Mokady et al., 2023) using the diffusion denoiser $\boldsymbol{\epsilon}_\theta(\cdot)$ from $t = 0$ up to a predefined timestep $t = t'$. This inversion timestep is calibrated to corrupt the semantic attributes of the original image while preserving its structural information. In the second stage, we reverse the process from $t = t'$ to $t = 0$ using a classifier-guided, deterministic DDIM sampling. This reverse procedure is guided by our proposed mutual information (MI) maximization objective, ultimately yielding the synthesized image. This overall process is illustrated in Fig. 2(A). The complete synthesis procedure and a detailed algorithm are provided in Appendix B.

4 Experiments

4.1 Setup

Macaque IT Cortex Dataset. We use single-unit spiking responses from the IT cortex of two macaques (denoted as M1 and M2) during a passive object recognition task (Majaj et al., 2015). Each head-fixed macaque passively viewed a stream of grayscale naturalistic object images while electrophysiological activity in the IT cortex was recorded. Neural activity was recorded from 110 IT channels in M1 and 58 channels in M2. Each image was presented for 100 ms at the center of gaze, and neural responses were measured in a post-stimulus window of 70–170 ms. The stimulus set consisted of 5760 images spanning eight basic-level categories, as illustrated in Fig. 3. To introduce identity-preserving variation, images were simulated via ray-tracing with randomized object position, scale, and pose. As previous studies (Lindsey & Issa, 2024) have shown that IT neurons exhibit a degree of invariance to background context, we segment out the foreground object in each stimulus image.

Figure 3: **Macaques IT cortex dataset.** (A) Experimental Setting. (B) Example images from the eight object categories in the dataset.

Model Settings. (1) Neural VAE. We adopt a two-layer MLP for both the encoder and the decoder. We set the neural latent dimension number to $D = 24$ and
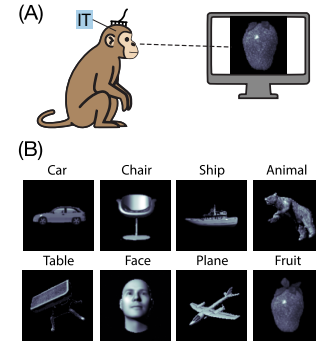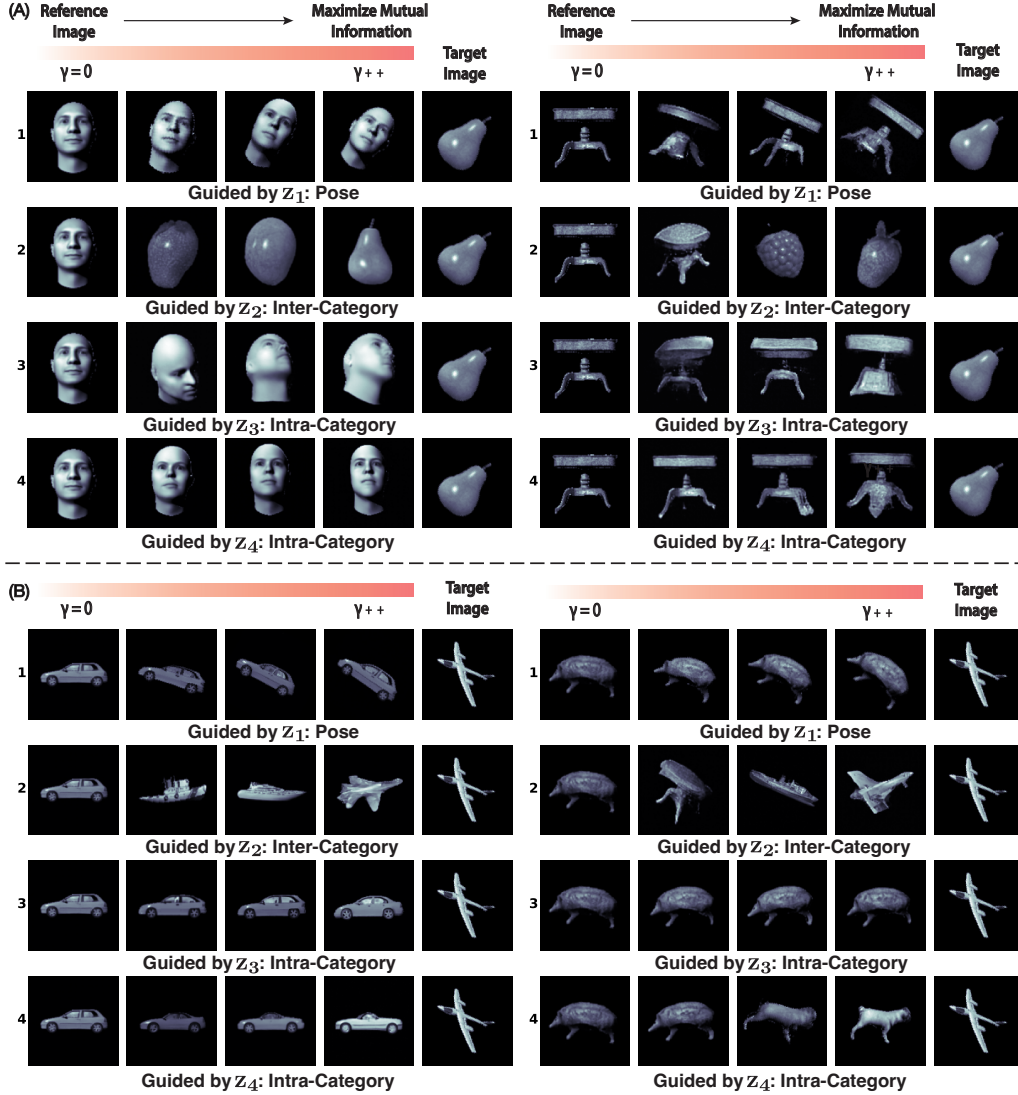
Figure 4: **Synthesized images by MIG-Vis under varying guidance strengths**. **(A)** Results using a frontal-view face and table as the original image, guiding toward a pear. **(B)** Results using a car and hedgehog object as the original image, guiding toward a plane.

