HI-DREAM: HIERARCHICAL DIFFUSION WITH ROI-AWARE REPRESENTATIONS FOR FMRI-TO-IMAGE RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstructing natural images from fMRI requires bridging neural activity with both the structural and semantic representations used by modern generative models. Existing diffusion-based decoders often condition on a single global fMRI embedding, which limits their ability to exploit the hierarchical organization of the visual cortex and makes the contribution of different visual areas difficult to inspect. We propose **HI-DREAM**, a brain-inspired hierarchical diffusion framework that structures fMRI conditioning according to early, middle, and late visual Regions of Interest (ROI) streams. A ROI adapter converts these streams into a multi-scale cortical pyramid, and a lightweight ROI-conditioned ControlNet injects the resulting anatomy-aware priors into matched U-Net depths during denoising. Rather than learning a single undifferentiated neural representation, HI-DREAM preserves subject-specific cortical structure throughout the generator-side conditioning pathway. Experiments on the Natural Scenes Dataset (NSD) show that HI-DREAM achieves state-of-the-art high-level semantic reconstruction while retaining strong low-level structure. Further ablation and attribution analyses show that the proposed hierarchy-aware conditioning is effective, and that different ROI streams provide complementary, inspectable contributions to reconstruction.

1 INTRODUCTION

Reconstructing natural images from human brain activity is a central problem in visual decoding, with potential implications for understanding perception, cognition, and brain–AI alignment. Functional Magnetic Resonance Imaging (fMRI) (Brown et al., 2014; McRobbie et al., 2007) provides non-invasive access to distributed cortical responses (Ashburner & Friston, 2000; Fischl, 2012), and recent diffusion-based generative models have substantially improved the visual quality of fMRI-to-image reconstruction (Takagi & Nishimoto, 2023; Quan et al., 2024; Lin et al., 2022). However, despite this progress, a key challenge remains: most existing methods compress fMRI signals into a single global embedding before conditioning a generative model (Scotti et al., 2024; Wang et al., 2024; Chen et al., 2023b;a). While such embeddings are effective for semantic guidance, they largely ignore the hierarchical organization of the visual cortex and make it difficult to inspect how different cortical regions contribute to different aspects of the reconstruction (Nishimoto et al., 2011; Naselaris et al., 2009; Shen et al., 2019).

Human visual processing is not monolithic (Haxby et al., 2001). Early visual areas are closely associated with local and retinotopic structural information (Joo et al., 2024; Jiang et al., 2021), mid-level areas integrate contours and object parts, and late ventral regions are more related to category-level semantics (See Fig. 1). This hierarchy suggests that different fMRI signals may be most useful at different stages of image generation (Luo et al., 2024; Zhu et al., 2024). In latent diffusion models, the U-Net backbone also contains a natural hierarchy of feature resolutions and denoising depths, where shallow layers tend to preserve spatial structure and deeper layers increasingly support semantic refinement. These observations motivate a more structured form of fMRI conditioning: instead of using one undifferentiated neural representation, the decoder should organize cortical signals into interpretable streams and inject them into matched generative depths.

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.

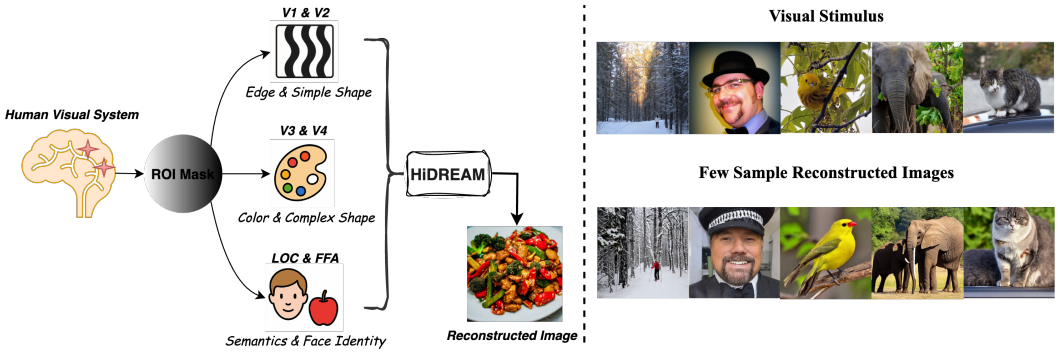

Figure 1: **Brain-inspired decoding pipeline.** HI-DREAM mirrors the visual hierarchy by grouping fMRI signals into early/mid/late ROIs (V1/V2, V3/hV4, and LOC/FFA) and converting them into a multi-scale cortical pyramid. Right: example stimuli and reconstructions, showing preserved structure and semantics with compact, interpretable ROI conditioning.

