# MindStrata: Learning Robust Multi–Subject Cortical Visual Representations via Stratified Layout–Semantics Decoding

**Anonymous authors**
Paper under double-blind review




Figure 1: MindStrata generates target-, surrounding-, and scene–level reconstructions from visual fMRI responses, capturing object details, contextual regions, and holistic scene layout.

## Abstract

Decoding cortical visual representations that generalize across human subjects remains challenging due to (i) limited modeling of compositional semantics and structured layout in multi–object scenes, (ii) brain functional differences across subjects, and (iii) evaluation protocols dominated by feature–based scores that are often misaligned with human perception. To address these challenges, we propose *MindStrata*, a neuro–cognitively inspired framework for multi–subject brain–to–scene decoding via stratified modeling. *MindStrata* maps fMRI into structured brain–scene representations using two complementary branches: a ventral–like branch decodes multi–level semantics (object→instances→scene) in a CLIP–aligned space, while a dorsal–like branch decodes multi–scale spatial organization (target→surrounding→global) within latent layout space. A Brain–Scene Routing Controller integrates these stratified semantic and spatial cues for structure–guided reconstruction. To mitigate multi–subject variability, we further introduce Consistency–based Functional Alignment, combining subject–specific normalization with a relational consistency loss. We also propose the *Semantic Faithfulness Evaluation Protocol* that evaluates description-, category-, count-, and position–level correctness and aligns more closely with human preferences Experiments demonstrate fine–grained layout and semantic decoding across single–subject, multi–subject, and data–limited settings, reveal ROI–specific spatial localization, and show robust reconstruction performance under progressively reduced fMRI spatial resolution. Code will be publicly released.

## 1 Introduction

*How can we decipher the neural code of visual perception from human brains?* Functional Magnetic Resonance Imaging (fMRI) provides a non–invasive window into cortical activity with high spatial resolution (Naselaris et al., 2011). Together with advances in diffusion models (Ho et al., 2020; Song et al., 2020; Dhariwal & Nichol, 2021; Rombach et al., 2022) and large–scale fMRI datasets (Allen et al., 2022), recent studies (Scotti et al., 2023; Gong et al., 2024; Dai et al., 2025; Beliy et al., 2025) have made substantial progress in reconstructing visual content from brain activity. Yet existing methods still struggle to faithfully decode complex natural scenes involving multiple instances or complex layout (Fig. 2a).

Faithfully decoding complex natural scenes requires more than recovering their overall appearance. Natural scenes are inherently compositional, with visual information spanning multiple levels, from individual objects to contextual relations and holistic scene structure. Recent approaches (Li et al.,



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

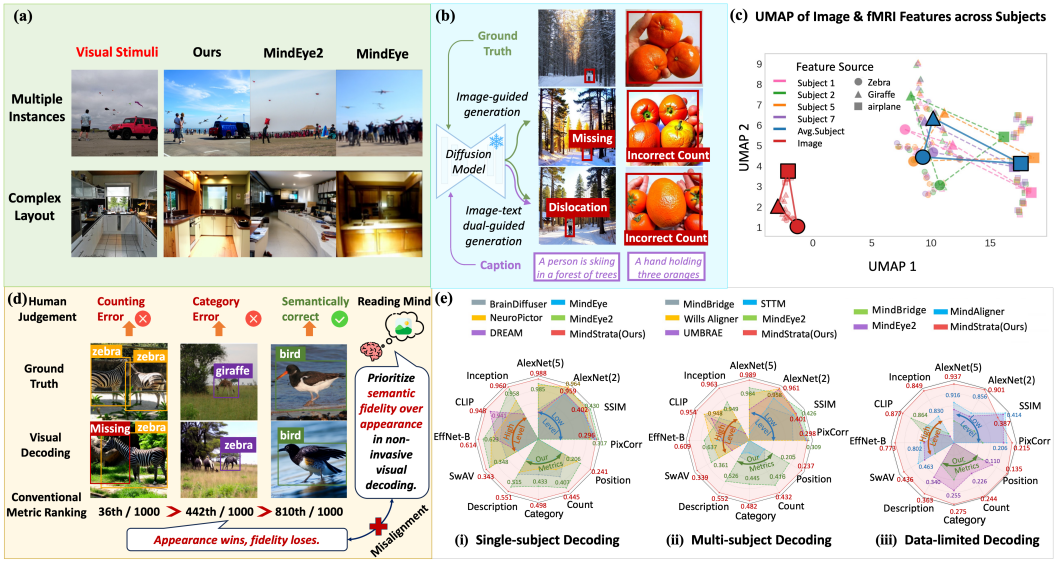



Figure 2: **Analysis of multi–subject cortical visual decoding and performance comparison.** (**a**) Qualitative comparison of decoding scenes with multiple instances or complex layout. (**b**) Failure cases of the diffusion model under image–only or image–text dual guidance. (**c**) UMAP visualization of fMRI and image feature distributions, showing multi–subject heterogeneity and cross–modal structural mismatch. (**d**) Evaluation of 1,000 MindEye (Scotti et al., 2023) outputs revealing the misalignment between semantic faithfulness and conventional metrics. (**e**) Performance comparison with baseline methods across single–subject, multi–subject, and data–limited decoding settings.

2025; Lu et al., 2025) separate semantic and spatial information, while still relying on holistic scene representations that limit interaction between instance semantics and structured layouts. The challenge is further amplified in multi–subject decoding (Wang et al., 2024a; Bao et al., 2025b), where a single model is expected to generalize across human subjects despite variability in cortical responses. Meanwhile, assessing visual reconstructions poses another challenge, as conventional feature–based metrics emphasize global visual similarity and may overlook errors in semantic content. Recent efforts toward semantic–aware evaluation (Park et al., 2026) improve sensitivity to semantic fidelity, yet remain limited in assessing fine–grained properties.

**Key observations.** To understand the causes of these challenges, we investigate current brain decoding methods and observe three limitations in both generation and evaluation. (i) Conditioned on coarse global guidance, the versatile diffusion (Xu et al., 2023) still suffers from instance omission and spatial inaccuracies, although it is widely adopted for visual decoding (Fig. 2b). (ii) The UMAP visualization (Fig. 2c) highlights two alignment failures: persistent multi–subject heterogeneity and a structural mismatch between the aligned fMRI and image feature space. (iii) Conventional high–level metrics exhibit an appearance bias, which may overrate reconstructions that match global appearance while penalizing semantic accuracy (Fig. 2d). Taken together, these observations reveal that faithful multi–subject brain decoding is constrained by both coarse–grained visual representation and unreliable evaluation. This raises a key question: *How can we extract fine–grained visual representations across subjects to ensure semantic and structural faithfulness?*

**Our solutions.** To answer this question, we present **MindStrata**, a stratified visual decoding model going beyond the layout–semantics decomposition. Unlike previous coarse–grained approaches that directly learn holistic representations, our neuro–cognitively inspired framework adopts a dual–branch, multi–level stratified architecture that preserves fine–grained semantic and spatial structure. MindStrata transforms fMRI voxels into a structured brain-–scene representation: a ventral–like branch decodes multi–level semantics (object→instances→scene) in a CLIP–aligned space, while a dorsal–like branch decodes multi–scale spatial layouts (target→surrounding→global) in the diffusion–VAE latent space. A Brain–Scene Routing Controller integrates these stratified semantic and spatial cues into a frozen diffusion model, enabling controllable and globally consistent scene reconstruction.

Additionally, we propose Consistency–based Functional Alignment (CFA) to mitigate multi–subject heterogeneity and cross–modal discrepancies. CFA incorporates subject–specific normalization layers into a unified brain encoder to adapt subject–wise distributions, together with a relational consistency loss that aligns the similarity structures of multi–subject fMRI and CLIP embedding spaces. We further introduce the Semantic Faithfulness Evaluation Protocol (SFEP), which assesses reconstructions along interpretable dimensions including description, categories, counts, and positions.

We conduct extensive experiments to evaluate both layout and semantic decoding. Layout experiments quantify spatial localization across cortical ROIs, while semantic experiments evaluate fine–grained reconstruction under single–subject, multi–subject, and data–limited settings (Fig. 2e). Beyond these standard settings, we further evaluate MindStrata under progressively reduced fMRI spatial resolution to examine whether its stratified architecture learns robust representations that remain informative under spatial degradation. The results consistently demonstrate strong semantic fidelity, spatial accuracy, and robustness across these settings.

To summarize, our contributions are as follows:

- **MindStrata.** We propose a new stratified layout–semantics decoding framework with consistency–based functional alignment to mitigate multi–subject heterogeneity.
- **SFEP.** A Semantic Faithfulness Evaluation Protocol, providing a human–aligned evaluation benchmark that measures description-, category-, count-, and position–level fidelity.
- **Empirical gains.** Extensive experiments demonstrate fine–grained layout and semantic decoding, and robust reconstruction performance across evaluation settings.

## 2 Background and Motivation

In this section, we review related work in Sec. 2.1 with additional discussion in Appendix E. We then introduce the preliminaries in Sec. 2.2 and the motivation for the model design in Sec. 2.3.

### 2.1 Related Work

**Generative Cortical Visual Decoding** Diffusion–based visual decoding initially mapped fMRI voxels into generative or CLIP–aligned latent spaces using linear (Ozcelik & VanRullen, 2023; Takagi & Nishimoto, 2023; Chen et al., 2023; Sun et al., 2023), MLP (Scotti et al., 2023; Gong et al., 2025), or Transformer–based models (Zhou et al., 2024d; Xia et al., 2024b). Later methods further incorporated language or multimodal priors to strengthen semantic guidance and refine visual outputs (Xia et al., 2024a; Huo et al., 2024). Beyond improving perceptual quality, recent studies have increasingly explored how to separate semantic and structural information in brain decoding, motivated by the dual–stream hypothesis of human vision (Goodale et al., 1991). One line of work follows an **implicit** dual–pipeline design, combining low–level structural cues with high–level semantic alignment in CLIP or text space (Ozcelik & VanRullen, 2023; Scotti et al., 2024; Beliy et al., 2025). A more recent line makes this principle **explicit**: DecoFuse (Li et al., 2025) decomposes fMRI representations into semantic, spatial, and motion components, while VCFLOW (Lu et al., 2025) partitions early visual, ventral, and dorsal cortical features for alignment. Beyond diffusion, MindHier (Zhang et al., 2026) increases the autoregressive generation resolution to refine outputs. However, these methods still primarily operate on coarse–grained scene representations, with limited modeling of fine–grained instance semantics, structured layouts, and their interactions.

### 2.2 Preliminaries

**Contrastive learning** aligns representations across modalities by pulling positive pairs together and pushing negative pairs apart (Radford et al., 2021). SoftCLIP instead uses intra–batch similarities of ground–truth CLIP embeddings as soft targets to capture inter–sample semantic relations:

$$\mathcal{L}_{\texttt{SoftCLIP}}(p,t) = -\sum_{i=1}^{N}\sum_{j=1}^{N}\left[\frac{\exp(t_i \cdot t_j/\tau)}{\sum_{m=1}^{N}\exp(\frac{t_i \cdot t_m}{\tau})} \cdot \log\left(\frac{\exp(p_i \cdot t_j/\tau)}{\sum_{m=1}^{N}\exp(\frac{p_i \cdot t_m}{\tau})}\right)\right] \tag{1}$$

where $p$ represents the predicted outputs and $t$ represents the target ground–truth CLIP embeddings in a batch of size $N$, while $\tau$ is a temperature hyperparameter.

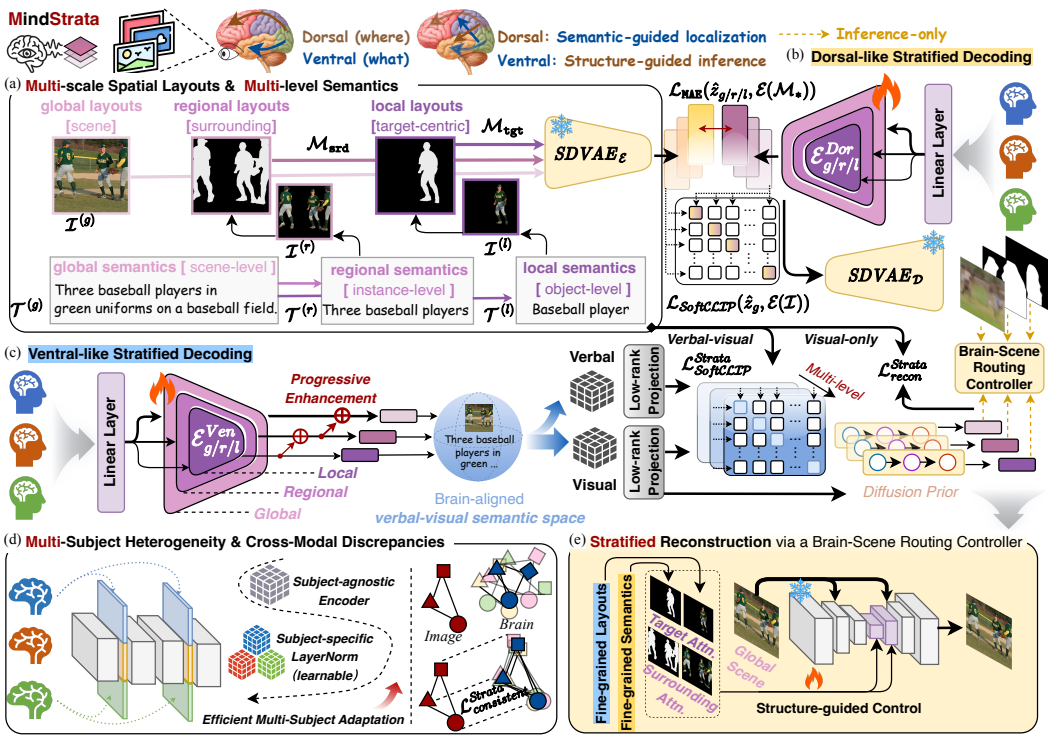



Figure 3: **Overview of MindStrata.** (**a**) Decoupling of multi–scale spatial layouts (global / surrounding / target–centric) and multi–level semantics (scene / instance / object). (**b**) A dorsal–like stratified decoder learns multi–scale layouts with MAE and SoftCLIP guidance. (**c**) A ventral–like stratified decoder refines local–to–global semantics with verbal–visual supervision. (**d**) Learnable subject–specific LayerNorms within a subject–agnostic encoder reduce subject heterogeneity, while structure alignment improves cross–subject consistency. (**e**) A Brain–Scene Routing Controller fuses stratified semantic and spatial cues to enable fine–grained, structure–guided scene reconstruction.

**Diffusion prior** models the conditional distribution of a target embedding given a source embedding from another modality, commonly used in text–to–image pipelines (Ramesh et al., 2022; Lee et al., 2024; Qu et al., 2024). In brain decoding, it transforms fMRI embeddings into aligned CLIP (Radford et al., 2021) embeddings for image synthesis (Scotti et al., 2023).

### 2.3 Motivation

Natural scene reconstruction requires multi–level visual modeling, motivating a stratified extension of the dorsal–ventral view grounded in visual neuroscience.

**Stratified attention mechanism.** Complementing the dual–stream structure, visual processing progresses from attended object information to contextual integration and scene-level representations. Object-based attention enhances representations of attended objects (O'Craven et al., 1999), contextual cues facilitate recognition within surrounding scenes (Palmer, 1975; Bar, 2004), and scene-selective cortical systems support holistic scene perception (Epstein & Baker, 2019). This local-to-global organization motivates the multi-stream design within our dual-branch framework.