partition it into $G = 4$ latent groups, each with a group rank of $D_g = 6$. The first two groups are supervised: Group 1 is informed by the 3D rotation angles provided in the dataset, and Group 2 uses the 8-way one-hot category_id labels. The remaining groups, Group 3 and Group 4, are learned in an unsupervised manner. **(2)** Neural Encoder. We first extract the DINO embeddings (Caron et al., 2021) from the raw images. For DINO, we adopt the `ViT-B/16` architecture and its image embedding size is 384. The density ratio estimator $s_\phi(\mathbf{z}_g, \mathbf{y})$ is implemented as a three-layer convolutional neural network. **(3)** Diffusion model. We downsample images to a resolution of 128×128 and trained an image diffusion model based on a U-Net architecture (Ronneberger et al., 2015). The diffusion step T is set as 150 and the t' is set as 135. Further details on the model architecture and hyper-parameters are listed in the Appendix A.

4.2 DIFFUSION VISUALIZATION OF SEMANTIC SELECTIVITY IN LATENT GROUPS

To conduct the diffusion visualization with MIG-Vis, we select two key components: a **reference image** $\mathbf{y}_0$ that serves as the base for the editing process described in Section 3.3, and a **specific neural latent group** $\mathbf{z}_g$ **from a target image**, which serves as the target latent group for our MI-guided synthesis. We visualize images synthesized by MIG-Vis under varying guidance strengths, $\gamma \in \{2, 5, 10\}$. In Fig. 4(A), a frontal-view face and a table are used as the reference images. Both
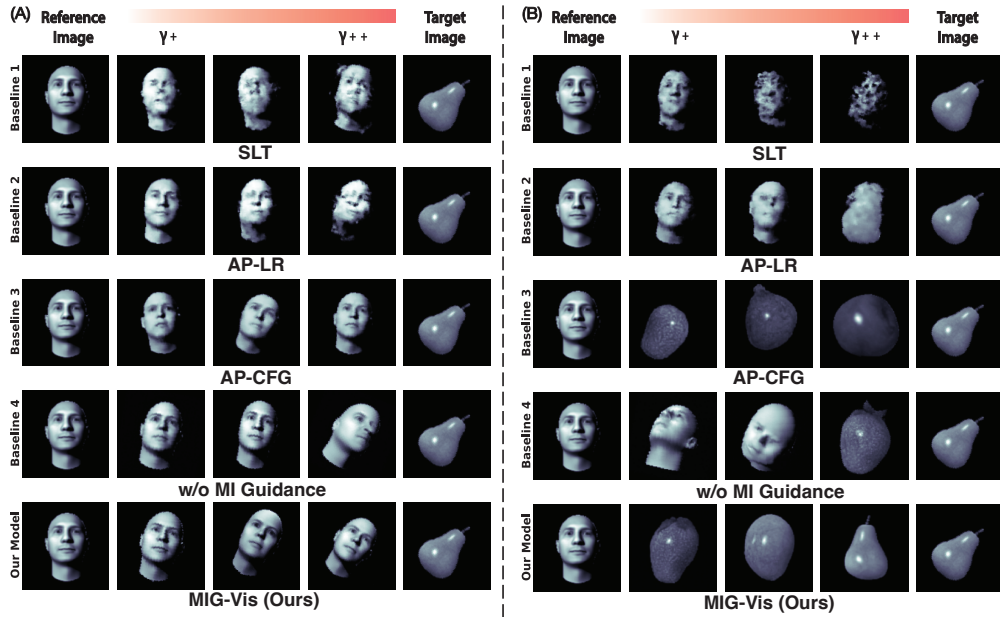
Figure 5: **Comparison of images synthesized by MIG-Vis and baseline methods**. **(A)** Results from probing Latent Group 1 (pose). **(B)** Results from probing Latent Group 2 (inter-category).