We propose **HI-DREAM**, a brain-inspired hierarchical diffusion framework for fMRI-to-image reconstruction. HI-DREAM groups fMRI responses into early, middle, and late visual Regions of Interest (ROI) streams and uses a ROI adapter to convert them into a multi-scale cortical pyramid. A lightweight ROI-conditioned ControlNet (Zhang et al., 2023) then injects these anatomy-aware priors (Gong et al., 2025; Huo et al., 2024; Qu et al., 2024) into corresponding U-Net depths during denoising within a latent diffusion backbone (Rombach et al., 2022; Ho et al., 2020). This design allows low-, mid-, and high-level cortical signals to guide the generative process where they are most relevant, while avoiding the computational overhead of conditioning on full 3D fMRI volumes. The resulting framework is compact, depth-aware, and interpretable: each conditioning stream has a clear cortical correspondence, and its influence can be inspected across reconstruction stages.

We evaluate HI-DREAM on the Natural Scenes Dataset (NSD) (Allen et al., 2022), reporting results on subjects 1, 2, 5, and 7, who completed the full scanning schedule. Quantitatively, HI-DREAM achieves state-of-the-art (SoTA) performance on high-level semantic metrics, including Inception and CLIP, while retaining strong low-level structural fidelity. Qualitatively, the reconstructions preserve object identity, scene layout, and salient visual attributes more consistently than prior diffusion-based decoders. Beyond reconstruction performance, our ablation and attribution analyses show that the proposed hierarchy-aware conditioning is effective and interpretable: early, middle, and late ROI streams provide complementary cues; random or reversed hierarchy-depth assignments degrade performance; and the learned depth-wise contributions align with the intended cortical-to-generative hierarchy. Our **contributions** are summarized as follows:

- We introduce **HI-DREAM**, a subject-specific fMRI-to-image reconstruction framework that organizes cortical responses into interpretable Early, Middle, and Late ROI streams and uses them as structured spatial conditions for diffusion generation.
- We propose a lightweight **ROI adapter** and **depth-matched ROI-ControlNet** that transform ROI activity into a multi-scale cortical pyramid and inject anatomy-aware priors into corresponding U-Net depths through residual-FiLM modulation.
- We show that this structured generator-side conditioning improves reconstruction quality on NSD and provides an inspectable pathway for analyzing how different ROI streams contribute across denoising depths, supported by ROI-stream, hierarchy-control, and attribution analyses.

2 RELATED WORK

2.1 FMRI-BASED IMAGE RECONSTRUCTION.

A growing line of work decodes natural images from fMRI by coupling neural representations with powerful image generators (Takagi & Nishimoto, 2023; Aggarwal et al., 2024; Li et al., 2024). Early

diffusion-era methods such as Mind-Vis (Chen et al., 2023a) map fMRI features into latent diffusion models, while MindEye2 (Scotti et al., 2024), Psychometry (Quan et al., 2024), DREAM (Xia et al., 2024), and MindBridge (Wang et al., 2024) improve semantic alignment, perceptual quality, or cross-subject transfer through stronger encoders and generator-side conditioning. Recent diffusion-based decoders further explore structured guidance: MindDiffuser (Lu et al., 2023) uses decoded semantic and structural features, Animate Your Thoughts (Lu et al., 2025) extends reconstruction to video by disentangling semantic, structural, and motion factors, and NeuralDiffuser (Li et al., 2025) introduces primary-visual gradient guidance for fine-detail fidelity. Most of these methods condition the generator using a compact global fMRI embedding or decoded visual features.