## 3 Method

### 3.1 Decoupling Multi–Granularity Layouts and Semantics

To enable stratified layout–semantics decoding, we construct multi–granularity scene supervision from each natural image. Specifically, each scene is decomposed into three complementary levels: a global level describing the overall scene, a regional level capturing surrounding instances and

contextual relations, and a local level focusing on a target object. This yields paired image–text triplets $\{\mathcal{I}^{(g)}, \mathcal{I}^{(r)}, \mathcal{I}^{(l)}\}$ and $\{\mathcal{T}^{(g)}, \mathcal{T}^{(r)}, \mathcal{T}^{(l)}\}$, which provide aligned supervision for global, regional, and local representations (Fig. 3a). Concretely, we use GIT (Wang et al., 2022) to generate the global scene caption, an LLM to parse semantic entities from the caption, and Grounded–SAM (Ren et al., 2024) to obtain their spatial masks. These outputs are then used to construct the regional and local semantic–layout supervision. The full preprocessing pipeline is described in Appendix B.1.

### 3.2 Dorsal–like Stratified Layout Decoding

The Layout Decoding Branch extracts spatial layouts from fMRI voxels to provide positional constraints for semantic reconstruction. The branch comprises three stratified streams (Fig. 3b), operating at global, regional, and local levels. Each stream uses an MLP backbone and a CNN decoder to model multi–scale spatial layouts that dynamically guide the diffusion process.

Specifically, the global stream maps the fMRI embedding $e_h^{(g)}$ to a predicted latent $\hat{z}_g$. This prediction is then aligned with the latent code of the original image $\mathcal{E}(\mathcal{I})$, which is obtained from the pretrained Stable Diffusion Variational AutoEncoder (SDVAE) encoder $\mathcal{E}$. The alignment is supervised by mean absolute error (MAE) loss $\mathcal{L}_{\texttt{MAE}}$ and SoftCLIP loss $\mathcal{L}_{\texttt{SoftCLIP}}$, yielding a coarse structural layout that serves as the initial prior for the diffusion process. The regional and local streams similarly align fMRI features $e_h^{(r)}$ and $e_h^{(l)}$ with the VAE latents of the surrounding mask $\mathcal{E}(\mathcal{M}_{\texttt{srd}})$ and target mask $\mathcal{E}(\mathcal{M}_{\texttt{tgt}})$, respectively, and their predictions $\hat{z}_r$ and $\hat{z}_l$ are trained using only $\mathcal{L}_{\texttt{MAE}}$. Accordingly, the overall spatial loss $\mathcal{L}_{\texttt{spatial}}$ is formulated as:

$$\mathcal{L}_{\texttt{spatial}} = \underbrace{\mathcal{L}_{\texttt{MAE}}(\hat{z}_g, \mathcal{E}(\mathcal{I})) + \lambda_{sc}\mathcal{L}_{\texttt{SoftCLIP}}(\hat{z}_g, \mathcal{E}(\mathcal{I}))}_{\text{Global Stream}} + \underbrace{\mathcal{L}_{\texttt{MAE}}(\hat{z}_r, \mathcal{E}(\mathcal{M}_{\texttt{srd}}))}_{\text{Regional Stream}} + \underbrace{\mathcal{L}_{\texttt{MAE}}(\hat{z}_l, \mathcal{E}(\mathcal{M}_{\texttt{tgt}}))}_{\text{Local Stream}} \tag{2}$$

where $\lambda_{sc}$ is a weighting hyperparameter set to 0.1, with specifications in Appendix D.2. During inference, the spatial masks are predicted directly from fMRI, without using any pretrained detector.

### 3.3 Ventral–like Stratified Semantic Decoding

While the Layout Decoding Branch provides geometric constraints, the Semantic Decoding Branch enriches it with multi–level semantics. As shown in Fig. 3c, the ventral–like stratified decoder progressively refines semantics from local to global levels, guided by verbal–visual supervision. Unlike prior methods (Ozcelik & VanRullen, 2023; Zeng et al., 2024; Gong et al., 2025) that directly project fMRI features into the CLIP image space, our semantic decoder operates in a unified verbal–visual semantic space to better capture the brain's semantic response. Mirroring the Layout Decoding Branch, it disentangles scene semantics into three stratified streams. Details are as follows.

**Verbal–visual semantic alignment.** Inspired by evidence that natural–language–supervised visual representations better capture high–level cortical semantics (Wang et al., 2023), we introduce a unified verbal–visual semantic space that jointly aligns fMRI $\mathcal{X}$, image $\mathcal{I}$, and text $\mathcal{T}$.

For cross–modal alignment, we employ a residual MLP to encode fMRI voxels $\mathcal{X}$ into a 4096-d latent vector $e_h$. The vector is then projected into the shared CLIP semantic space to yield brain–semantic tokens $e_{\mathcal{X}_s}$, with CLIP–compatible dimensions (e.g., $N \times 768$ for ViT–L/14). Low–rank factorized linear projectors (more details in Appendix B.2) further transform $e_{\mathcal{X}_s}$ into an fMRI–text embedding $e_{\mathcal{X}_t}$ and an fMRI–image embedding $e_{\mathcal{X}_i}$ (e.g., $77 \times 768$ and $257 \times 768$ for ViT–L/14).

With these embeddings, we incorporate two complementary losses that jointly enhance semantic consistency and generative fidelity. (i) **Stratified alignment loss:** for cross–modal alignment, the projected embeddings $e_{\mathcal{X}_i}$ and $e_{\mathcal{X}_t}$ are passed through the image alignment head (IAH) and text alignment head (TAH), producing refined embeddings $e_{\mathcal{I}}$ and $e_{\mathcal{T}}$, respectively. We then employ the SoftCLIP loss on stratified levels ($\mathcal{L}^{Strata}_{\texttt{SoftCLIP}}$) to align these outputs with their corresponding ground–truth CLIP embeddings, $c_{\mathcal{I}}$ and $c_{\mathcal{T}}$. The stratified alignment loss ($\mathcal{L}^{Strata}_{\texttt{align}}$) is defined as:

$$\mathcal{L}^{Strata}_{\texttt{align}} = \mathcal{L}^{Strata}_{\texttt{SoftCLIP}}(e_{\mathcal{I}}, c_{\mathcal{I}}) + \mathcal{L}^{Strata}_{\texttt{SoftCLIP}}(e_{\mathcal{T}}, c_{\mathcal{T}}) \tag{3}$$

where $e_{\mathcal{I}}$, $e_{\mathcal{T}}$, $c_{\mathcal{I}}$ and $c_{\mathcal{T}}$ span stratified levels $\{g, r, l\}$. (ii) **Stratified reconstruction loss** ($\mathcal{L}^{Strata}_{\texttt{recon}}$) is used to train diffusion prior modules in each semantic stratum. This objective guides the model to

reconstruct the ground–truth image embeddings $c_{\mathcal{I}}^{(k)}$ from the fMRI–image representation $e_{\mathcal{X}_i}^{(k)}$. Let $\widehat{c_{\mathcal{I}}^{(k)}}$ be the predicted embedding at a specific level, the loss is defined as:

$$\mathcal{L}_{\texttt{recon}}^{Strata} = \sum_{k\in\{g,r,l\}} \|\widehat{c_{\mathcal{I}}^{(k)}} - c_{\mathcal{I}}^{(k)}\|_2^2 \tag{4}$$

**Progressive stream enhancement.** Motivated by the bottom–up integration characteristic of biological visual pathways (Miller, 1999; Rutishauser et al., 2004), we propose a Progressive Stream Enhancement (PSE) strategy to model stratified information flow. Moving beyond a purely parallel architecture, PSE uses additive residual connections to propagate local details to the regional stream and further enrich the global stream with contextual cues, following ($e_h^{(l)} \rightarrow e_h^{(r)} \rightarrow e_h^{(g)}$).

### 3.4 Consistency–based Functional Alignment

Traditional contrastive learning overlooks inter–sample structure (Wu et al., 2024a), leading to inconsistent representations in multi–subject scenarios despite the use of subject–specific adapters. We address this issue through a Consistency–based Functional Alignment (CFA) strategy.

The CFA strategy comprises two key designs, as shown in Fig. 3d. (i) **Subject–specific normalization layers**: motivated by the use of conditional normalization (Huang & Belongie, 2017; Peebles & Xie, 2023; Wu et al., 2024b), we adapt our encoder to individual subjects by making the normalization layers subject–specific. The remaining layers of the encoder are shared across all subjects to learn a common feature representation. (ii) **Relational consistency loss**: inspired by (Zhou et al., 2024d), we introduce the relational consistency loss $\mathcal{L}_{\texttt{cons}}$, which minimizes the discrepancy between the relational structures of fMRI and CLIP embeddings. For each batch, $e_{\mathcal{I}}^{(s)}$ and $e_{\mathcal{T}}^{(s)}$ denote the fMRI–derived image and text embeddings of a subject $s$ respectively, while $c_{\mathcal{I}}^{(s)}$ and $c_{\mathcal{T}}^{(s)}$ represent their corresponding ground–truth CLIP embeddings.

To enforce inter–subject consistency, embeddings from all $S$ subjects are concatenated to construct multi–subject representations $E_{\mathcal{I}}^{S} = \texttt{Concat}(e_{\mathcal{I}}^{(1)}, ..., e_{\mathcal{I}}^{(S)})$ and $E_{\mathcal{T}}^{S} = \texttt{Concat}(e_{\mathcal{T}}^{(1)}, ..., e_{\mathcal{T}}^{(S)})$. The corresponding CLIP embeddings are concatenated in the same order, producing $C_{\mathcal{I}}^{S}$ and $C_{\mathcal{T}}^{S}$. Using $\texttt{Sim}(\cdot)$ to denote the inner–product similarity matrix, we extend this formulation to our stratified triplet representation $\{g, r, l\}$, we define a stratified consistency loss $\mathcal{L}_{\texttt{cons}}^{Strata}$ as:

$$\mathcal{L}_{\texttt{cons}}^{Strata} = \sum_{k\in\{g,r,l\}} \sum_{m\in\{\mathcal{I},\mathcal{T}\}} \left\| \texttt{Sim}(E_m^{S^{(k)}}) - \texttt{Sim}(C_m^{S^{(k)}}) \right\|_1 \tag{5}$$

The overall loss for multi–subject training of the Stratified Semantic Decoding Branch is defined as:

$$\mathcal{L}_{\texttt{sem}} = \mathcal{L}_{\texttt{recon}}^{Strata} + \alpha\mathcal{L}_{\texttt{align}}^{Strata} + \beta\mathcal{L}_{\texttt{cons}}^{Strata} \tag{6}$$

where $\alpha$ and $\beta$ are set to 0.033 and 0.007, with specifications in Appendix D.2.

### 3.5 Brain–Scene Routing Controller

Previous decoding methods (Ferrante et al., 2023; Zeng et al., 2024; Xu et al., 2025) typically rely on pretrained ControlNet (Zhang et al., 2023) to inject text–aligned fMRI features, limiting the use of richer image–modality representations. To overcome this limitation, we propose a new Brain–Scene Routing Controller, which is: (i) **lightweight** ($\sim$ 14M parameters); (ii) **compatible** with image–only conditioned models such as Versatile Diffusion and SDXL; and (iii) **capable of providing multi–granularity semantics–layout guidance** (Fig. 3e). Here we focus on the core design principles of the controller, with detailed architectural specifications in Appendix B.3.

**Brain–inspired attention mechanism.** To address the limitations of coarse global guidance, we introduce a brain–inspired attention mechanism. While standard cross–attention broadcasts global semantics across the entire image, our approach spatially routes specific features to their relevant regions using two specialized modules: **Target Attention** (TA) and **Surrounding Attention** (SA). The TA module focuses on injecting fine–grained instance details into the focal region, whereas

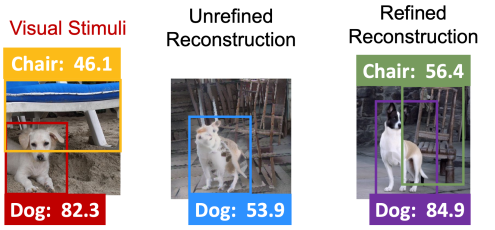
the SA module handles the contextual surroundings. By utilizing the layout masks, these modules constrain attention within boundaries to ensure accurate semantic grounding.

**Stratified semantics–layout interaction.** To generate coherent scene structure, the controller integrates stratified semantic representations with spatial layouts and injects the resulting structure–guided control into the middle and first upsampling cross–attention blocks of the UNet, which are important for semantic and layout formation (Chen et al., 2024; Kamb & Ganguli, 2024).

# 4 Experiments

## 4.1 Semantic Faithfulness Evaluation Protocol

Conventional visual brain decoding metrics rely on pixel–level similarity or feature–space matching, which may fail to faithfully capture semantic correctness. Therefore, we propose the Semantic Faithfulness Evaluation Protocol (SFEP), which evaluates reconstructions at the description, category, count, and position levels. To minimize potential circularity between the models used for constructing training supervision and those used for evaluation, we adopt independent evaluator backbones outside the CLIP/DINO model families, using Emu3 (Wang et al., 2024b) for image captioning and RT–DETR (Zhao et al., 2024) for object detection. Specifically, description–level similarity is measured by the cosine similarity between caption representations of the ground–truth and reconstructed images. Category-, count-, and position–level metrics are computed from detected instances by comparing categories, counts, and bounding–box overlap. Further details are provided in Appendix C.2.




| High-level Metrics | Unrefined | Refined | SFEP Metrics | Unrefined | Refined |
|---|---|---|---|---|---|
| Inception↑ | **0.98** | 0.95 | Description↑ | 0.51 | **0.67** |
| CLIP↑ | **0.99** | 0.93 | Category↑ | 0.67 | **1.00** |
| EffNet↓ | **0.79** | 0.96 | Count↑ | 0.67 | **1.00** |
| SwAV↓ | 0.49 | **0.45** | Position↑ | 0.48 | **0.73** |

Figure 4: **Impact of refinement on semantic fidelity.** Bounding boxes display the class labels and confidence scores.

**Validation of SFEP.** We validate SFEP from both metric consistency perspectives and human preference. First, we revisit the SDXL–based refinement in MindEye2 (Scotti et al., 2024), which improves visual realism and local semantics but may lower traditional feature–based scores. This paradox shows that conventional metrics may penalize semantically improved reconstructions, whereas SFEP captures gains in category-, count-, and position–level correctness (Fig. 4). Details are provided in Appendix C.3. Second, we conduct a user study with 51 participants evaluating reconstructions for 50 visual stimuli from MindStrata and three baselines, yielding 2,550 ranking trials per criterion and 30,600 pairwise comparisons in total. Based on these pairwise preferences, most SFEP metrics align better with human judgments than conventional metrics (Fig. 5).

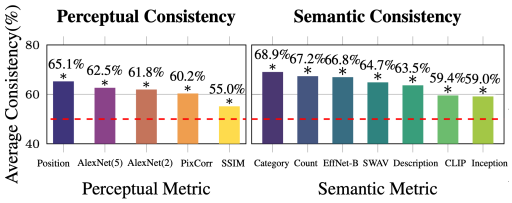



Figure 5: **User study results.** Red lines denote random chance. * indicates significantly above chance (one–sided tests using a binomial generalized linear mixed–effects model, $p < 0.01$).