of them are guided towards the neural latent groups of a target rotated pear. In Fig. 4(B), a car and a hedgehog serve as the reference images, and both are guided toward the neural latent groups of a target plane. Our visual-semantic selectivity findings are as follows:

(1) Latent Group 1: Intra-Category Pose. Probing Latent Group 1 modulates intra-category, pose-related features, for instance the rotation of the face and car objects. This finding is consistent with the rotation supervision applied to this group's dimensions. Crucially, the objects' category remain stable throughout all these synthesis, indicating that this latent group disentangles pose variations from semantic content.

(2) Latent Group 2: Inter-Category Semantic Attributes. Notably, despite being supervised only with high-level category identity and no explicit semantic features, Latent Group 2 learns to control inter-category semantic attributes. For example, our MI-guidance transforms the face image into a strawberry and subsequently into a pear. We also observe a direct relationship between activation strength and semantic distance; a larger magnitude of γ results in a generated image that is more semantically distinct from the original, demonstrating that the axes of this group subspace encode high-level categorical information.

(3) Latent Group 3 and 4: Intra-Category Content Details. Latent Groups 3 and 4, which were discovered without supervision, both encode intra-category content variations. Specifically, Group 3 primarily modulates the appearance of the face and the chair, whereas Group 4 significantly alters the car and hedgehog's appearance. Importantly, we note that the selectivity of these two groups to intra-category content details is discovered without supervision. Moreover, Group 3 primarily modulates the appearance of the face with minimal effect on the car. Conversely, Group 4 significantly alters the appearance of the car and hedgehog while inducing little change in the face. This emergent separation likely reflects the distinct visual patterns of variation inherent to these two categories.

Furthermore, as shown in both Fig. 4 (A) and (B), when different original images are guided by the latent group of the same reference image, the synthesized results exhibit consistent semantic variations. This observation verifies the efficacy of our MI guidance objective. Additional visualizations using other reference objects or categories are provided in the Appendix C.

4.3 NEURAL GUIDANCE BASELINE METHODS FOR COMPARISON

As prior works are designed for different types of neural data (e.g., fMRI), here we isolate their core neural guidance approaches during image synthesis for a fair comparison with our method:

- **Standard Latent Traversal (SLT)**: a standard approach for latent variable manipulation (Chen
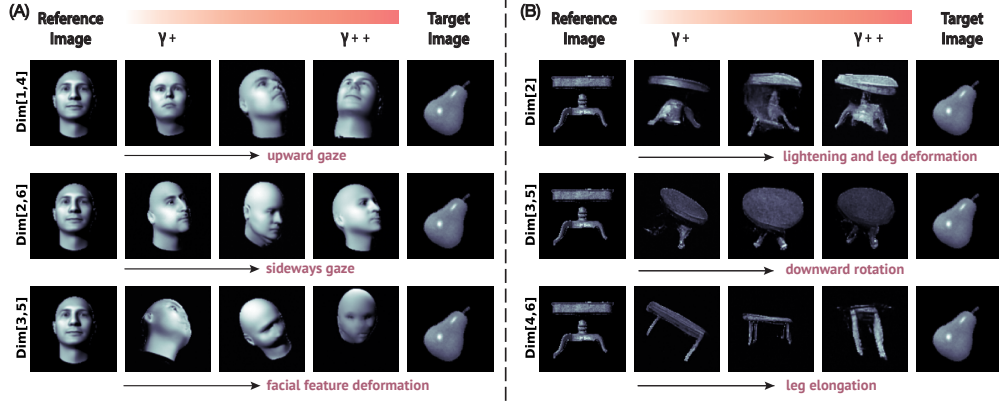
Figure 6: **Pair-wise visual-semantic feature investigation within latent Group 3.** **(A)** Synthesized images of MIG-Vis from probing a frontal-view human face. **(B)** Synthesized images of MIG-Vis from probing a table object. These results demonstrate that various dimension pairs within a latent group further encode distinct and fine-grained visual-semantic attributes.

et al., 2018; Esmaeili et al., 2019), which trains a neural decoder to map images. We perform latent traversal by interpolating a reference image's latent groups toward the latent of the target image.