2.2 Conditional diffusion and control.

Modern diffusion models (Rombach et al., 2022; Vaswani et al., 2017; Li et al., 2021) admit a spectrum of conditioning mechanisms, including cross-attention to token-like sequences, feature injection at selected layers, and auxiliary networks that modulate intermediate activations (Ho et al., 2020; Rombach et al., 2022; Ho & Salimans, 2022). ControlNet (Zhang et al., 2023) exemplifies the latter, adding a lightweight branch trained to translate structured conditions (edges, depth, poses) into spatial guidance for the U-Net with minimal changes to the sampler. For fMRI decoding, prior works mainly condition on global or low-resolution neural embeddings that act analogously to text tokens (Takagi & Nishimoto, 2023; Scotti et al., 2024; Wang et al., 2024; Quan et al., 2024), which offers semantic steering but provides limited spatial specificity. Using structured spatial conditions as guidance remains comparatively under-explored in this domain.

2.3 Neuroanatomical priors and ROI modeling.

Neuroimaging and systems neuroscience establish a hierarchical organization of human visual cortex (Bandettini, 2009; Baillet, 2017; Cohen, 2017): early areas (V1/V2/V3) encode local, retinotopic structure, mid-level regions integrate contours/parts, and late ventral stream areas represent categorical semantics (Grill-Spector & Malach, 2004; DiCarlo et al., 2012; Wandell & Winawer, 2015). Computational decoders that explicitly respect this hierarchy are rare; most pipelines flatten the neural signal into a single embedding (Takagi & Nishimoto, 2023; Scotti et al., 2024), obscuring how different regions contribute. A few works (Kay et al., 2013; Finn et al., 2015; Huth et al., 2016) inject ROI information as features (Grill-Spector & Weiner, 2014; Benson et al., 2018) or masks (Wandell et al., 2007), but rarely model cooperative interactions between ROI groups across network depths. In contrast, we partition ROIs into early/mid/late groups with group-specific transformations and couple them to the generator via a depth-matched, ControlNet-based injection (Zhang et al., 2023), supplying spatial priors that localize structure. Together these choices yield a decoder that is not only competitive in accuracy but also more neuro-aligned than purely data-driven alternatives.

3 Method

3.1 Motivation and Overview

Motivation. Human visual cortex is hierarchically organized: early areas encode local and retinotopic structure, mid-level areas integrate contours and object parts, and late ventral regions represent category-level semantics. We use this organization as an inductive bias for fMRI-to-image reconstruction. Instead of mapping all fMRI activity into a single global embedding, we structure the conditioning signal into early, middle, and late ROI streams and inject them into matched depths of a diffusion backbone. Our focus is generator-side organization of subject-specific cortical signals: how anatomically distinct ROI responses can be transformed into compact spatial priors and injected into the diffusion process at appropriate depths.

Overview. As shown in Fig. 2, HI-DREAM contains two lightweight components. First, a ROI adapter (Sec. 3.3) converts trial-wise fMRI responses and subject-specific ROI masks into a multi-scale cortical pyramid. Second, a ROI-conditioned ControlNet (Sec. 3.4) translates this pyramid into depth-matched hints and injects them into the corresponding U-Net blocks during denoising. Given ROI responses and masks, the adapter produces scale-wise priors $\{\Xi_s\}_{s\in\mathcal{S}}$, each Ξ_s is converted into a hint ξ_s and injected through residual + FiLM modulation. The latent diffusion backbone then