## 4.2 Main Results

**Layout decoding.** To investigate the feasibility of decoding spatial information from cortical representations and to examine the spatial sensitivity of different Regions of Interest (ROIs), we conduct extensive experiments covering: (i) **ROI selection**, (ii) **mask regions** (target / surrounding), (iii) **mask types**, and (iv) **subject settings**. Fig. 6 presents ROI–wise rankings, with comprehensive details provided in Appendix D.3. Specifically, we utilize fMRI activations from the `Dorsal` and `Nsdgeneral` ROIs to predict spatial masks, which preserve object shape and location. The `Nsdgeneral` ROI refers to the fMRI voxels most responsive to the visual stimuli. Considering that segmentation maps may carry implicit semantic cues, we utilize bounding box maps as a stricter control. Experiments are conducted under single- and multi–subject scenarios.

Table 1: **Quantitative evaluation of visual decoding performance.** For methods based on SDXL, we adopt the refinement procedure following (Scotti et al., 2024). Rows marked † show unrefined results and are shaded gray; each is immediately followed by its refined counterpart. * denotes results taken from (Quan et al., 2024) for multi–subject visual decoding. Red indicates the best performance within each decoding setting.

| Method | Low–Level | | | | High–Level | | | | SFEP (Ours) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PixCorr ↑ | SSIM ↑ | Alex(2) ↑ | Alex(5) ↑ | Incep ↑ | CLIP ↑ | Eff ↓ | SwAV ↓ | Description ↑ | Category ↑ | Count ↑ | Position ↑ |
| ***Single–subject visual decoding*** | | | | | | | | | | | | |
| Brain–Diffuser | 0.254 | 0.356 | 94.2% | 96.2% | 87.2% | 91.5% | 0.775 | 0.423 | - | - | - | - |
| MindEye | 0.301 | 0.328 | 93.8% | 96.9% | 93.9% | 93.6% | 0.653 | 0.373 | 0.515 | 0.433 | 0.407 | 0.172 |
| NeuroPictor | 0.229 | 0.374 | 96.4% | 98.4% | 94.7% | 93.4% | 0.637 | 0.350 | 0.510 | 0.419 | 0.392 | 0.206 |
| DREAM | 0.274 | 0.328 | 93.9% | 96.7% | 93.4% | 94.1% | 0.645 | 0.418 | - | - | - | - |
| MindStrata (w/ VD) | 0.296 | 0.344 | 95.9% | 98.2% | 94.8% | 94.8% | 0.614 | 0.343 | 0.534 | 0.481 | 0.428 | 0.223 |
| MindStrata (w/ SDXL)† | 0.228 | 0.340 | 95.3% | 98.8% | 96.0% | 94.2% | 0.644 | 0.393 | 0.551 | 0.460 | 0.407 | 0.236 |
| MindStrata (w/ SDXL) | 0.283 | 0.402 | 95.8% | 98.6% | 95.5% | 94.7% | 0.622 | 0.372 | 0.548 | 0.498 | 0.445 | 0.241 |
| ***Multi–subject visual decoding*** | | | | | | | | | | | | |
| MindEye* | 0.129 | 0.255 | 84.2% | 89.2% | 84.1% | 85.0% | 0.812 | 0.487 | - | - | - | - |
| MindBridge | 0.148 | 0.283 | 88.3% | 95.8% | 92.7% | 94.8% | 0.700 | 0.412 | 0.479 | 0.395 | 0.377 | 0.143 |
| MindEye2† | 0.243 | 0.347 | 93.5% | 98.4% | 94.9% | 92.2% | 0.651 | 0.370 | 0.526 | 0.386 | 0.372 | 0.205 |
| MindEye2 | 0.309 | 0.426 | 94.2% | 98.3% | 94.4% | 91.4% | 0.637 | 0.361 | 0.519 | 0.445 | 0.416 | 0.201 |
| UMBRAE | 0.283 | 0.341 | 95.5% | 97.0% | 91.7% | 93.5% | 0.700 | 0.393 | - | - | - | - |
| STTM | 0.288 | 0.304 | 92.1% | 95.4% | 92.5% | 92.1% | 0.697 | 0.406 | 0.471 | 0.374 | 0.358 | 0.142 |
| Wills Aligner | 0.271 | 0.328 | 95.8% | 98.0% | 94.3% | 94.8% | 0.649 | 0.373 | - | - | - | - |
| MindStrata (w/ VD) | 0.298 | 0.344 | 96.0% | 98.4% | 95.5% | 95.4% | 0.609 | 0.339 | 0.541 | 0.476 | 0.424 | 0.225 |
| MindStrata (w/ SDXL)† | 0.238 | 0.340 | 95.5% | 98.9% | 96.3% | 94.2% | 0.645 | 0.394 | 0.552 | 0.443 | 0.395 | 0.231 |
| MindStrata (w/ SDXL) | 0.292 | 0.401 | 96.1% | 98.9% | 95.8% | 94.5% | 0.622 | 0.373 | 0.548 | 0.482 | 0.432 | 0.237 |

As shown in Fig. 6, fMRI–based decoding consistently outperforms the random baseline on all metrics, confirming the feasibility of extracting valid spatial layouts from cortical activity. Notably, while `Nsdgeneral` outperforms the `Dorsal` ROI on segmentation maps, the `Dorsal` ROI surpasses `Nsdgeneral` on target–centric bounding box maps. Since `Nsdgeneral` contains semantic–selective voxels, the shift in performance on bounding boxes implies that removing shape bias reveals the positional specialization of the dorsal stream. Furthermore, while multi–subject decoding with `Nsdgeneral` surpasses its single–subject counterpart on target masks, the `Dorsal` stream shows limited multi–subject transferability. These results may suggest that semantic representations are more robust across individuals than dorsal spatial features.

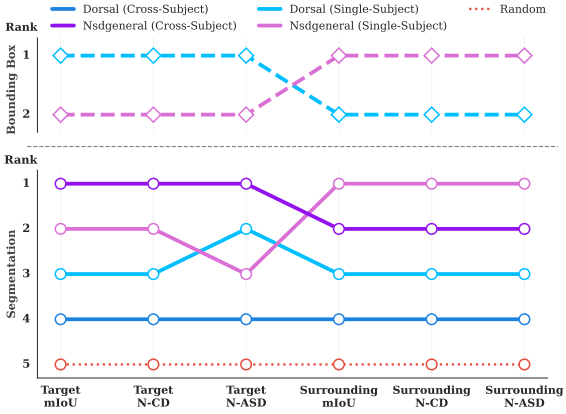



Figure 6: **ROI–wise layout decoding rankings.**

**Semantic decoding.** We evaluate semantic decoding under single- and multi-subject settings, and further examine robustness under limited-data and reduced-resolution conditions. In the multi-subject setting, a single model is trained on four subjects and tested individually. Although multi-subject training quadruples the data, it does not necessarily improve decoding, due to inter-subject brain variability. This observation is well validated by the performance comparison of MindEye across the single- and multi-subject sections. As summarized in the multi-subject section of Tab. 1, MindStrata achieves stronger semantic performance than competing methods while remaining comparable to its single–subject counterpart, demonstrating robust decoding across multiple subjects.

To examine whether MindStrata learns robust representations beyond standard training conditions, we further evaluate it under limited–data and reduced–resolution settings. For data–limited decoding, we train the multi–subject model using only 1 hour of fMRI data. Despite this severe data constraint, MindStrata maintains strong semantic and spatial decoding performance, as shown in Fig. 2e. For reduced–resolution decoding, we progressively apply Gaussian smoothing to fMRI voxels to simulate lower–spatial–resolution medical imaging signals and test whether the learned representations remain informative as spatial resolution decreases. MindStrata remains more robust under increasing spatial degradation, as shown in Fig. 7. More results are provided in Appendix D.5 and D.6.

Further implementation details, such as training cost, are provided in Appendix D.1. Despite adopting a three–stream architecture, our model maintains a parameter budget comparable to the single–stream

Table 2: **Ablations on the modules of MindStrata.** LS: Local Stream; RS: Regional Stream; CFA: Consistency-based Functional Alignment. Red indicates the best performance.

| Method | Low–Level | | | | High–Level | | | | SFEP (Ours) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PixCorr ↑ | SSIM ↑ | Alex(2) ↑ | Alex(5) ↑ | Incep ↑ | CLIP ↑ | EffNet–B ↓ | SwAV ↓ | Description ↑ | Category ↑ | Count ↑ | Position ↑ |
| Backbone & Diffusion Prior | 0.350 | 0.365 | 96.5% | 98.4% | 94.9% | 95.2% | 0.619 | 0.348 | 0.520 | 0.462 | 0.418 | 0.201 |
| + LS | 0.343 | 0.356 | 97.2% | 98.6% | 95.1% | 95.0% | 0.609 | 0.341 | 0.536 | 0.480 | 0.429 | 0.224 |
| + LS + RS | 0.341 | 0.352 | 97.7% | 98.9% | 95.9% | 95.3% | 0.600 | 0.331 | 0.542 | 0.489 | 0.435 | 0.236 |
| + LS + RS + CFA | 0.342 | 0.352 | 97.7% | 99.1% | 96.1% | 96.1% | 0.596 | 0.328 | 0.553 | 0.494 | 0.440 | 0.242 |

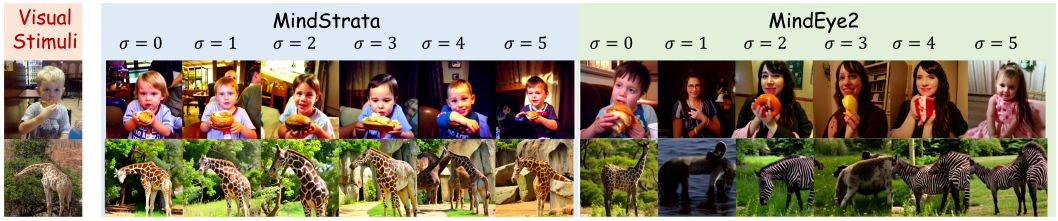



Figure 7: **Qualitative comparison under reduced spatial resolution.** We apply Gaussian smoothing to the fMRI voxels during both training and testing; $\sigma = 0$ denotes the original fMRI voxels, while larger $\sigma \in \{1, 2, 3, 4, 5\}$ simulates progressively lower spatial resolution.

MindEye2 (2,186 M vs. 2,379 M) and outperforms existing approaches across semantic–related metrics, with particularly notable gains on the SFEP metrics. Extensive visual decoding results are provided in Appendix D.4.

### 4.3 Ablation Study

For efficiency, ablation experiments are based on versatile diffusion to verify the effectiveness of each component. Additional comparisons and results are provided in Appendix D.7.

**Ablations on the modules of MindStrata.** Tab. 2 displays our progressive ablation study. The introduction of the local stream consistently improves semantic metrics, with particularly notable gains on the SFEP metrics. This indicates that the local stream effectively guides the model to focus on target objects and their spatial details. After further incorporating the regional stream, the SFEP semantic metrics are comprehensively improved again, with gains observed at the description level (0.520→0.542), category level (0.462→0.489), and position level (0.201→0.236). These results suggest that the contextual semantics provided by the regional stream contribute more substantially to enhancing fine–grained semantic fidelity. Additionally, the consistency–based functional alignment strategy brings overall semantic performance gains despite multi–subject intrinsic variability.

Interestingly, we observe that while Position–level accuracy improves significantly, PixCorr and SSIM decline slightly. However, as shown in Fig. 5, the Position–level metric aligns significantly better with human preference than PixCorr or SSIM, suggesting that our method prioritizes meaningful structural fidelity over rigid pixel–to–pixel correspondence.

## 5 Conclusion and Discussion

We present MindStrata, a new stratified layout–semantics decoding framework with Consistency–based Functional Alignment (CFA). We further introduce the Semantic Faithfulness Evaluation Protocol (SFEP) for interpretable evaluation of semantic and spatial fidelity. Extensive experiments demonstrate fine–grained semantic and layout decoding, ROI–specific spatial localization, and robust reconstruction under multi–subject, data–limited, and reduced–resolution settings.

**Discussion.** Our ventral–like and dorsal–like branches should be interpreted as functional proxies, using the dorsal–ventral distinction only as a coarse prior for separating semantics from spatial organization rather than modeling biological pathways. MindStrata also uses pretrained components such as GIT and Grounded–SAM for offline preprocessing of the visual data, while inference relies solely on fMRI voxels. SFEP adopts evaluator backbones outside the CLIP/DINO families to reduce training–evaluation circularity. Nevertheless, external pretrained components may still influence the construction and quality of the training supervision, which we leave for further investigation.

### AI USE STATEMENT

In this work, we used AI tools to assist with the construction of the stratified training dataset. During offline preprocessing, GIT was used to generate image captions, an LLM to parse semantic entities, and Grounded–SAM to obtain segmentation masks, from which multi–level verbal and visual supervision was derived. We did not use generative AI tools for other research tasks requiring mandatory disclosure, and the remaining required tasks are not applicable to this work.

Additionally, we used generative AI tools to polish the writing and improve readability, and for literature retrieval and discovery, including identifying relevant work on the dorsal–ventral dual–stream hypothesis and the neuroscience literature motivating our stratified attention mechanism. Generative AI was also used to assist in creating the visualization of the user–study results; no other figures in the paper were created or modified with generative AI.

All AI–assisted outputs were reviewed and verified by the authors. In particular, retrieved references were manually checked against their original sources, and AI–assisted visualization did not alter the underlying experimental data or results. We take responsibility for the final content of this work, including text, claims, data, and artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

We conduct a low–risk user study to evaluate the agreement between metric–based evaluation and human preferences. Participant consent was obtained before the study, and the study protocol, participant instructions, and interface details are provided in Appendix C.4. We report only aggregated study results and do not disclose identifiable participant information. The fMRI data used in this work are obtained from publicly available datasets and are used in accordance with their respective data–use policies.

### REPRODUCIBILITY STATEMENT

We provide the details of the proposed MindStrata framework, including the stratified layout–semantics decoding architecture, Consistency–based Functional Alignment, and Brain–Scene Routing Controller in Section 3. Additional details on datasets, preprocessing, training procedures, hyper-parameter settings, evaluation metrics, and ablation studies are provided in the Appendix. Upon acceptance, we will release the full code on GitHub to facilitate reproducibility.

## 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.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. eDiff-I: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Guangyin Bao, Qi Zhang, Zixuan Gong, Zhuojia Wu, and Duoqian Miao. MindSimulator: Exploring brain concept localization via synthetic fMRI. *arXiv preprint arXiv:2503.02351*, 2025a.

Guangyin Bao, Qi Zhang, Zixuan Gong, Jialei Zhou, Wei Fan, Kun Yi, Usman Naseem, Liang Hu, and Duoqian Miao. Wills Aligner: Multi-subject collaborative brain visual decoding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14194–14202, 2025b.

Moshe Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629, 2004.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. MultiDiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023.

Roman Beliy, Amit Zalcher, Jonathan Kogman, Navve Wasserman, and Michal Irani. Brain-IT: Image reconstruction from fMRI via brain-interaction transformer. *arXiv preprint arXiv:2510.25976*, 2025.

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.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9): 10850–10869, 2023.

Yuqin Dai, Zhouheng Yao, Chunfeng Song, Qihao Zheng, Weijian Mai, Kunyu Peng, Shuai Lu, Wanli Ouyang, Jian Yang, and Jiamin Wu. MindAligner: Explicit brain functional alignment for cross-subject visual decoding from limited fMRI data. *arXiv preprint arXiv:2502.05034*, 2025.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5(10): 1097–1107, 2023.

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

Russell A Epstein and Chris I Baker. Scene perception in the human brain. *Annual review of vision science*, 5(1):373–397, 2019.

Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991.

Matteo Ferrante, Tommaso Boccato, Furkan Ozcelik, Rufin VanRullen, and Nicola Toschi. Generative multimodal decoding: Reconstructing images and text from human fMRI. In *Deep Generative Models for Health Workshop NeurIPS 2023*, 2023.