- **Activation Probing via Linear Regressor (AP-LR)**: the neural manipulation method used in **BrainDIVE** (Luo et al., 2023a), which synthesizes images from diffusion with the goal of maximizing the predicted activation of a target neural region. It combines a pre-trained diffusion model with a linear regressor for neural guidance.

- **Activation Probing via Classifier-Free Guidance (AP-CFG)**: the neural manipulation method used in **BrainACTIV** (Cerdas et al., 2024), which synthesize images guided to maximize or minimize the activation of a target cortical region. Compared to our approach, it employs a classifier-free guidance diffusion (Ho & Salimans, 2022) that jointly models the denoising and guidance gradients.

- **Ours w/o MI Guidance**: an ablation of our method where the MI guidance objective is replaced with the activation (first-order) guidance of the latent towards the reference image's neural latents.

Comparison Results Analyses. In Fig. 5, we compare MIG-Vis with the four baselines introduced above, using a frontal-view face as the reference image. We focus on two representative latent groups for this analysis: Group 1 and Group 2, which have been identified as controlling pose and inter-category factors, respectively. We first find that images manipulated by **SLT** fall outside the natural image distribution and are structurally blurry. **AP-LR** suffers from a similar limitation, as its gradient is derived from a limited linear model. **AP-CFG** performs reasonably well in capturing inter-category variations, Nonetheless, it is incapable of uncovering rotation variations due to the entangled gradient of the denoiser and the guidance. Importantly, for the ablation of our framework (**w/o MI guidance**), we find that it captures rotation semantics effectively. However, for inter-category semantic features, due to the complex underlying structure of this latent subspace, linear activation probing alone is insufficient. This leads to synthesized results that are inconsistent with the realistic categorical variations. In contrast, **our MI guidance objective** is capable of learning inter-category variations and conducting smooth transitions across object instances.

4.4 FINE-GRAINED NEURAL SELECTIVITY WITHIN LATENT GROUP

In Figures 6 and 7(A), we further investigate the semantic structure of the group subspace by probing single-dimension or pair-wise dimensions within the group. Focusing on Group 2 ("inter-category") and Group 3 ("intra-category"), we find that these pair-wise subspaces also encode distinct semantic attributes. For example, within the face subspace of Group 3, shown in Fig. 6(A), different dimension pairs control specific visual features: pair $[1, 4]$ alters the gaze direction, pair $[3, 5]$ modulates the formation of facial features.

An interesting finding from our latent dimension manipulations is that certain combinations exhibit object-invariant effects, while others are object-specific. For instance, the two-dimensional subspace defined by dimensions $[3, 5]$ appears to encode a consistent visual-semantic variation across both face and table objects, one that progressively removes fine-grained detail to produce a more abstract, simplified form. In contrast, other combinations of dimensions, such as those involving dimensions
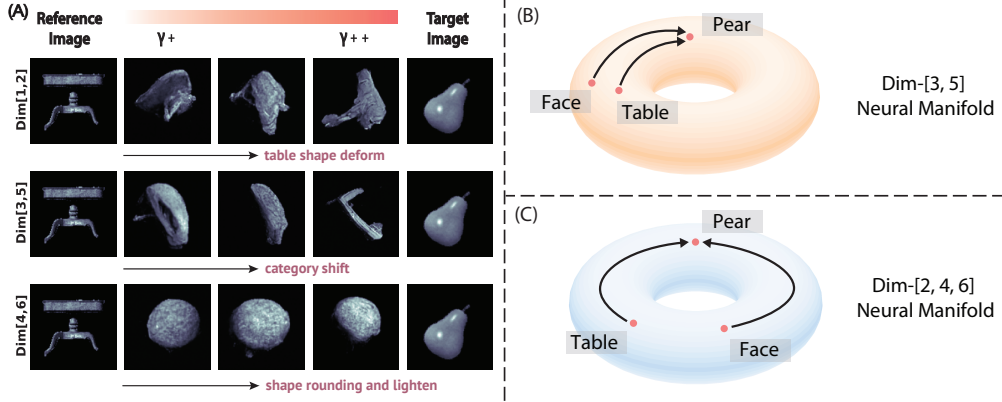
Figure 7: **(A) Pair-wise visual-semantic feature investigation within latent Group 2.** Results of MIG-Vis from probing a table object. **(B) Neural Manifolds of Pair-wise Dimensions.**

$[2, 4, 6]$, encode distinct, object-specific semantics. For the face, this subspace primarily governs pose and rotation. For the table, however, the same subspace modulates surface properties, including color and the thickness of its components.