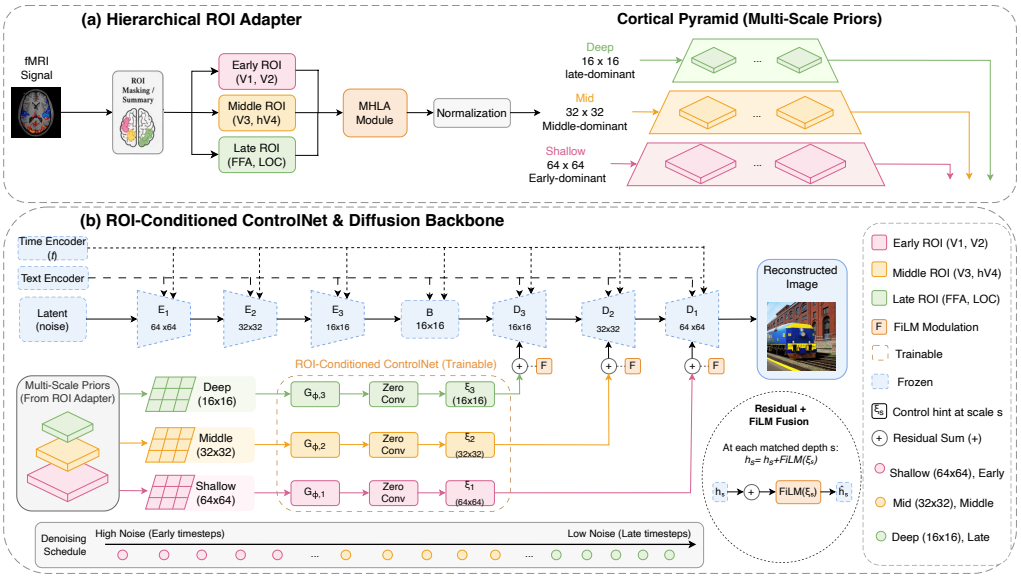

Figure 2: **Overview of HI-DREAM.** (a) The ROI adapter constructs early, middle, and late ROI streams from fMRI responses and converts them into a multi-scale cortical pyramid, with MHLA coordinating cross-stream interactions. (b) The ROI-conditioned ControlNet injects the resulting anatomy-aware priors into matched U-Net depths through residual–FiLM modulation, guiding diffusion reconstruction without full-volume fMRI conditioning and enabling model-level attribution.

denoises under these depth-specific hints to generate the reconstructed image. During inference (Sec. 3.5), we record group weights and injected responses for model-level attribution.

3.2 ROI STREAM CONSTRUCTION

Let the fMRI input be $\mathbf{X} \in \mathbb{R}^{T\times H\times W\times D}$, and let $\{\mathcal{R}^e, \mathcal{R}^m, \mathcal{R}^l\}$ denote the early, middle, and late ROI groups. For each ROI r, we compute a regional response $a_r \in [0,1]$ from the normalized trial-wise beta amplitude and use a subject-specific projected ROI mask $m_r \in \{0,1\}^{H\times W}$ aligned to the image plane. These quantities provide the activity strength and anatomical support for constructing ROI condition maps.

Scale-space construction. We build a spatial pyramid at resolutions $\mathcal{S} = \{64, 32, 16\}$ to match the shallow, middle, and deep U-Net blocks. At scale s, we define a soft ROI mask:

$$\tilde{m}_{r,s} = \mathrm{Down}_s(\mathcal{G}_{\sigma_s} * m_r) \in [0,1]^{h_s\times w_s}, \tag{1}$$

where $\mathcal{G}_{\sigma_s}$ is a Gaussian kernel and Down_s resizes the mask to the corresponding U-Net resolution. The blurred mask itself does not encode stimulus-specific edges; instead, it provides anatomy-constrained spatial scaffolding. Layout and detail preservation arise from the combination of trial-wise ROI activation, shallow depth-matched guidance, and the learned denoising prior.

Per-scale weighting and grouping. Each ROI contributes a trial-specific condition map

$$c_{r,s} = w_{r,s}\, a_r\, \tilde{m}_{r,s}, \tag{2}$$

where $w_{r,s} \in [0,1]$ is a learnable depth-dependent gate. Stacking maps within a ROI group $g \in \{e, m, l\}$ yields a group tensor $\mathbf{C}_{g,s} \in \mathbb{R}^{R_g\times h_s\times w_s}$.

3.3 HIERARCHICAL ROI ADAPTER

The ROI adapter converts the group tensors $\{\mathbf{C}_{g,s}\}$ into compact, scale-wise cortical priors for the diffusion model.

Lightweight channelization. A 1×1 Conv/MLP mixer $M_{g,s}$ compresses each group tensor $\mathbf{C}_{g,s}$ into guidance channels $\Xi_{g,s} \in \mathbb{R}^{d_s\times h_s\times w_s}$, where d_s is small in practice, e.g., $d_s \in \{8, 16, 32\}$. The group-specific guidance maps are then combined with learnable nonnegative mixing weights $\alpha_{g,s}$ to form the scale-wise cortical pyramid:

$$\Xi_s \;=\; \sum_{g\in\{e,m,l\}} \alpha_{g,s}\,\Xi_{g,s}, \qquad s \in \mathcal{S}. \tag{3}$$

Equation 3 allows each U-Net scale to receive a different mixture of early, middle, and late ROI evidence, supporting depth-aware conditioning rather than using a single fixed fMRI representation across all layers.