Matteo Ferrante, Tommaso Boccato, Furkan Ozcelik, Rufin VanRullen, and Nicola Toschi. Through their eyes: multi-subject brain decoding with simple alignment techniques. *Imaging Neuroscience*, 2:1–21, 2024.

Zixuan Gong, Guangyin Bao, Qi Zhang, Zhongwei Wan, Duoqian Miao, Shoujin Wang, Lei Zhu, Changwei Wang, Rongtao Xu, Liang Hu, et al. NeuroClips: Towards high-fidelity and smooth fMRI-to-video reconstruction. *Advances in Neural Information Processing Systems*, 37:51655–51683, 2024.

Zixuan Gong, Qi Zhang, Guangyin Bao, Lei Zhu, Rongtao Xu, Ke Liu, Liang Hu, and Duoqian Miao. MindTuner: Cross-subject visual decoding with visual fingerprint and semantic correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14247–14255, 2025.

Melvyn A Goodale, A David Milner, Lorna S Jakobson, and David P Carey. A neurological dissociation between perceiving objects and grasping them. *Nature*, 349(6305):154–156, 1991.

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. In *Advances in neural information processing systems*, 2020.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 1501–1510, 2017.

Jingyang Huo, Yikai Wang, Yun Wang, Xuelin Qian, Chong Li, Yanwei Fu, and Jianfeng Feng. NeuroPictor: Refining fMRI-to-image reconstruction via multi-individual pretraining and multi-level modulation. In *European Conference on Computer Vision*, pp. 56–73. Springer, 2024.

Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012.

Mason Kamb and Surya Ganguli. An analytic theory of creativity in convolutional diffusion models. *arXiv preprint arXiv:2412.20292*, 2024.

Hsin-Ying Lee, Hung-Yu Tseng, and Ming-Hsuan Yang. Exploiting diffusion prior for generalizable dense prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7861–7871, 2024.

Chong Li, Jingyang Huo, Weikang Gong, Yanwei Fu, Xiangyang Xue, and Jianfeng Feng. DecoFuse: Decomposing and fusing the "What", "Where", and "How" for brain-inspired fMRI-to-video decoding. *arXiv preprint arXiv:2504.00432*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jingyu Lu, Haonan Wang, Qixiang Zhang, and Xiaomeng Li. A cognitive process-inspired architecture for subject-agnostic brain visual decoding. *arXiv preprint arXiv:2511.02565*, 2025.

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, 2023.

Andrew Luo, Maggie Henderson, Michael Tarr, and Leila Wehbe. Brainscuba: Fine-grained natural language captions of visual cortex selectivity. In *International Conference on Learning Representations*, volume 2024, pp. 1336–1372, 2024a.

Andrew F Luo, Jacob Yeung, Rushikesh Zawar, Shaurya Dewan, Margaret M Henderson, Leila Wehbe, and Michael J Tarr. Brain mapping with dense features: Grounding cortical semantic selectivity in natural images with vision transformers. *arXiv preprint arXiv:2410.05266*, 2024b.

David Mayo, Christopher Wang, Asa Harbin, Abdulrahman Alabdulkareem, Albert Shaw, Boris Katz, and Andrei Barbu. BrainBits: How much of the brain are generative reconstruction methods using? *Advances in Neural Information Processing Systems*, 37:54396–54420, 2024.

John Mazziotta, Arthur Toga, Alan Evans, Peter Fox, Jack Lancaster, Karl Zilles, Roger Woods, Tomas Paus, Gregory Simpson, Bruce Pike, et al. A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (ICBM). *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1412):1293–1322, 2001.

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

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4296–4304, 2024.

Mu Nan, Muquan Yu, Weijian Mai, Jacob S Prince, Hossein Adeli, Rui Zhang, Jiahang Cao, Benjamin Becker, John A Pyles, Margaret M Henderson, et al. Meta-learning in-context enables training-free cross subject brain decoding. *arXiv preprint arXiv:2604.08537*, 2026.

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

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Kathleen M O'Craven, Paul E Downing, and Nancy Kanwisher. fMRI evidence for objects as the units of attentional selection. *Nature*, 401(6753):584–587, 1999.

Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023.

Stephen E Palmer. The effects of contextual scenes on the identification of objects. *Memory & cognition*, 3(5):519–526, 1975.

Juhyeon Park, Peter Kim, Jiook Cha, Shinjae Yoo, and Taesup Moon. SEED: Towards more accurate semantic evaluation for visual brain decoding. In *International Conference on Learning Representations*, volume 2026, pp. 11967–11986, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Jacob S Prince, Ian Charest, Jan W Kurzawski, John A Pyles, Michael J Tarr, and Kendrick N Kay. Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *Elife*, 11:e77599, 2022.

Yunpeng Qu, Kun Yuan, Kai Zhao, Qizhi Xie, Jinhua Hao, Ming Sun, and Chao Zhou. XPSR: Cross-modal priors for diffusion-based image super-resolution. In *European Conference on Computer Vision*, pp. 285–303. Springer, 2024.

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.

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

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded SAM: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025, 1999.

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.

Ueli Rutishauser, Dirk Walther, Christof Koch, and Pietro Perona. Is bottom-up attention useful for object recognition? In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pp. II–II. IEEE, 2004.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

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.

Guobin Shen, Dongcheng Zhao, Xiang He, Linghao Feng, Yiting Dong, Jihang Wang, Qian Zhang, and Yi Zeng. Neuro-vision to language: Enhancing brain recording-based visual reconstruction and language interaction. *Advances in Neural Information Processing Systems*, 37:98083–98110, 2024.

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.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023.

Alexis Thual, Quang Huy Tran, Tatiana Zemskova, Nicolas Courty, Rémi Flamary, Stanislas Dehaene, and Bertrand Thirion. Aligning individual brains with fused unbalanced gromov wasserstein. *Advances in neural information processing systems*, 35:21792–21804, 2022.

Aria Y Wang, Kendrick Kay, Thomas Naselaris, Michael J Tarr, and Leila Wehbe. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, 5(12):1415–1426, 2023.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

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

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b.

Brian R White and Joseph P Culver. Quantitative evaluation of high-density diffuse optical tomography: in vivo resolution and mapping performance. *Journal of biomedical optics*, 15(2): 026006–026006, 2010.

Sitong Wu, Haoru Tan, Zhuotao Tian, Yukang Chen, Xiaojuan Qi, and Jiaya Jia. SaCo loss: Sample-wise affinity consistency for vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27358–27369, 2024a.

Xiaoyang Wu, Zhuotao Tian, Xin Wen, Bohao Peng, Xihui Liu, Kaicheng Yu, and Hengshuang Zhao. Towards large-scale 3D representation learning with multi-dataset point prompt training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19551–19562, 2024b.

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, 2024a.

Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. Umbrae: Unified multimodal brain decoding. In *European Conference on Computer Vision*, pp. 242–259. Springer, 2024b.

Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. BoxDiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7452–7461, 2023.

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.

Yangyang Xu, Bangzhen Liu, Wenqi Shao, Yong Du, Shengfeng He, and Tingting Zhu. Cross-subject mind decoding from inaccurate representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15066–15075, 2025.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, et al. Qwen2. 5-1m technical report. *arXiv preprint arXiv:2501.15383*, 2025.

Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. ReCo: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14246–14255, 2023.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Muquan Yu, Mu Nan, Hossein Adeli, Jacob Prince, John A Pyles, Leila Wehbe, Maggie Henderson, Michael Tarr, and Andrew Luo. Meta-learning an in-context transformer model of human higher visual cortex. *Advances in Neural Information Processing Systems*, 38:168122–168180, 2026.

Bohan Zeng, Shanglin Li, Xuhui Liu, Sicheng Gao, Xiaolong Jiang, Xu Tang, Yao Hu, Jianzhuang Liu, and Baochang Zhang. Controllable mind visual diffusion model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 6935–6943, 2024.

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.

Xu Zhang, Ruijie Quan, Wenguan Wang, and Yi Yang. Moving beyond diffusion: Hierarchy-to-hierarchy autoregression for fMRI-to-image reconstruction. In *International Conference on Learning Representations*, volume 2026, pp. 68490–68515, 2026.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. DETRs beat YOLOs on real-time object detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16965–16974. IEEE, 2024.

Dewei Zhou, You Li, Fan Ma, Zongxin Yang, and Yi Yang. MIGC++: Advanced multi-instance generation controller for image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024a.

Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. MIGC: Multi-instance generation controller for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6818–6828, 2024b.

Dewei Zhou, Ji Xie, Zongxin Yang, and Yi Yang. 3DIS: Depth-driven decoupled instance synthesis for text-to-image generation. *arXiv preprint arXiv:2410.12669*, 2024c.

Qiongyi Zhou, Changde Du, Shengpei Wang, and Huiguang He. CLIP-MUSED: CLIP-guided multi-subject visual neural information semantic decoding. *arXiv preprint arXiv:2402.08994*, 2024d.

## APPENDIX OVERVIEW

This material provides supplementary details to the main paper, including the following sections:


- (A) **Dataset Elaboration**
  - (A.1) NSD Dataset
  - (A.2) Preprocessing
- (B) **Method Details**
  - (B.1) Decoupling Scene Semantics
  - (B.2) Low–Rank Factorized Linear Projector
  - (B.3) Brain–Scene Routing Controller
- (C) **Metric Details**
  - (C.1) Existing Evaluation Methods
  - (C.2) Semantic Faithfulness Evaluation Protocol
  - (C.3) Evaluation Paradox Analysis
  - (C.4) User Study
- (D) **Experiment Details**
  - (D.1) Implementation Details
  - (D.2) Hyperparameter Selection
  - (D.3) Layout Decoding
  - (D.4) Semantic Decoding
  - (D.5) Data–limited Decoding
  - (D.6) Impact of Spatial Resolution
  - (D.7) Ablation Study
  - (D.8) Stratified Semantic Decoding
  - (D.9) Fine–Grained Brain Concept Localization
- (E) **Related Work**
  - (E.1) Controllable Image Diffusion
  - (E.2) Multi–Subject Brain Alignment


## A DATASET ELABORATION

### A.1 NSD DATASET

We conduct our brain decoding experiments on the widely–used Natural Scenes Dataset (NSD) (Allen et al., 2022), which contains 7-Tesla fMRI scans from eight subjects viewing 9,000–10,000 unique natural images (totaling 22,000-30,000 trials) from the MS–COCO dataset. For region–specific analyses, we adopt two standard Region of Interest (ROI) definitions provided by the dataset authors. 1) `nsdgeneral`: a broad visual–cortex mask defined by the intersection of a cortical mask and a viewer–agnostic functional responsiveness mask. It covers the posterior visual areas, spanning early visual cortex (dorsal and ventral subdivisions) and higher visual cortex. 2) `streams`: a finer parcellation dividing the cortex into the early visual cortex (V1-V3), mid–level regions, and higher–level functional streams. `streams` include the ventral stream for object recognition and the lateral and parietal streams involved in spatial processing and motion perception.

### A.2 PREPROCESSING

Following prior work (Wang et al., 2024a; Gong et al., 2025; Quan et al., 2024; Bao et al., 2025b; Dai et al., 2025), we select the four subjects (1, 2, 5, and 7) who completed all 40 scan sessions and use 1.8mm voxel resolution beta–activations from GLMSingle (Prince et al., 2022) with `nsdgeneral` mask. Each session consists of 750 fMRI trials, where each image was presented for three trials. For preprocessing, we apply z–score normalization to the fMRI activations on a per–session basis. During

training, the three fMRI trials for the same image are treated as individual samples. For inference, we average the responses across the three repetitions. This protocol yields a training set of 27,000 samples and a test set of 1,000 samples for each selected subject.

Table 3: **Prompt template for entity extraction.** The prompt template is used to extract semantic entities from image captions using Qwen2.5-14B–Instruct. We explicitly instruct the model to retain "stuff" while discarding abstract scene descriptors.

**System Prompt for Entity Extraction**

**Objective:** To extract all salient and tangible object nouns from a given image caption.
**Instructions:**
1. **Identify Tangible Entities:** Extract all concrete, physical elements. This includes countable objects (e.g., animals, vehicles, furniture) and uncountable semantic "stuff" (e.g., water, grass, beach).
2. **Include Compound Nouns:** Capture both simple nouns (e.g., `clock`) and compound nouns (e.g., `baseball player`).
3. **Filter Abstract Concepts:** Exclude modifiers, quantifiers, and prepositions. Crucially, omit abstract nouns and generic scene descriptors (e.g., `view`, `background`, `scenery`) that serve to frame the image rather than denote a specific physical entity.
4. **Output Format:** Return the extracted nouns as a single string, with each noun separated by " . " and the entire string terminating with a period.

**Few–shot Examples:**
- Caption: "three giraffes in a zoo enclosure with rocks"
  Objects: "giraffes . rocks ."
- Caption: "a plane flying with buildings in the background."
  Objects: "plane . buildings ."
- Caption: "a large building with a clock on top of it."
  Objects: "building . clock ."
- Caption: "a kitchen with a window, sink, and counter."
  Objects: "window . sink . counter ."
- Caption: "a vase with red roses on top of a piano."
  Objects: "vase . roses . piano ."

**Input Caption:** "{caption}"
**Extracted Objects:**

# B Method Details

## B.1 Decoupling Scene Semantics

We construct stratified semantic and layout supervision through a lightweight offline preprocessing pipeline. **Caption generation is optional**: because NSD images originate from MS–COCO, this step can be replaced by the original COCO captions. In our implementation, we use GIT–generated captions $\mathcal{T}^{(g)}$ (Wang et al., 2022) mainly to avoid additional caption cleaning, which takes 0.18s per sample. Given the global caption, we employ Qwen2.5-14B–Instruct (Yang et al., 2025) as a semantic parser to extract stratified semantic cues, taking 0.08s per sample. Our objective is to isolate tangible entities, covering both countable objects (e.g., "dog" and "plane") and uncountable stuff (e.g., "water" and "grass"), while filtering out abstract concepts or generic scene descriptors (e.g., "view", "picture", "background") that lack specific spatial grounding. We facilitate this by providing the LLM with detailed instructions and few–shot examples. The full system prompt is detailed in Table 3.

Following extraction, the output string is parsed into a sequence of candidate entities. We then employ Grounded–SAM (Ren et al., 2024) to generate segmentation masks and binarize them into layout supervision, taking 0.63s per sample. To ensure robust view construction, we apply three key post–processing steps: (1) **Convex Hull Filling**: To address the issue of hollow segmentation common in complex objects, we implement convex hull–based semantic filling; (2) **Morphological Smoothing**: We apply morphological operations to refine mask boundaries; and (3) **Size Filtering**: We discard any entity whose bounding box exceeds 95% of the total image area to prevent global

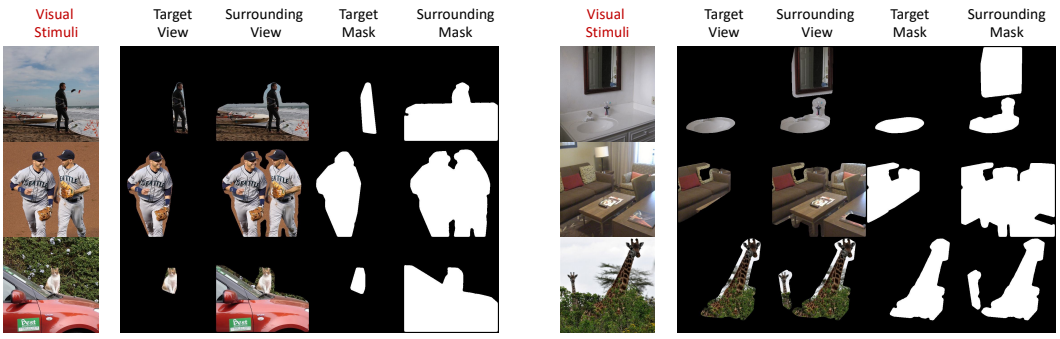



Figure 8: **Examples of decoupling scene semantics.** The mask image is generated by binarizing the segmentation map.

scene descriptions from dominating local views. Finally, the first valid entity in the sequence is selected to generate the target view $\mathcal{I}^{(l)}$, while the entire set of valid entities is utilized to construct the surrounding view $\mathcal{I}^{(r)}$. Overall, the full preprocessing requires less than 1 second per sample and is performed only once offline. We will release both the preprocessing code and the processed data for reproducibility. The segmentation examples are shown in Fig. 8.