We think that semantic probing with MIG-Vis can provide insights into the high-dimensional neural space of the higher visual cortex. One hypothesis is that different latent subspaces encode objects and semantic attributes differently. For example, in Fig. 7(B), on a neural manifold defined by dimensions $[3, 5]$, the encodings of "table" and "face" may be close, so MI-guided traversal to "pear" follows similar paths, yielding similar semantic transitions. In contrast, on the manifold of dimensions $[2, 4, 6]$ shown in Fig. 7(C), "table" and "face" are far apart, producing divergent paths to "pear" and thus distinct semantic variations. This represents just one possible hypothesis. With further MIG-Vis manipulations across more objects and dimensions, we can build a richer geometric intuition of how objects are positioned in neural subspaces, their relative distances, and the semantic attributes each area on the neural manifold encodes. MIG-Vis serves as an intuitive tool for visualizing neural manifolds and generating hypotheses, while also pointing toward future neuroscience research on formally characterizing the geometry of neural subspaces in higher visual cortex.

4.5 NEURAL RECONSTRUCTION EVALUATION

To verify that our neural VAE module maintains high-quality neural reconstruction with the partial correlation regularization and weak label supervision, we quantitatively evaluated its performance. Table 1 presents the R-squared (R^2 in %) values for both macaques, comparing our module to standard unsupervised VAE. These results demonstrate that the neural reconstruction quality is well-preserved, incurring only a marginal performance drop compared to the standard VAE. We hypothesize that the weakly-supervised labels induce a rotation of the neural latent subspace while preserving the information necessary for reconstruction. From the R^2 results of ablation studies on our VAE module's two introduced terms, we observe that each has only a minimal negative impact on neural reconstruction.

Table 1: Neural reconstruction quality comparisons on IT cortex dataset. We report the mean and standard deviation of explained variance (R^2 in %) over five runs.

Subject	Method	$R^2(\%) \uparrow$
M1	Standard VAE	78.62 ($\pm$0.58)
	Ours w/o Sup.	76.90 ($\pm$0.53)
	Ours w/o PC.	77.30 ($\pm$0.62)
	Ours	76.58 ($\pm$0.64)
M2	Standard VAE	83.72 ($\pm$0.47)
	Ours w/o Sup.	82.39 ($\pm$0.59)
	Ours w/o PC.	82.21 ($\pm$0.55)
	Ours	81.86 ($\pm$0.51)

5 CONCLUSION

Our contributions can be summarized into the following points: (1) Leveraging advanced neural latent variable models, this is the first work to explore neural representations with semantic selectivity in the higher visual cortex from electrophysiological data. (2) To interpret the visual-semantic features encoded within a specific neural latent group, we use a DDIM-based deterministic image editing approach with a proposed mutual information maximization objective. (3) The edited stim-

uli faithfully demonstrate distinct and high-level visual features, verifying the semantic selectivity within these inferred neural latent groups. Our work offers a critical step toward understanding the compositional, multi-dimensional nature of visual coding in the primate visual cortex.

REFERENCES

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Diego García Cerdas, Christina Sartzetaki, Magnus Petersen, Gemma Roig, Pascal Mettes, and Iris Groen. Brainactiv: Identifying visuo-semantic properties driving cortical selectivity using diffusion-based image manipulation. *bioRxiv*, pp. 2024–10, 2024.

Le Chang and Doris Y Tsao. The code for facial identity in the primate brain. *Cell*, 169(6):1013–1028, 2017.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11472–11481, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, Narayanaswamy Siddharth, Brooks Paige, Dana H Brooks, Jennifer Dy, and Jan-Willem Meent. Structured disentangled representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2525–2534. PMLR, 2019.

Winrich A Freiwald and Doris Y Tsao. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, 330(6005):845–851, 2010.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Chengrui Li, Yunmiao Wang, Yule Wang, Weihan Li, Dieter Jaeger, and Anqi Wu. A revisit of total correlation in disentangled variational auto-encoder with partial disentanglement. *arXiv preprint arXiv:2502.02279*, 2025.

Jack W Lindsey and Elias B Issa. Factorized visual representations in the primate visual system and deep neural networks. *Elife*, 13:RP91685, 2024.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019.

Andrew Luo, Maggie Henderson, Leila Wehbe, and Michael Tarr. Brain diffusion for visual exploration: Cortical discovery using large scale generative models. *Advances in Neural Information Processing Systems*, 36:75740–75781, 2023a.

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*, 2023b.