Multi-Head Latent Attention. The ROI streams are not treated as independent modules. After group-specific features are obtained, we use a Multi-Head Latent Attention (MHLA) (Vaswani et al., 2017) module to model cooperative interactions among early, middle, and late streams. U-Net features act as queries, while ROI-derived latents provide keys and values. The resulting gated attention features modulate the depth-specific ControlNet hints, allowing information from one ROI group to influence how other groups are used at a given denoising depth. Thus, the ROI-ControlNet provides spatial, depth-matched guidance, while MHLA coordinates cross-ROI interactions.

Stability and regularization. To stabilize the learned mixtures, we constrain the ROI gates $w_{r,s}$ and group weights $\alpha_{g,s}$ to be nonnegative and normalize them across groups at each scale. We further apply mild regularization to avoid degenerate single-stream solutions and use a weak anatomical prior that encourages shallow scales to rely more on early ROIs and deeper scales to rely more on late ROIs, while still allowing the final mixing weights to be learned from data.

3.4 ROI-Conditioned ControlNet: Selective Depth Injection

The ROI-conditioned ControlNet translates the adapter pyramid $\{\Xi_s\}$ into resolution-aligned hints that guide image generation. Rather than using ROI features as generic conditioning, we inject each hint into the U-Net depth where the corresponding scale is most informative.

Resolution-aligned injection. For each scale s matched to a U-Net block with activation h_s, a shallow stack $G_{\phi,s}$ of 3×3 convolutions produces a hint $\xi_s = G_{\phi,s}(\Xi_s)$. We inject it using residual-FiLM fusion:

$$\hat{h}_s \;=\; h_s \;+\; \lambda_s\Big(A_s(\xi_s) \;+\; \gamma_s(\xi_s)\odot h_s \;+\; \beta_s(\xi_s)\Big), \tag{4}$$

where A_s is a 1×1 convolution, (γ_s, β_s) are per-channel affine parameters, and $\lambda_s \in [0, 1]$ controls the guidance strength at depth s.

Hint normalization and calibration. We normalize each hint to zero mean and unit variance, and bound the total injected energy by a soft budget:

$$\sum_s \lambda_s \cdot \mathbb{E}[\|A_s(\xi_s)\|_1] \le \eta, \tag{5}$$

implemented by on-the-fly rescaling of $\{\lambda_s\}$. This prevents any single scale from overwhelming the denoising backbone.

Depth and time scheduling. We encourage early, middle, and late ROI evidence to contribute more strongly at shallow, intermediate, and deep U-Net blocks, respectively. Guidance is also scheduled over diffusion time:

$$\lambda_s(t) = \lambda_s^{\max} \cdot \tfrac{1}{2}\big(1 + \cos(\pi \cdot \mathrm{clip}(t/T, 0, 1))\big)^{\rho_s}, \tag{6}$$

so structural hints can guide early denoising while deeper semantic hints refine the reconstruction later.

3.5 Training and Inference

Objective. We optimize a diffusion objective augmented with lightweight structural regularization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}} + \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_{\text{grad}} + \lambda_3 (1 - \text{SSIM}), \tag{7}$$

where $\mathcal{L}_1$, $\mathcal{L}_{\text{grad}}$, and $1 - \text{SSIM}$ encourage pixel consistency, edge/gradient preservation, and structural similarity, respectively. We additionally retain regularizers on α and w to stabilize ROI mixing and prevent single-stream dominance.