It is worth noting that external tools like Grounded–SAM and the LLM are strictly employed to construct stratified supervision signals during training. During inference, the visual reconstruction is solely decoded from fMRI voxels without accessing any image or text ground truth, ensuring the results reflect genuine neural decoding capabilities.

### B.2 Low–Rank Factorized Linear Projector

As described in the main text, the brain–semantic tokens $e_{\mathcal{X}_s} \in \mathbb{R}^{N\times D}$ serve as the intermediate representation, where $N$ denotes the number of brain tokens and $D = 768$ is the feature dimension for ViT–L/14. To map $e_{\mathcal{X}_s}$ to the target CLIP embeddings $e_{\mathcal{X}_k} \in \mathbb{R}^{M_k\times D}$, we employ the Low–Rank Factorized Linear Projector. Here $k \in \{i, t\}$ denotes the image and text modality respectively, with $M_i = 257$ representing the image sequence length and $M_t = 77$ representing the text sequence length.

A standard dense projection from the brain–semantic tokens to the CLIP embeddings would require a parameter matrix of size $(N \cdot D) \times (M_k \cdot D)$. Such a high–dimensional mapping is computationally expensive and prone to overfitting given the limited availability of fMRI data. Therefore, we decompose the projection into two decoupled linear transformations defined as

$$e_{\mathcal{X}_k} = \left(\mathbf{W}^{(k)}_{seq} e_{\mathcal{X}_s}\right) \mathbf{W}^{(k)}_{chn}, \tag{7}$$

where $\mathbf{W}^{(k)}_{seq} \in \mathbb{R}^{M_k\times N}$ is the Token–Mixing matrix that operates along the sequence dimension to resample the $N$ brain–semantic tokens into $M_k$ tokens, effectively aligning the sequence dimensions between the brain and CLIP space. The matrix $\mathbf{W}^{(k)}_{chn} \in \mathbb{R}^{D\times D}$ serves as the Channel–Mixing matrix operating along the feature dimension to align the semantic distributions. This factorized design imposes a low–rank constraint that reduces the parameter complexity from $\mathcal{O}(NM_kD^2)$ to $\mathcal{O}(NM_k + D^2)$, thereby facilitating efficient learning while mitigating overfitting.

### B.3 Brain–Scene Routing Controller

Here we present a comprehensive framework diagram (Fig. 9) to intuitively illustrate the stratified routing mechanism and the dynamic interaction between the global, regional, and local streams. Complementing this visualization, we detail the theoretical formulation of the Brain–Scene Routing Controller as follows.

**Brain–inspired attention mechanism.** Standard image–to–image pipelines employ cross–attention to condition the latent feature maps $\mathbf{f}$ on the guidance from a CLIP image embedding. However, this

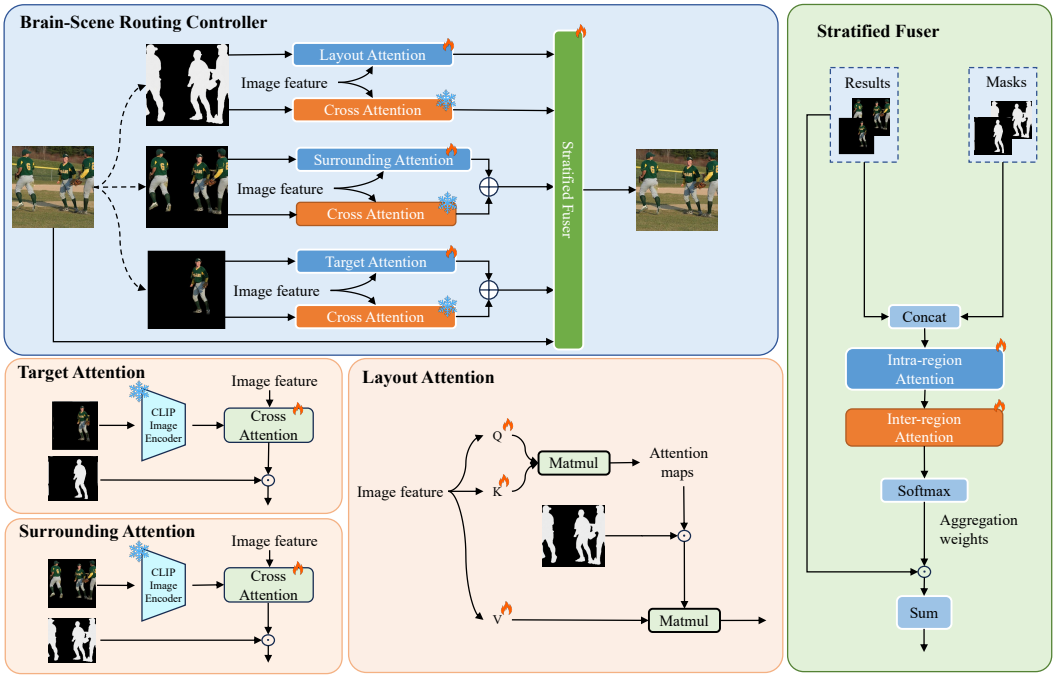



Figure 9: **Overview of the Brain–Scene Routing Controller.** It employs Target, Surrounding and Layout attention modules to extract specific visual cues. The **Stratified Fuser** then integrates these representations to compute aggregation weights. Bottom panels detail the specific designs of each attention sub–module.

coarse global guidance often fails to preserve fine–grained details (Fig. 2a). To address this, our brain–scene routing controller implements a brain–inspired attention mechanism with stratified and trainable modules: **Target Attention** (TA) for local semantics and **Surrounding Attention** (SA) for contextual semantics.

To enable this fine–grained control, we first encode the global scene, surrounding context, and local target images to obtain the corresponding CLIP embeddings: $\mathbf{c}_\mathcal{I} = [c_\mathcal{I}^{(g)}, c_\mathcal{I}^{(r)}, c_\mathcal{I}^{(l)}]$. Inside the UNet, the feature maps $\mathbf{f}$ first interact with the global embedding $c_\mathcal{I}^{(g)}$ in the pretrained cross–attention layers. Subsequently, our TA and SA modules provide fine–grained guidance by operating on the regional $c_\mathcal{I}^{(r)}$ and local $c_\mathcal{I}^{(l)}$ embeddings. Crucially, these modules incorporate explicit spatial constraints by element–wise multiplication ($\odot$) of the attention output with spatial masks $\mathbf{M} = [\mathcal{M}_{\texttt{srd}}, \mathcal{M}_{\texttt{tgt}}]$ derived from our Layout Decoding Branch. The general operation is formulated as:

$$\mathcal{A}(\mathbf{f}, \mathbf{c}, \mathcal{M}) = \text{Softmax}\left(\frac{Q(\mathbf{f})K(\mathbf{c})^\top}{\sqrt{d}}\right) V(\mathbf{c}) \odot \mathcal{M} \tag{8}$$

where $(\mathbf{c}, \mathcal{M}) = (c_\mathcal{I}^{(l)}, \mathcal{M}_{\texttt{tgt}})$ and $(c_\mathcal{I}^{(r)}, \mathcal{M}_{\texttt{srd}})$ denote TA and SA, respectively. Fine–grained attention outputs are then separately added to the global attention output. This biologically inspired dual–module design provides precise, spatially–aware guidance while minimizing the computational overhead and redundancy of multiple independent object attention blocks.

**Stratified semantics–layout interaction.** While the TA and SA modules focus on fine-grained semantics, they do not explicitly enforce spatial consistency across different layout regions. To generate a coherent and complete scene, the Brain–Scene Routing Controller further incorporates a **Layout Attention** (LA) module and a **Stratified Fuser** (SF) module. The LA module employs region-constrained self-attention to restrict information exchange across different spatial regions. Specifically, we denote the spatial mask set as $\mathbf{M} = [\mathcal{M}_{\texttt{srd}}, \mathcal{M}_{\texttt{tgt}}]$ and construct an additive attention mask $\mathcal{A}_{\texttt{LA}}$ according to region membership:

$$\mathcal{A}_{\texttt{LA}}(i,j) = \begin{cases} 0, & \exists\, \mathcal{M} \in \mathbf{M},\ \mathcal{M}(i) = \mathcal{M}(j) = 1, \\ -\infty, & \text{otherwise.} \end{cases} \tag{9}$$

The Layout Attention is then formulated as

$$\mathrm{LA}(\mathbf{f}) = \mathrm{Softmax}\left(\frac{Q(\mathbf{f})K(\mathbf{f})^\top}{\sqrt{d}} + \mathcal{A}_{\mathrm{LA}}\right) V(\mathbf{f}), \tag{10}$$

where the additive mask assigns $-\infty$ to cross-region attention logits before normalization, ensuring that each spatial token attends only to tokens within the same region. This region-constrained interaction reduces attribute leakage across different areas while preserving spatial consistency. The SF module then aggregates the outputs from the specialized attention modules through cross-level attention, producing a unified global feature representation. The entire controller is trained end-to-end with the latent diffusion objective (Rombach et al., 2022):

$$\min_{\theta'} \mathcal{L}_{\mathrm{LDM}} = \mathbb{E}_{(\mathbf{z},\mathbf{c},\mathbf{M}),\, \epsilon\sim\mathcal{N}(0,\mathbf{I}),\, t}\left[\|\epsilon - f_{\theta,\theta'}(\mathbf{z}_t, t, \mathbf{c}, \mathbf{M})\|_2^2\right], \tag{11}$$

where $(\mathbf{z}, \mathbf{c}, \mathbf{M})$ denotes the latent, semantic condition, and spatial mask set associated with each training sample. Only the controller parameters $\theta'$ are optimized, while the pretrained diffusion model parameters $\theta$ remain frozen.

# C Metric Details

This section details the Brain–to–Image evaluation metrics. Appendix C.1 reviews the traditional metrics, including their evaluation methodologies and inherent limitations. Appendix C.2 provides a detailed description of our proposed Semantic Faithfulness Evaluation Protocol. Finally, the specifics of our user study are outlined in Appendix C.4.

## C.1 Existing Evaluation Methods

Traditional evaluation metrics for brain–to–image synthesis, as summarized in Table 4, are generally categorized into low–level and high–level assessments. Notably, two–way identification accounts for half of the evaluation metrics. It determines whether the ground–truth image embedding is closer to the brain embedding than a random negative. Despite their widespread use, these metrics suffer from inherent limitations that may hinder an accurate assessment of semantic faithfulness.

First, feature–based high–level metrics struggle to capture instance–level semantic errors, misaligning the evaluation with the primary goal of brain decoding. This often manifests as a prototype bias, where reconstructions are rewarded for matching the global appearance or texture of a category, even if they contain significant semantic errors like incorrect object counts or swapped attributes. Second, the two–way identification metrics do not directly measure the similarity between the prediction and the ground truth. This relative judgment means that a model can achieve a high identification accuracy even if its generated image is of poor quality, as long as it is marginally better than sufficiently dissimilar negative samples.

## C.2 Semantic Faithfulness Evaluation Protocol

Brain–Diffuser (Ozcelik & VanRullen, 2023) typically classifies visual features into high–level (e.g., descriptive scene semantics and object categories) and low–level attributes (e.g., color, shape, texture, and spatial layout). However, as stated in Appendix C.1, conventional evaluation metrics struggle with category–level semantic errors. To address this issue, we introduce a Semantic Faithfulness Evaluation Protocol that measures description-, category-, count-, and position–level correctness, serving as a rigorous complement to semantic fidelity.

**Description–level metric.** To assess the preservation of descriptive scene semantics, we convert the visual information from the ground–truth $\mathcal{I}_{gt}$ and the prediction $\mathcal{I}_{pred}$ into textual captions, $\mathcal{T}_{gt}$ and $\mathcal{T}_{pred}$ respectively, using an image–to–text model $\mathcal{G}$. Direct string comparison is inadequate, as it may penalize semantically equivalent but lexically different descriptions (e.g., "a pair of dogs" vs. "two dogs"). We therefore leverage a pretrained encoder, denoted as $\mathcal{E}$, to map the semantic content of each caption into a vector: $v_{gt} = \mathcal{E}(\mathcal{T}_{gt})$ and $v_{pred} = \mathcal{E}(\mathcal{T}_{pred})$. The final semantic alignment $S_{desc}$ is quantified by the cosine similarity between these vectors:

$$S_{desc}(\mathcal{I}_{gt}, \mathcal{I}_{pred}) = \frac{v_{gt} \cdot v_{pred}}{\|v_{gt}\|\|v_{pred}\|} \tag{12}$$

Table 4: **Existing evaluation metrics.** Description for both high–level and low–level metrics.

| Level | Metrics | Two–way identification | Description |
|---|---|---|---|
| Low–level | PixCorr | | Pixel–wise correlation between reconstruction and ground truth. |
| | SSIM | | Structural similarity assessing luminance, contrast, and structure. |
| | AlexNet(2) | ✓ | Feature comparison from AlexNet's second layer. |
| | AlexNet(5) | ✓ | Feature comparison from AlexNet's fifth layer. |
| High–level | Inception | ✓ | Feature comparison from InceptionV3's final pooling layer. |
| | CLIP | ✓ | Semantic similarity via CLIP–Vision's output layer. |
| | EffNet–B | | Distance metric from EfficientNet–B1. |
| | SwAV | | Distance metric from SwAV–ResNet50. |

We employ Emu3 (Wang et al., 2024b) as $\mathcal{G}$. By evaluating semantic similarity through linguistic descriptions rather than holistic image features, this metric explicitly captures the semantic content conveyed by the reconstructed scene.

**Category–level and count–level metrics.** In addition to descriptive scene semantics, a more fine–grained evaluation is needed to assess the faithfulness of individual objects and their quantities. Whereas description–level metrics operate in an open–vocabulary manner, category- and count–level metrics require explicit detection and categorization of discrete instances. Since the NSD dataset is built upon MS–COCO, where each ground–truth image is annotated with one or more of 80 object category labels, we apply a pretrained object detector to the ground–truth and reconstructed images. For each detected instance, the detector outputs a bounding box $B$, a predicted class label $L_p$, and a confidence score $\tau_c$, enabling precise quantitative comparison at the category and count levels.

To obtain a reliable and threshold–agnostic score, we evaluate the detected instances across a range of confidence thresholds $\tau$ from $\tau_{\min}$ to 1.0. For each threshold, we compute precision and recall using a task–specific function $\mathcal{F}$ (category–level or count–level). Precision measures how many predicted instances are correct, reflecting the accuracy of the reconstruction. Recall measures how many ground–truth instances are successfully recovered, thereby penalizing omissions of the reconstruction. Iterating multiple thresholds reduces sensitivity to any single confidence cutoff. We average precision and recall across all thresholds and report their harmonic mean as the final score. The full procedure is summarized in Algorithm 1.