Najib J Majaj, Ha Hong, Ethan A Solomon, and James J DiCarlo. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience*, 35(39):13402–13418, 2015.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6038–6047, 2023.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention– MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Nicole C Rust and James J DiCarlo. Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area v4 to it. *Journal of Neuroscience*, 30(39):12978–12995, 2010.

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

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Yule Wang, Zijing Wu, Chengrui Li, and Anqi Wu. Extraction and recovery of spatio-temporal structure in latent dynamics alignment with diffusion models. *Advances in Neural Information Processing Systems*, 36:38988–39005, 2023.

Yule Wang, Chengrui Li, Weihan Li, and Anqi Wu. Exploring behavior-relevant and disentangled neural dynamics with generative diffusion models. *Advances in Neural Information Processing Systems*, 37:34712–34736, 2024.

Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. Uncovering the disentanglement capability in text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1900–1910, 2023.

Yudi Xie, Weichen Huang, Esther Alter, Jeremy Schwartz, Joshua B Tenenbaum, and James J DiCarlo. Vision cnns trained to estimate spatial latents learned similar ventral-stream-aligned representations. *arXiv preprint arXiv:2412.09115*, 2024.

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

APPENDIX

A DETAILS OF THE MIG-VIS IMPLEMENTATION

We provide the code and dataset used in this work at the following repository: `https://github.com/yulewang97/MIG-Vis`. The same codebase is also included in the supplementary ZIP file for convenience.

A.1 MODEL TRAINING DETAILS

For the partially disentangled neural VAE, we employ an architecture consisting of a two-layer MLP for both the probabilistic encoder and the probabilistic decoder networks. Each MLP uses ReLU activations. For the neural latent dimension number k to both the two macaques, we set $k = 24$ and partition it into four distinct latent groups, each of dimension 6 (i.e., group rank equals to 6). This group rank setting provides the expressiveness to model complex visual semantic factors (e.g., inter-category variations) within a unique latent group. For latent group 1, we weakly supervise its learning with the 3D rotation angles of the object provided in the dataset; for latent group 2, we use the 8-way category identity labels (e.g., "Animals" as integer label 0, "Boats" as integer label 1); latent groups 3 and 4 are learned in a fully unsupervised manner. We use a learning rate of 1×10^{-3} and set partial correlation penalty `pc_weight` to 5×10^{-5} on both two macaques. We use the Adam Optimizer Kingma & Ba (2014) for optimization.

For the image diffusion model, we downsample the grayscale images to a resolution of $1 \times 128 \times 128$ and train the diffusion model on the resulting inputs. We augment the dataset using random horizontal flips to improve model generalization. We adopt the architecture of the image diffusion denoiser as the U-Net (Ronneberger et al., 2015). We adopt the ϵ-parameterization in our diffusion model for improved training stability, given the relatively small dataset size and diffusion timestep settings. The diffusion timestep is set as 150. The embedding input dimension to the U-Net architecture is set as 64 and the U-Net has three down-sampling and up-sampling layers. The training batch size is set as 64. We train the U-Net denoiser model on $40,000$ iterations with a learning rate of 2×10^{-4}. All experiments are conducted using PyTorch on a compute cluster equipped with NVIDIA A40 GPUs. The training phase of the image diffusion model takes approximately 20 hours. We also apply exponential moving average (EMA) with a decay rate of 0.98 to stabilize training and improve generalization.

A.2 DERIVATION OF THE NOISE-FREE PREDICTION $\hat{\mathbf{y}}_0$

We here elaborate the approximation of the clean image sample $\hat{\mathbf{y}}_0$, as used in Section 3.2. It is computed from the perturbed data point $\mathbf{y}_t$ and the prediction of the denoiser model $\epsilon_\theta(\mathbf{y}_t, t)$.

As in the forward process of diffusion models, given a clean stimulus image $\mathbf{y}_0$, it generates a noisy sample $\mathbf{y}_t$ by gradually adding Gaussian noise through the following closed-form equation:

$$\mathbf{y}_t = \sqrt{\alpha_t}\,\mathbf{y}_0 + \sqrt{1 - \alpha_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{7}$$

The above re-parameterized forward process represents $\mathbf{y}_t$ as a linear combination of the clean image $\mathbf{y}_0$ and the Gaussian noise $\boldsymbol{\epsilon}$. By rearranging the above equation, we obtain an exact expression for the original image sample $\mathbf{y}_0$:

$$\mathbf{y}_0 = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{y}_t - \sqrt{1 - \alpha_t}\,\boldsymbol{\epsilon}\right). \tag{8}$$

As the true noise $\boldsymbol{\epsilon}$ is unknown during the sampling phase, it is approximated by the learned denoiser neural network $\epsilon_\theta(\mathbf{y}_t, t)$. Inserting this estimate into the expression above provides an approximation of the clean image $\mathbf{y}_0$:

$$\hat{\mathbf{y}}_0 = \frac{\mathbf{y}_t - \left(\sqrt{1 - \alpha_t}\right)\boldsymbol{\epsilon}_\theta(\mathbf{y}_t, t)}{\sqrt{\alpha_t}},$$

this formula is used consistently across both DDPM and DDIM frameworks to obtain a noise-free approximation of the original image from the noisy sample at each timestep.