**Position–level metric.** To evaluate the spatial accuracy of reconstructed objects beyond presence or count, we introduce a position–level metric based on bounding–box overlap. After applying the same confidence threshold $\tau_{\min}$ to detections in both ground–truth and reconstructions, we perform per–class greedy matching: for each shared category, we repeatedly pair the ground–truth and reconstructed boxes yielding the highest IoU (Intersection over Union) until no further positive–IoU match exists. The resulting mean IoU directly quantifies localization fidelity, with higher values indicating that correctly recognized objects are placed more accurately in the reconstruction. The procedure is summarized in Algorithm 2.

### C.3 Evaluation Paradox Analysis

Following MindEye2 (Scotti et al., 2024), SDXL–based methods employ a two–stage generation process to enhance the fidelity and realism of the reconstructed images. While the initial outputs from the SDXL model capture the global layout and high–level semantics, they often suffer from localized artifacts and structural distortions. This phenomenon is likely attributed to the higher dimensionality of the target OpenCLIP ViT–bigG/14 latent space.

**Algorithm 1** Category–level / Count–level Scores

**Input:** Ground–truth labels $L_g$ and scores $S_g$
**Input:** Predicted labels $L_p$ and scores $S_p$
**Input:** Minimum confidence threshold $\tau_{\min}$
**Input:** Task–specific function $\mathcal{F}$

**function** COMPUTESCORE($L_g, S_g, L_p, S_p, \tau_{\min}, \mathcal{F}$)
    Initialize recall list $R$ and precision list $P$
    **for** each threshold $\tau$ from $\tau_{\min}$ to 1.0 **do**
        Select ground–truth instances with $S_g \geq \tau$
        Select predicted instances with $S_p \geq \tau$
        Use $\mathcal{F}$ to obtain matched objects and count
        **if** ground–truth set is not empty **then**
            Append recall $r_\tau$ to $R$
        **end if**
        **if** prediction set is not empty **then**
            Append precision $p_\tau$ to $P$
        **end if**
    **end for**
    Compute average recall $\bar{r}$ and precision $\bar{p}$
    Return the harmonic mean as final score
**end function**

**Algorithm 2** Position–level Score

**Input:** Ground–truth boxes $B_g$, labels $L_g$, scores $S_g$
**Input:** Predicted boxes $B_p$, labels $L_p$, scores $S_p$
**Input:** Confidence threshold $\tau_{\min}$

**function** POSSCORE($B_g, L_g, S_g, B_p, L_p, S_p, \tau_{\min}$)
    Preserve $B_g$ and $B_p$ with $S_g \geq \tau_{\min}$, $S_p \geq \tau_{\min}$
    Initialize total IoU to 0
    Identify labels shared by both sets
    **for** each shared label **do**
        Collect $B_g$ and $B_p$ of this label
        Compute the IoU matrix
        **for** each possible match **do**
            Find the largest IoU and its indices
            Add the IoU value to total IoU
            Remove the matched row and column
        **end for**
    **end for**
    Compute position score as total IoU over the number of valid $B_g$
    **return** position score
**end function**

While MindEye2 reports that the refinement stage often leads to a decline in most high–level metrics, our experiments reveal a significant boost in Category–level and Count–level performance. To investigate this discrepancy, we analyze a representative case from MindEye2. The unrefined output presents a distorted, ambiguous animal figure, whereas the refined counterpart clearly resolves into a dog.

Quantitative analysis of this sample highlights the limitation of traditional metrics: while the Category–level and Count–level scores both leap from 0.67 to 1.00 due to the restored structural coherence, standard high–level metrics fail to capture this perceptual improvement. Specifically, both Inception and CLIP scores suffer significant drops (0.98 to 0.95 and 0.99 to 0.93, respectively), and EffNet–B actually degrades (distance increases from 0.79 to 0.96), while SwAV shows only a marginal improvement (0.49 to 0.45). These counter–intuitive fluctuations underscores the lack of interpretability

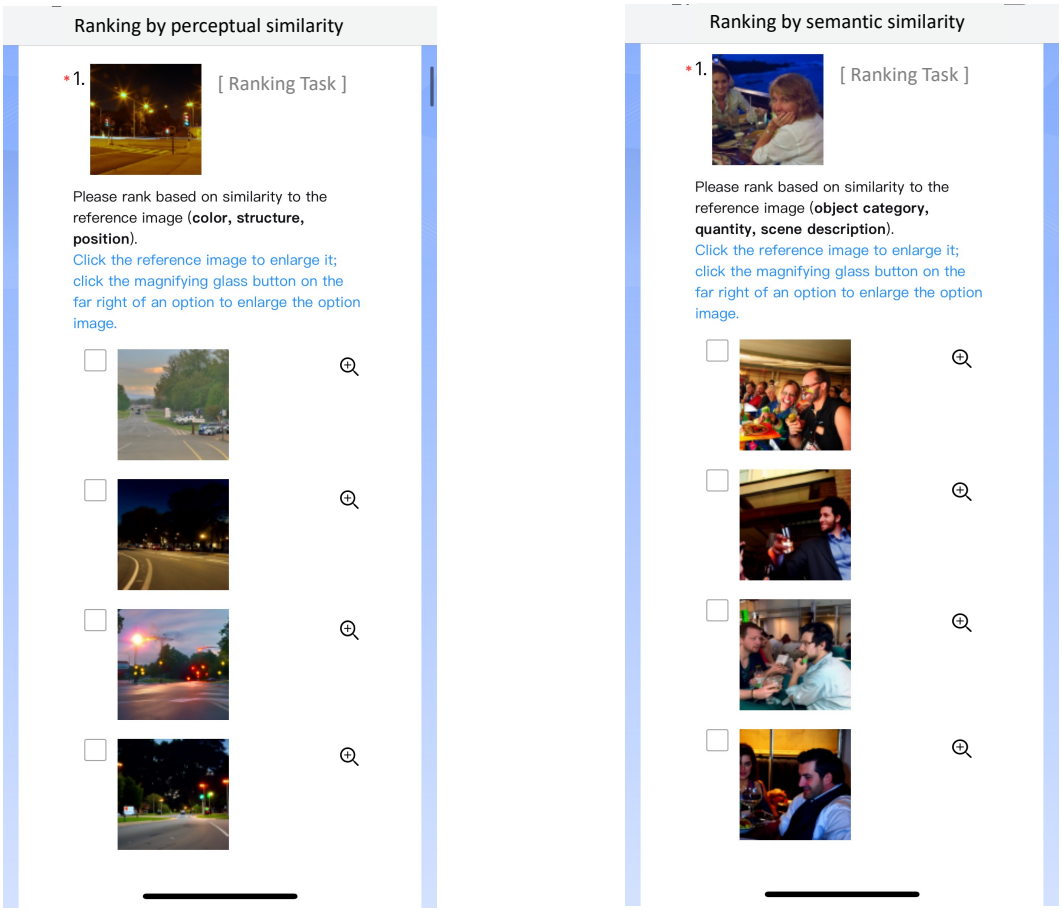



Figure 10: **User study interface.** The left panel evaluates perceptual similarity, while the right panel focuses on semantic similarity.

in traditional high–level metrics, as they fail to align with the visible restoration of semantic content. Consequently, the substantial improvements in object–centric metrics indicate that the refinement process successfully enhances semantic fidelity, allowing object detectors to recognize features with higher confidence, which aligns much closely with human perceptual intuition.

### C.4 User Study

We conducted a user study involving 51 participants to analyze the correlation between metric evaluation and human preferences. We randomly sampled 50 ground–truth images from the NSD dataset and presented their corresponding reconstructions from our model along with those from three baseline models, yielding 50 quadruples. The baselines included MindEye, MindBridge, and NeuroPictor. The model labels were anonymized throughout the human evaluation process.

Participants performed two separate rankings for each image quadruple: one for perceptual similarity and one for semantic similarity. Perceptual similarity focuses on how closely the reconstruction matches the ground–truth in low–level visual attributes such as color, structure, spatial layout, and overall visual appearance. Semantic similarity evaluates how well the reconstruction preserves the main object categories, object count, and overall scene description. Fig. 10 displays our user study interface. To quantify the alignment between metrics and human preferences, we compute a consistency score. For each of the image quadruples, we consider all possible pairs of reconstructions. The consistency score is defined as the fraction of human ranking pairs that agree with the metric–based ranking under the corresponding criterion. We average these scores across all quadruples to obtain the final result.

# D EXPERIMENT DETAILS

## D.1 IMPLEMENTATION DETAILS

All models are trained independently. For both the Layout and Semantic Decoding Branches, we use AdamW (Loshchilov & Hutter, 2017) with a OneCycle learning rate schedule, 5e-4 for the Layout Decoding Branch, and 3e-4 / 1e-4 for the Semantic Decoding Branch (ViT–L/14 / OpenCLIP ViT–bigG/14), while the Brain–Scene Routing Controller uses a constant learning rate of 1e-4. (i) **Layout Decoding Branch**: Trained for 120 epochs with batch sizes of 24 (single) or 96 (cross) on 1 or 4 GPUs. (ii) **Semantic Decoding Branch**: Trained for 150 epochs on the same GPU setup. Batch sizes are 24/96 for ViT–L/14 and 16/64 for OpenCLIP ViT–bigG/14. (iii) **Brain–Scene Routing Controller**: Trained for 5 epochs on a single GPU using data from four subjects, with batch sizes of 16 for the VD version and 4 for the SDXL version.

## D.2 HYPERPARAMETER SELECTION

We follow MindEye2 (Scotti et al., 2024) and set $\lambda_{sc}$ to 0.1.

As described in Section 3, we define the semantic reconstruction loss as

$$L_{\text{sem}} = L_{\text{recon}} + \alpha L_{\text{align}} + \beta L_{\text{cons}},$$

where $\alpha$ controls the alignment loss and $\beta$ controls the consistency loss.

The value of $\alpha$ was set empirically, while $\beta$ was selected through a small hyperparameter sweep. Specifically, following (Scotti et al., 2024), we fixed the reconstruction–to–alignment ratio at $30:1$. This choice reflects that reconstruction is the primary objective, while alignment serves as an auxiliary constraint to regularize the learned semantic representation. The ratio corresponds to $\alpha \approx 1/30 = 0.033$, which we report as $\alpha = 0.03$ in our experiments.

We then tuned $\beta$ relative to $\alpha$ by sweeping the ratio $\beta/\alpha$ over $\{0, 0.1, 0.2, 0.3, 0.5, 0.8, 1, 1.5\}$. To reduce computational cost, this sweep was conducted under a reduced–capacity setting, using one brain token and one–third of the Subject-1 training data.

As shown in Table 5, setting $\beta/\alpha = 0.2$ achieves the best overall trade–off across the evaluated metrics. This setting obtains the best results on PixCorr, AlexNet, InceptionV3, CLIP, EffNet–B, and SwAV, while remaining competitive on SSIM. When $\beta$ becomes too large, the consistency term may over–regularize the representation, suppressing subject–specific semantic details and degrading reconstruction quality. Therefore, we set $\beta/\alpha = 0.2$ in the main experiments, which gives $\beta \approx 0.007$.

Table 5: Experimental results of various evaluation metrics under different $\beta/\alpha$ ratios.

| weights | 0 | 0.1 | 0.2 | 0.3 | 0.5 | 0.8 | 1 | 1.5 |
|---|---|---|---|---|---|---|---|---|
| PixCorr ↑ | 0.2710 | 0.2758 | **0.2789** | 0.2724 | 0.2735 | 0.2747 | 0.2744 | 0.2723 |
| SSIM ↑ | 0.3315 | 0.3312 | 0.3307 | 0.3275 | 0.3289 | 0.3266 | **0.3325** | 0.3278 |
| AlexNet (2) ↑ | 93.38% | 93.67% | **93.76%** | 93.59% | 93.43% | 93.33% | 93.21% | 93.24% |
| AlexNet (5) ↑ | 97.17% | 97.26% | **97.38%** | 97.25% | 97.13% | 96.71% | 96.76% | 96.77% |
| InceptionV3 ↑ | 90.16% | 90.82% | **91.48%** | 91.05% | 90.36% | 89.78% | 89.47% | 88.58% |
| CLIP ↑ | 90.19% | 90.57% | **90.85%** | 90.68% | 90.24% | 89.67% | 89.27% | 88.51% |
| EffNet–B ↓ | 0.6637 | 0.6604 | **0.6587** | 0.6624 | 0.6678 | 0.6732 | 0.6796 | 0.6827 |
| SwAV ↓ | 0.3884 | 0.3875 | **0.3830** | 0.3878 | 0.3969 | 0.4007 | 0.4045 | 0.4093 |

## D.3 LAYOUT DECODING

**Motivation.** As discussed in Appendix. A.1, the visual cortex is functionally organized into streams, and the lateral and parietal streams are collectively known as the dorsal stream or the "where" pathway, which is specialized for spatial awareness. Motivated by this biological organization, we move beyond standard semantic decoding and conduct a quantitative analysis of spatial localization capability. Our goal is to examine how different brain regions, spanning early visual areas and higher–level

Table 6: **Performance on layout decoding.** Multi–subject methods are highlighted. N–CD indicates Normalized Centroid Distance; N–ASD indicates Normalized Average Surface Distance. Red indicates the best performance.

| Configuration | Mask Map | Voxel Num (Avg.) | Target Region | | | Surrounding Region | | |
|---|---|---|---|---|---|---|---|---|
| | | | mIoU ↑ | N–CD ↓ | N–ASD ↓ | mIoU ↑ | N–CD ↓ | N–ASD ↓ |
| Random | Segmentation | 13013 | 6.98% | 16.20% | 13.43% | 44.09% | 11.50% | 8.88% |
| Dorsal | Segmentation | 13013 | 35.53% | 11.24% | 6.75% | 54.61% | 8.74% | 6.52% |
| Dorsal | Segmentation | 13013 | 31.94% | 12.26 % | 7.23% | 53.43% | 9.07% | 6.78% |
| Nsdgeneral | Segmentation | 13931 | 37.50% | 10.88% | 6.81% | 57.01% | 8.19% | 6.09% |
| Nsdgeneral | Segmentation | 13931 | 38.20% | 10.80% | 6.59% | 56.58% | 8.46% | 6.23% |
| Dorsal | Bounding Box | 13013 | 40.62% | 11.03% | 6.97% | 68.23% | 7.11% | 5.52% |
| Nsdgeneral | Bounding Box | 13931 | 35.48% | 11.25% | 7.60% | 70.24% | 6.57% | 5.14% |

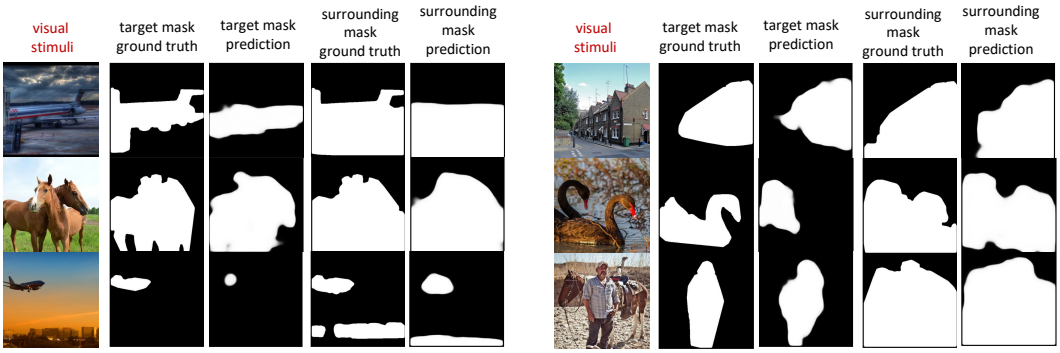



Figure 11: **Qualitative visualization of layout decoding.** We display the reconstruction results for the segmentation masks.

functional streams, contribute to the encoding of positional information, and to determine whether the dorsal pathway indeed exhibits superior performance in spatial reconstruction tasks compared to semantic–focused regions.

**Experimental Setup.** To isolate spatial processing capabilities from semantic understanding, we designed a series of controlled decoding experiments. We varied the input brain regions and the output representations to evaluate the localization performance of the model. We reconstruct both target and surrounding regions to mitigate subjective bias in target annotations. Furthermore, detailed segmentation contours can inadvertently leak semantic identity. To address this, we introduce a Bounding Box setting where masks are replaced by rectangles. This eliminates shape priors and forces the model to rely exclusively on spatial localization.

**ROI Configurations.** We analyze 3 distinct ROI configurations to dissect spatial encoding capabilities. (1) *Random Baseline* uses random values to establish a lower bound for performance. (2) `Dorsal` aggregates lateral and parietal dorsal regions across multiple functional levels. (3) `Nsdgeneral` encompasses the general visual cortex and serves as a comprehensive reference.

**Evaluation Metrics.** We employ three metrics to quantify reconstruction quality. Mean Intersection over Union measures the overlap accuracy between the predicted and ground truth masks. Normalized Centroid Distance assesses the displacement of the predicted object center. Normalized Average Surface Distance evaluates the precision of the predicted boundaries. We normalize both distance metrics by the image diagonal to ensure scale invariance.

**Result Analysis.** Table 6 presents the quantitative results, while Fig. 11 provides a qualitative comparison of the reconstructed masks. We summarize our 4 main findings as follows:

- Positional information can be effectively decoded from fMRI signals recorded during natural scene viewing. Across both target and surrounding regions, the trained ROI configurations

consistently outperform the random baseline across all metrics, confirming that cortical responses contain decodable spatial information.

- `Nsdgeneral` performs better than `Dorsal` in the segmentation setting. This suggests that when explicit object contours are preserved, broader visual representations can provide additional cues for spatial reconstruction beyond those captured by the dorsal pathway alone.
- Removing shape bias reveals the positional specialization of the dorsal pathway. On target–region bounding–box masks, `Dorsal` outperforms `Nsdgeneral`, reversing the trend observed in the segmentation setting. This suggests that the advantage of `Nsdgeneral` in segmentation partly benefits from shape–related cues, whereas `Dorsal` more strongly preserves position–specific information.
- Multi–subject results further reveal different generalization behaviors across ROI configurations. `Nsdgeneral` benefits from multi–subject training, whereas `Dorsal` shows a performance decline, suggesting that broader visual representations are more transferable across subjects than spatial features.

### D.4 Semantic Decoding

To provide a more intuitive assessment of reconstruction quality, we present extensive qualitative comparisons between MindStrata and 4 baselines: MindEye2, Neuropictor, MindEye, and MindBridge. All baseline results are reproduced under our experimental settings. Specifically, Fig. 12 displays reconstructions using the Versatile Diffusion (VD), while Fig. 13 visualizes results using the SDXL.




Figure 12: **Qualitative comparison of reconstructions using Versatile Diffusion.** We compare MindStrata (w/ VD) against four reproduced baselines.




Figure 13: **Qualitative comparison of reconstructions using SDXL**. We compare MindStrata (w/ SDXL) against four reproduced baselines.

Table 7: **Quantitative evaluation of reconstruction performance with limited fMRI data.** Red indicates the best performance.

| Method | Low–Level | | | | High–Level | | | | SFEP (Ours) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PixCorr ↑ | SSIM ↑ | Alex(2) ↑ | Alex(5) ↑ | Incep ↑ | CLIP ↑ | Eff ↓ | SwAV ↓ | Description ↑ | Category ↑ | Count ↑ | Position ↑ |
| MindBridge | 0.127 | 0.233 | 80.2% | 87.5% | 81.7% | 86.4% | 0.836 | 0.525 | 0.295 | 0.201 | 0.206 | 0.083 |
| MindEye2 | 0.201 | 0.406 | 85.0% | 90.8% | 81.5% | 80.3% | 0.806 | 0.465 | 0.340 | 0.255 | 0.226 | 0.110 |
| MindAligner | 0.206 | 0.414 | 85.6% | 91.6% | 83.0% | 81.2% | 0.802 | 0.463 | - | - | - | - |
| MindStrata (w/ VD) | 0.215 | 0.334 | 89.7% | 93.4% | 84.6% | 87.5% | 0.773 | 0.436 | 0.361 | 0.272 | 0.240 | 0.122 |
| MindStrata (w/ SDXL) | 0.203 | 0.387 | 90.1% | 93.7% | 84.9% | 87.7% | 0.784 | 0.449 | 0.363 | 0.275 | 0.244 | 0.135 |

### D.5 Data–limited Decoding

The acquisition of fMRI data is inherently costly and time–intensive, often constraining real–world decoding applications to scenarios with limited data. To address this challenge, we investigate a data–limited multi–subject decoding setting. Prior approaches typically adopt a subject–transfer paradigm, where a model is first pre–trained on data–rich source subjects to establish strong priors before being fine–tuned on a data–poor target subject. Departing from this paradigm, we take a step further by training a unified multi–subject model entirely from scratch using limited data from multiple subjects. As discussed in Appendix A, while the NSD dataset contains four subjects with complete 40-session scans ($\sim$ 40h), we intentionally utilize only 1 hour of data to train MindStrata, thereby vigorously testing its generalization performance under severe data constraints.

We compared MindStrata against three SOTA multi–subject decoding models in the limited–data setting, as detailed in Tab. 7. Even with restricted training data, MindStrata (SDXL) demonstrates generalized semantic decoding capabilities, achieving the highest scores across high–level metrics and all human–aligned SFEP metrics. Notably, in terms of spatial organization, our method achieves a Position–level score of 0.132, significantly surpassing the baselines. MindStrata (SDXL) outperforms MindEye2, by substantial margins in semantic fidelity and layout accuracy.

### D.6 Impact of Spatial Resolution

To assess the robustness of MindStrata under reduced spatial resolution, we simulate progressively degraded spatial resolution by applying Gaussian smoothing to the fMRI voxels. Fig. 14 visualizes the signals under varying degrees of smoothing, illustrating the progressive loss of fine-grained spatial details.

**Experimental Setup.** We conducted extensive experiments to evaluate decoding performance as spatial resolution decreases, employing the Semantic Faithfulness Evaluation Protocol (Fig. 15), as well as standard high–level (Fig. 16) and low–level (Fig. 17). We employ normalized performance to evaluate the overall variations, while also presenting the specific changes for each individual metric. MindStrata utilizes Versatile Diffusion (Xu et al., 2023) for inference. Consistent with our simulation protocol, all models were trained and evaluated on the smoothed fMRI data, with reported results representing the average across 4 subjects.

Specifically, to investigate whether spatial smoothing mitigates the heterogeneity of semantic responses across individuals, we trained MindEye2 on the smoothed data from all 4 subjects jointly before evaluating each subject. For computational efficiency, the shared–subject latent space in MindEye2 is set to 1024 dimensions.

**Result Analysis.** While decoding performance naturally declines across all models as resolution degrades, MindStrata demonstrates superior robustness, preserving more recognizable visual structures than baselines as qualitatively shown in Fig. 18. Specifically, it consistently outperforms the baselines on SFEP and high–level metrics, with a minor exception at the most extreme smoothing level ($\sigma$=5). In terms of low–level metrics, MindStrata achieves performance comparable to MindEye2 while significantly surpassing MindEye.

Furthermore, regarding the multi–subject training strategy, our results suggest that while spatial smoothing may facilitate multi–subject alignment by reducing individual variability, the information loss incurred by reduced spatial fidelity dominates, ultimately outweighing any benefits from shared latent modeling.

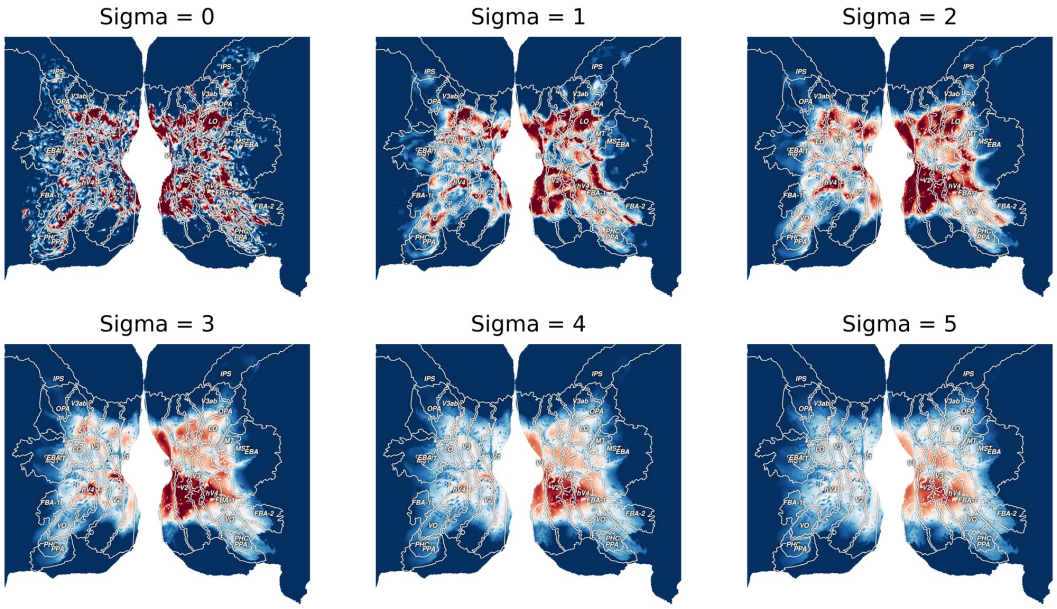



Figure 14: **Simulating reduced spatial resolution.** We visualize fMRI signals under varying degrees of Gaussian smoothing ($\sigma = 1, 2, 3, 4, 5$ voxels) to mimic degraded spatial fidelity.

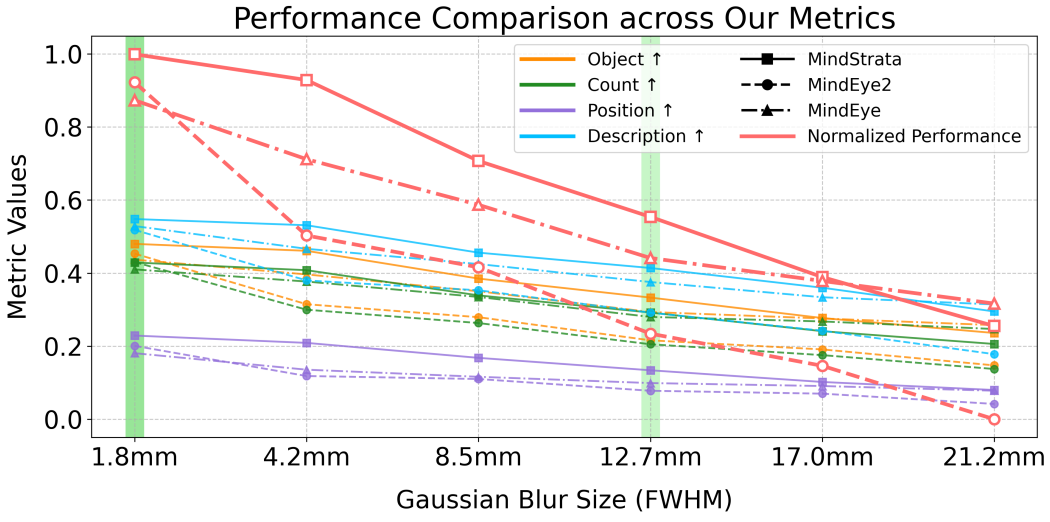



Figure 15: **SFEP decoding performance at lower spatial resolutions.** fMRI data were smoothed with Gaussian kernels ($\sigma = 1, 2, 3, 4, 5$ voxels), corresponding to full–width–at–half–maximum (FWHM) of 4.2, 8.5, 12.7, 17.0, and 21.2 mm. The dark green line marks the current fMRI spatial resolution, and the light green line marks the estimated resolution of current portable systems (White & Culver, 2010).

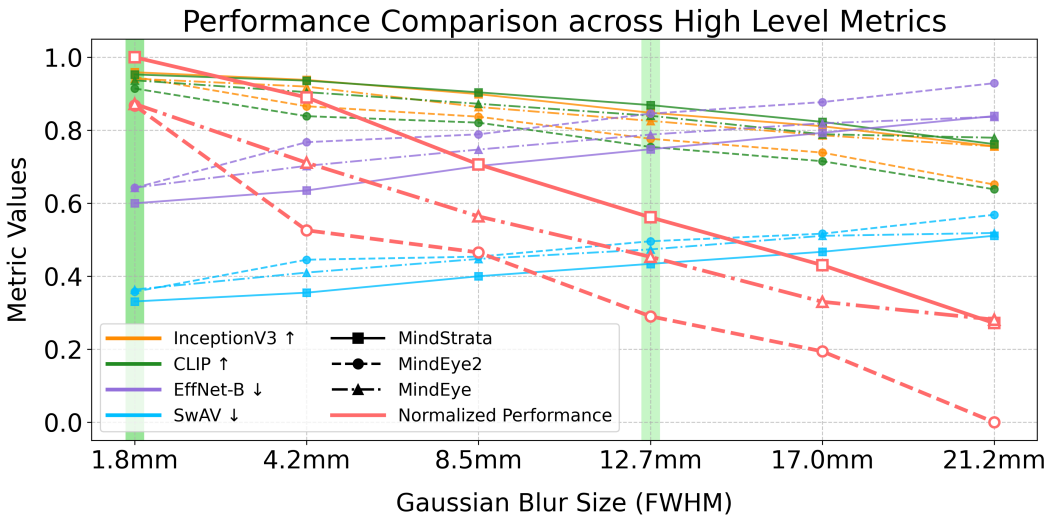



Figure 16: **High–level decoding performance at lower spatial resolutions.** The experimental setup, smoothing parameters, and resolution markers follow Fig. 15.

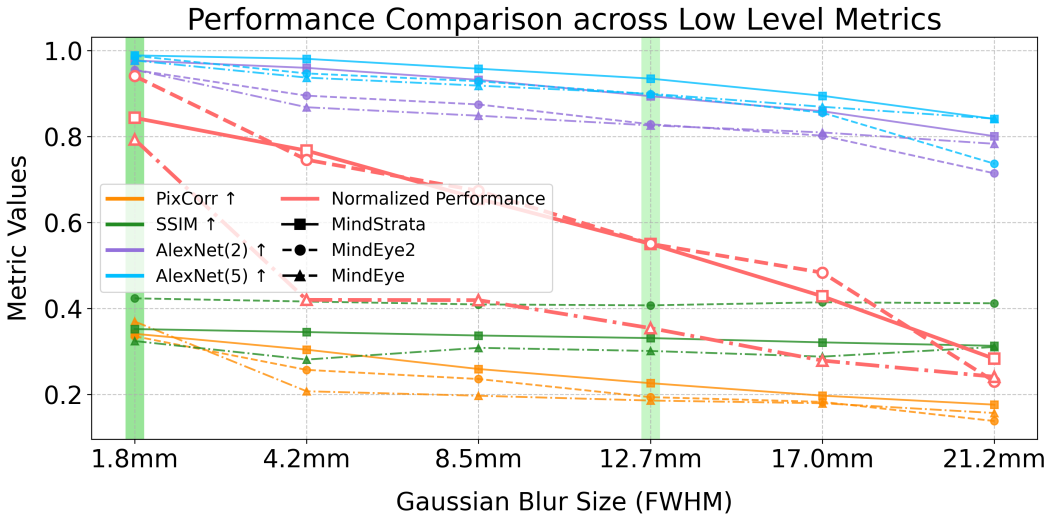



Figure 17: **Low–level decoding performance at lower spatial resolutions.** The experimental setup, smoothing parameters, and resolution markers follow Fig. 15.