A.3 GROUP-WISE DISENTANGLED NEURAL LATENT SUBSPACE INVESTIGATION

To ensure that the neural VAE module within the MIG-Vis framework maintains high-quality neural reconstruction along with the partial correlation disentanglement regularization and weakly super-vised label guidance, we present the quantitative neural reconstruction results of both two macaques in Table 2, which records the R-squared values (R^2, in %) and RMSE of our module and the standard unsupervised VAE. These results indicate that the neural reconstruction quality is well-preserved, with only a slight performance drop compared to the standard VAE.

A possible explanation is that the weakly-supervised labels rotate the neural latent subspace in a way that preserves most of the information necessary for reconstructing the neural activity. This observation is reasonable, given that these labels can be accurately decoded from the neural activity itself. Furthermore, we evaluate the disentanglement quality of the inferred latent subspace using the widely adopted Mutual Information Gap (MIG) metric (Chen et al., 2018), also reported in Table 2. The results demonstrate that the neural latents learned by MIG-Vis are significantly more disentangled than those of the standard VAE.

Table 2: Performance report of our neural VAE module and a standard VAE on IT cortex neural activity from macaque M1 and M2. We report explained variance (R^2), root mean squared error (RMSE), and mutual information gap (MIG) to assess reconstruction accuracy and latent disentan-glement. Results are averaged over 5 runs; bold numbers indicate the highest MIG score on each subject.

Metrics \ Method	M1		M2	
	Standard VAE	**Ours**	Standard VAE	**Ours**
$R^2(\%)$ ↑	78.62 ($\pm$0.58)	76.58 ($\pm$0.64)	83.72 ($\pm$0.47)	81.86 ($\pm$0.51)
RMSE ↓	48.49 ($\pm$0.39)	50.47 ($\pm$0.39)	40.77 ($\pm$0.28)	43.44 ($\pm$0.35)
MIG(%) ↑	33.27 ($\pm$0.82)	**44.23** ($\pm$0.61)	28.65 ($\pm$0.71)	**49.85** ($\pm$0.55)

B DETAILED MUTUAL-INFORMATION MAXIMIZATION GUIDANCE ALGORITHM

Algorithm 1: Synthesize Stimulus Image for Semantic Latent Group Discovery

---

Input: Target Neural Latent Group $\mathbf{z}_g$, Guidance Scale γ, Sampling Step t'
Output: Synthesized Image Stimulus with Guidance $\tilde{\mathbf{y}}_0$

1 Initiate Original Image Stimulus $\mathbf{y}_0$;

2 **for** $t = 1$ **to** t' **do**

3     $\mathbf{y}_t = \sqrt{\frac{\alpha_t}{\alpha_{t-1}}} \cdot \mathbf{y}_{t-1} + \left(\sqrt{\frac{1}{\alpha_t} - 1} - \sqrt{\frac{1}{\alpha_{t-1}} - 1} \right) \cdot \boldsymbol{\epsilon}_\theta(\mathbf{y}_{t-1}, t)$ ▷ DDIM Inversion

4 **for** $t = t'$ **to** 1 **do**

5     $\hat{\boldsymbol{\epsilon}}(\mathbf{y}_t, t) = \boldsymbol{\epsilon}_\theta(\mathbf{y}_t, t) + \gamma \nabla_{\mathbf{y}_t} \mathcal{L}_{\mathrm{N}}(\boldsymbol{\phi})$ ▷ Mutual Information Guided Gradient ;

6     $\mathbf{y}_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} \cdot \mathbf{y}_t + \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \cdot \hat{\boldsymbol{\epsilon}}(\mathbf{y}_t, t)$ ▷ Guided DDIM Sampling