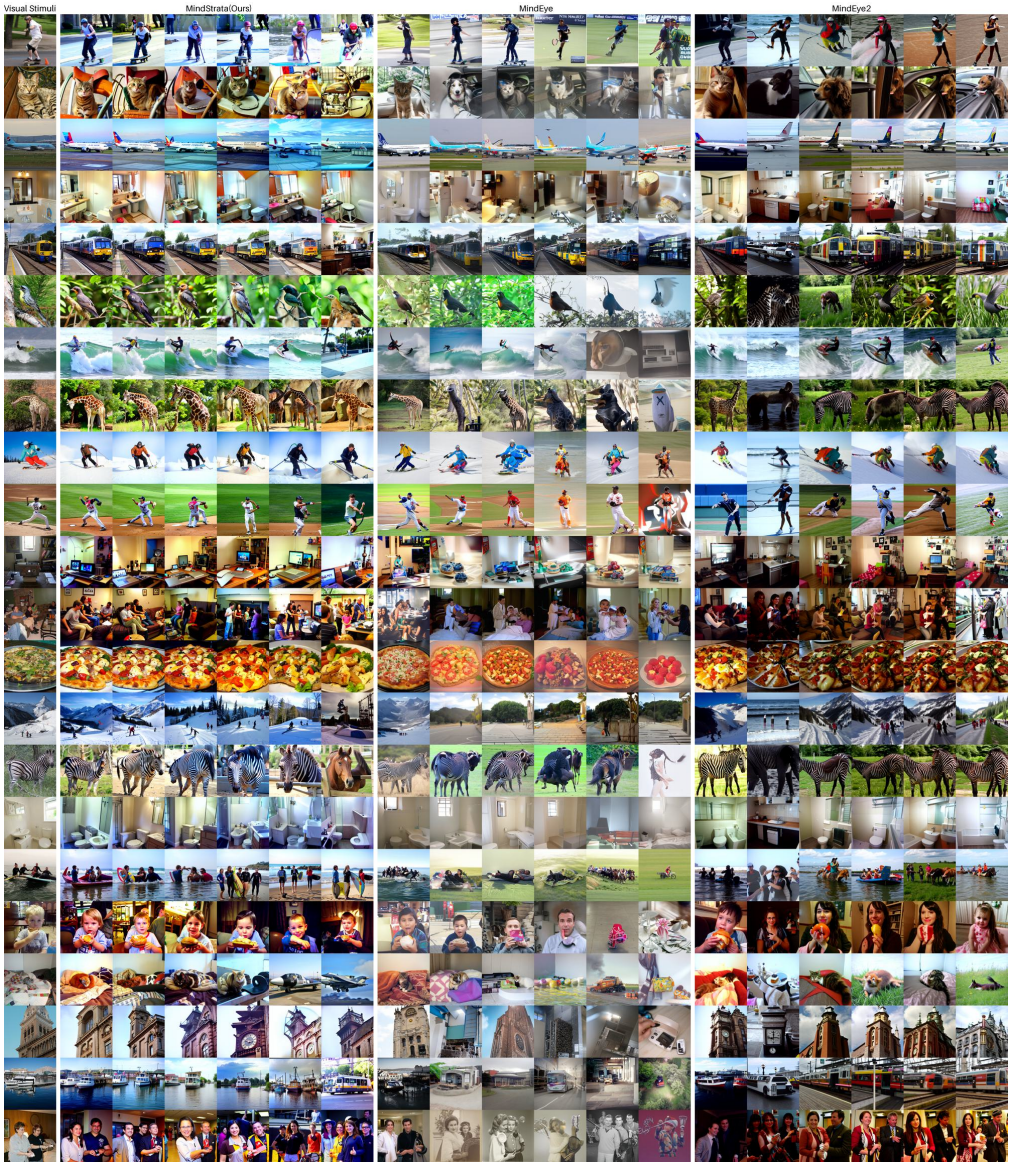



Figure 18: **Qualitative comparison of reconstructions under degrading spatial fidelity.** The leftmost column displays the ground truth visual stimuli. The remaining columns are grouped into three main blocks corresponding to MindStrata, MindEye, and MindEye2, arranged from left to right. Within each model's block, the six sub–columns represent increasing levels of Gaussian smoothing applied to the fMRI data, with kernel sizes $\sigma \in \{0, 1, 2, 3, 4, 5\}$ voxels, where $\sigma = 0$ indicates the original spatial resolution. We recommend zooming in for better visualization.

Table 8: **Ablations on the brain–semantic tokens.** Red indicates the best performance.

| $[N_{e^{(g)}_{x_s}}, N_{e^{(i)}_{x_s}}, N_{e^{(l)}_{x_s}}]$ | Params (M) | Low–Level | | | | High–Level | | | | SFEP (Ours) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | PixCorr ↑ | SSIM ↑ | Alex(2) ↑ | Alex(5) ↑ | Incep ↑ | CLIP ↑ | EffNet–B ↓ | SwAV ↓ | Description ↑ | Category ↑ | Count ↑ | Position ↑ |
| [1, 1, 1] | 620 | 0.333 | 0.347 | 96.6% | 98.3% | 94.6% | 94.3% | 0.628 | 0.351 | 0.517 | 0.445 | 0.393 | 0.221 |
| [4, 4, 4] | 648 | 0.338 | 0.350 | 97.3% | 98.6% | 95.4% | 94.8% | 0.615 | 0.343 | 0.531 | 0.472 | 0.414 | 0.228 |
| [16, 16, 16] | 762 | 0.341 | 0.353 | 97.6% | 98.8% | 95.3% | 94.9% | 0.609 | 0.335 | 0.538 | 0.484 | 0.431 | 0.236 |
| [64, 32, 16] (Ours) | 963 | 0.341 | 0.352 | 97.7% | 98.9% | 95.9% | 95.3% | 0.600 | 0.331 | 0.542 | 0.489 | 0.435 | 0.236 |
| [64, 64, 64] | 1,214 | 0.342 | 0.353 | 97.8% | 99.0% | 96.1% | 95.8% | 0.598 | 0.329 | 0.540 | 0.491 | 0.438 | 0.239 |

### D.7 Ablation Study

Beyond ablating MindStrata's core modules, we investigate the trade–off between performance and model size by varying the number of brain–semantic token triplets. As shown in Tab. 8, performance improves across all metrics as the number of tokens increases. Notably, a model with only one token triplet retains close performance while utilizing $\sim 50\%$ fewer parameters than a 64-triplet. This finding supports the efficacy of sparse representations for fMRI, consistent with prior work (Mayo et al., 2024; Chen et al., 2023; Sun et al., 2023). To balance performance and efficiency, we ultimately adopt a configuration with progressively fewer tokens per stream, which is comparable in parameter count to MindEye. Conversely, a baseline without brain–semantic tokens that directly projects fMRI latents into the CLIP space is computationally impractical, increasing the parameter count to 3.76G ($> 3\times$ our model) and requiring infeasible GPU memory.

### D.8 Stratified Semantic Decoding

Mirroring the stratified nature of human visual perception (Felleman & Van Essen, 1991; Riesenhuber & Poggio, 1999), our model reconstructs not just the global scene but also the target object of a subject's perception and its corresponding contextual surroundings, as showcased in Fig. 19, as the local and regional streams culminate in the global stream's output, which provides a faithful and holistic reconstruction of the visual scope.

The local stream isolates the primary semantic object (e.g., a person, a bird, or a laptop), demonstrating MindStrata's ability to pinpoint the focus of perception. Complementing this, the intermediate stream captures the broader environmental context while specifically maintaining the target object. These two components culminate in the global stream's output, which provides a faithful and holistic reconstruction of the entire visual scene.

### D.9 Fine–Grained Brain Concept Localization

Localizing a single visual concept from brain activity is challenging, especially when the stimuli are complex natural scenes that contain multiple semantic elements. While MindSimulator (Bao et al., 2025a) introduces a generative fMRI encoding approach to synthesize brain activity for concept–oriented stimuli, the reconstructed fMRI signals for complex scenes often remain entangled with neural patterns from co–occurring concepts. This can be revealed by decoding the synthesized signals via decoding models. This raises an important question: *How can we generate concept–exclusive brain responses from complex scenes to achieve fine–grained localization of specific visual concepts?*

A straightforward solution is to encode a target image containing only one object. Leveraging this idea, our stratified model successfully decodes single–concept representations by reconstructing the target concept encoded from the synthesized fMRI. As illustrated in Fig. 20, this approach effectively disentangles the object's neural signature from its environmental context. While the global stream fails by incorrectly imposing background priors learned from natural scenes, the regional and local streams successfully filter out this contextual interference (grassland) to reconstruct the single semantic concept (zebra). This finding confirms that a brain encoder, despite being trained on complex natural scenes, can effectively process segmented, single–concept images, thereby paving the way for fine–grained localization of conceptual representations in the brain.




Figure 19: **Stratified decoding.** MindStrata's three stratified streams successfully decode target object, contextual surroundings and global scene.

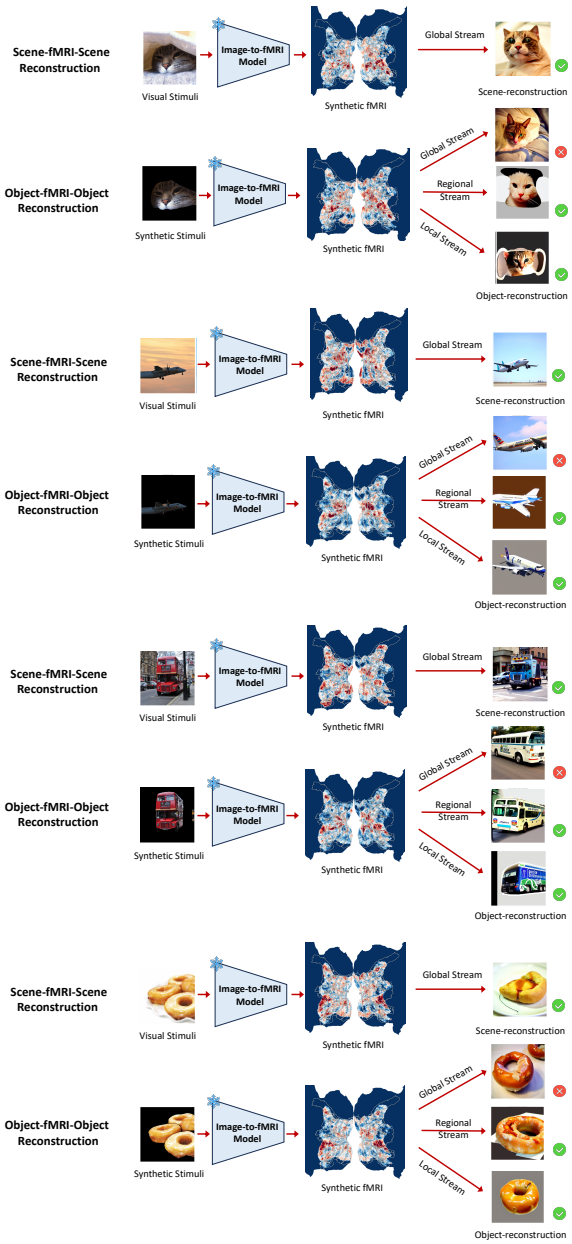



Figure 20: **Additional examples of fine–grained brain localization.** We illustrate the Scene–fMRI–Scene and Object–fMRI–Object reconstruction across diverse samples. The results reinforce that regional and local streams are crucial for correcting the contextual bias compared with global stream in object–centric tasks. We recommend zooming in for better visualization.

# E Related Work

## E.1 Controllable Image Diffusion

With the rapid progress of diffusion models (Song et al., 2020; Ho et al., 2020; Dhariwal & Nichol, 2021; Ho & Salimans, 2022; Rombach et al., 2022; Croitoru et al., 2023), they have become the SOTA paradigm for high–resolution image synthesis. However, the common text–only guidance provides weak control over the generative process (Rombach et al., 2022; Balaji et al., 2022; Ramesh et al., 2022; Nichol et al., 2021; Saharia et al., 2022). The problem becomes more evident in multi–instance scenes, frequently leading to instance omission, attribute confusion, and chaotic layouts. To address these limitations, a significant body of work (Zhang et al., 2023; Ye et al., 2023; Mou et al., 2024; Yang et al., 2023; Bar-Tal et al., 2023; Xie et al., 2023; Zhou et al., 2024b;a;c) has focused on adding explicit constraints. ControlNet (Zhang et al., 2023) introduces strong spatial conditioning via inputs like canny edges, depth maps, or human poses. The more challenging task of controlling multi–instance generation has been approached by fine–tuning the diffusion model on position–specific data or by incorporating guidance from bounding boxes to enforce layout constraints (Yang et al., 2023; Zhou et al., 2024b;a). Distinct from the traditional text–to–image paradigm, Versatile Diffusion (Xu et al., 2023) introduces a multi–modal framework, enabling both image and text features to jointly guide the synthesis process. However, it still struggles to preserve consistent spatial structure and multi–instance coherence.

## E.2 Multi–Subject Brain Alignment

Multi–subject brain decoding faces substantial challenges due to individual variability in brain anatomy and neural responses (Mazziotta et al., 2001), often leading to negative transfer when training across subjects (Quan et al., 2024). Most existing approaches rely on anatomical or functional alignment. Anatomical alignment methods (Bao et al., 2025b; Shen et al., 2024) map individual brain volumes onto a standard template, which is computationally expensive and may distort functional patterns. By contrast, functional alignment projects fMRI signals from different subjects into a shared latent space. Explicit functional alignment methods (Thual et al., 2022; Ferrante et al., 2024; Dai et al., 2025) learn pairwise mappings between subjects but fail to exploit shared inter–subject structures. Implicit alignment methods (Défossez et al., 2023; Tang et al., 2023; Wang et al., 2024a; Scotti et al., 2024; Gong et al., 2025) train a shared encoder across subjects, yet neglect the relational structure among latent representations, leading to suboptimal or compromised alignment. More recently, meta–learning approaches have enabled in–context adaptation to novel subjects, ranging from few–shot cortical response modeling (Yu et al., 2026) to training–free cross–subject semantic decoding (Nan et al., 2026); however, these methods do not address visual reconstruction. To address the challenge, we propose a Consistency–based functional alignment strategy that preserves intrinsic multi–subject and multi–modal representational structures within a unified latent space.

## E.3 Fine–Grained Cortical Interrogation

Complementary to visual reconstruction, prior work has used encoding models to probe fine–grained cortical organization under naturalistic stimuli (Huth et al., 2012). More recent methods leverage generative or vision–language models: BrainDiVE (Luo et al., 2023) synthesizes preferred stimuli, BrainSCUBA (Luo et al., 2024a) describes voxel–level semantic selectivity, BrainSAIL (Luo et al., 2024b) localizes image regions associated with cortical responses, and MindSimulator (Bao et al., 2025a) localizes concept–selective regions using synthetic fMRI. These approaches primarily interrogate which visual concepts or regions drive cortical responses, rather than reconstructing perceived scenes from measured fMRI. In contrast, MindStrata decodes fine–grained semantic and spatial representations from fMRI for faithful scene reconstruction.