---

C ADDITIONAL VISUALIZATIONS OF DISENTANGLED SEMANTIC LATENT GROUPS
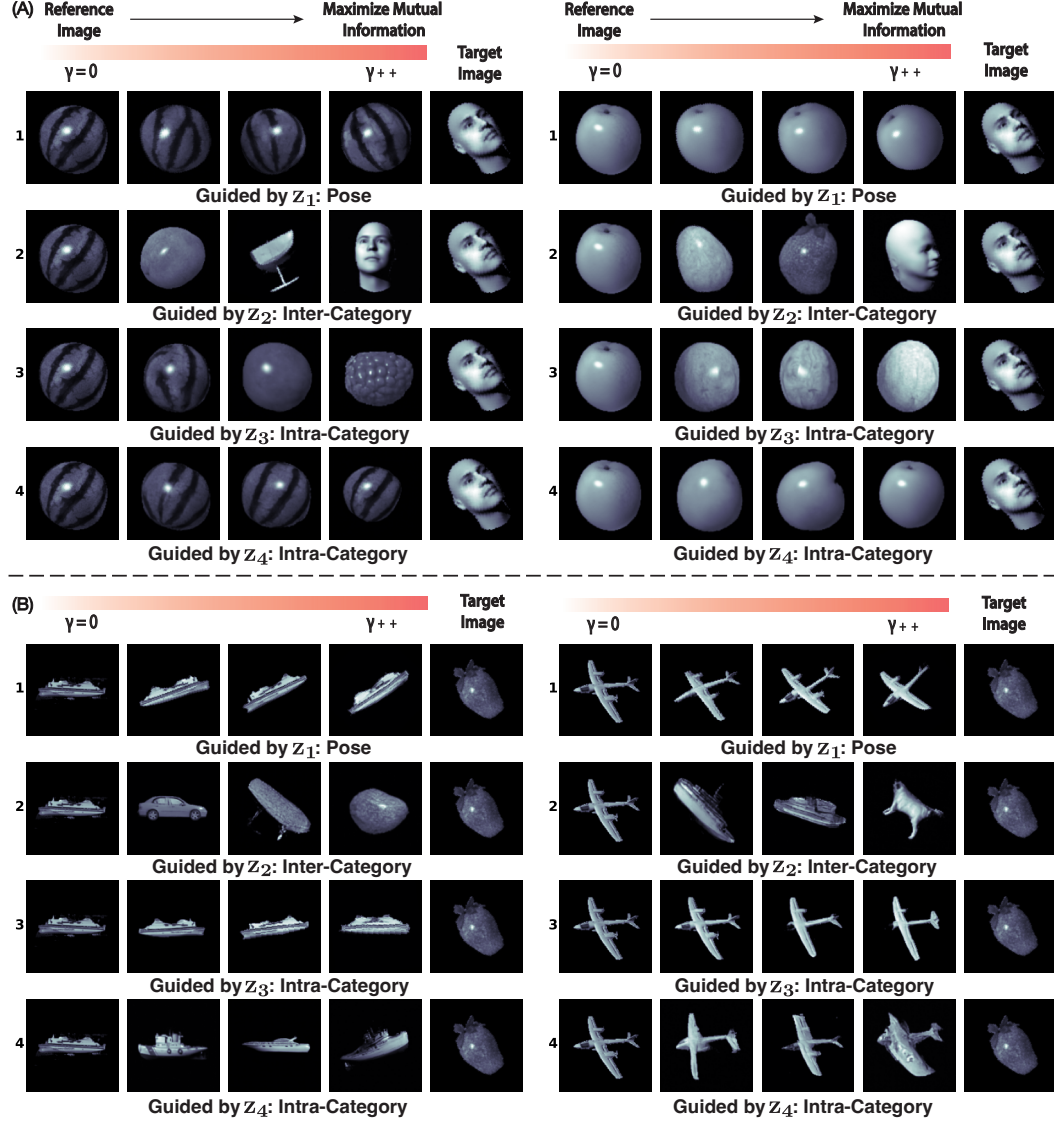

Figure 8: **Synthesized images by MIG-Vis under varying guidance strengths**. **(A)** Results using a watermelon and a peach as the reference image, guiding toward a rotated face. **(B)** Results using a ship and a airplane object as the reference image, guiding toward a strawberry. Group 1 captures pose-related variations (e.g., object rotation), Group 2 controls inter-category semantic transformations, while Groups 3 and 4 govern distinct intra-category content variations. Image transitions along guidance strength γ demonstrate smooth and interpretable semantic information.

Figure 9: **Synthesized images by MIG-Vis under varying guidance strengths**. **(A)** Results using a face and a berry as the reference image, guiding toward a rotated table. **(B)** Results using a chair and a boat object as the reference image, guiding toward a car. Group 1 captures pose-related variations (e.g., object rotation), Group 2 controls inter-category semantic transformations, while Groups 3 and 4 govern distinct intra-category content variations. Image transitions along guidance strength γ demonstrate smooth and interpretable semantic information.