Two-stage schedule. In Stage A, we freeze the diffusion backbone and warm up the ROI adapter to produce stable cortical pyramids. The mixing weights $\alpha_{g,s}$ are initialized conservatively, and λ_s is ramped up from small values to avoid unstable early conditioning. In Stage B, we enable the ROI-conditioned ControlNet and selectively unfreeze only the U-Net blocks aligned with each scale. The ControlNet learns to translate $\{\Xi_s\}$ into effective guidance signals, while the partially unfrozen backbone adapts locally and preserves the pretrained generative prior. All experiments use the training objective in Eq. 7, together with depth-wise guidance $(\lambda_{\text{shallow}}, \lambda_{\text{mid}}, \lambda_{\text{deep}}) = (1.2, 0.6, 0.3)$.

Inference and complexity. At test time, $\{\Xi_s\}$ is computed in one pass and DDIM/PLMS sampling proceeds with fixed $(\alpha_{g,s}, \lambda_s)$. The adapter and ControlNet operate on compact 2D pyramids, adding negligible overhead relative to the U-Net and avoiding the I/O cost of repeatedly conditioning on full 3D fMRI volumes.

4 Experiment

4.1 Dataset and Metrics

Dataset. We evaluate our method on the Natural Scenes Dataset (NSD) (Allen et al., 2022), a large-scale 7T fMRI dataset with natural-image stimuli. NSD provides trial-wise beta responses and subject-specific anatomical/ROI information, including retinotopic visual areas and higher-level category-selective regions. Following common practice in fMRI-to-image reconstruction, we report results on subjects 1, 2, 5, and 7, who completed the full scanning schedule. We group the visual ROIs into early areas (V1/V2), middle areas (V3/hV4), and late ventral areas (LOC/FFA), which serve as the cortical streams used by HI-DREAM.

Evaluation Metrics. We follow the standard evaluation protocol used in prior fMRI reconstruction work (Scotti et al., 2024; Wang et al., 2024) and report metrics covering low-level, perceptual, and semantic fidelity. Low-level fidelity is evaluated using PixCorr, SSIM (Wang et al., 2004) and AlexNet(2/5) (Krizhevsky et al., 2012) (↑), which measure pixel-, structure-, and early perceptual-level similarity. High-level semantic alignment is evaluated using Inception (Szegedy et al., 2016) and CLIP (Radford et al., 2021) (↑), together with EfficientNet-B distance Tan & Le (2019) and SwAV distance (Caron et al., 2020) (↓), to assess category-level consistency and feature-level discrepancy with the target stimulus. For fair comparison with previous methods, all reconstructions are evaluated at 256×256 resolution using the same preprocessing and metric definitions across baselines.

4.2 Qualitative & Quantitative Results

Main results.

Table 1 compares HI-DREAM with previous state-of-the-art (SoTA) fMRI-to-image reconstruction methods, including Takagi (Takagi & Nishimoto, 2023), MindBridge (Wang et al., 2024), DREAM (Xia et al., 2024), Psychometry (Quan et al., 2024), and MindEye2 (Scotti et al., 2024).

HI-DREAM achieves the strongest high-level semantic alignment, obtaining the best results on Incep, CLIP, Eff, and SwAV. Qualitatively, this means our reconstructions capture object identity and scene semantics more reliably, while avoiding the mode-mismatch sometimes observed when conditioning directly on full fMRI volumes. It also performs strongly on low-level and perceptual metrics, achieving the best AlexNet(2) score and second-best PixCorr and AlexNet(5) scores. These results

Table 1: Quantitative comparison between HI-DREAM and previous SoTA fMRI-to-image reconstruction methods on low-level and high-level metrics. All results are averaged across subjects 1, 2, 5, and 7 from the Natural Scenes Dataset. Best results are shown in **bold**, and second-best results are underlined.

Methods	Low-Level				High-Level			
	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
Takagi et al.	.246	.410	78.9%	85.6%	83.8%	82.1%	.811	.504
MindBridge	.148	.259	86.9%	95.3%	92.2%	94.3%	.713	.413
Psychometry	.297	.340	96.4%	**98.6%**	95.8%	96.8%	.628	.345
DREAM	.274	.328	93.9%	96.7%	93.4%	94.1%	.645	.418
MindEye2	**.322**	**.431**	96.1%	**98.6%**	95.4%	94.5%	.619	.344
HI-DREAM	.316	.407	**96.8%**	97.4%	**97.2%**	**98.3%**	**.541**	**.342**

suggest that hierarchy-aware ROI conditioning improves semantic reconstruction while preserving competitive structural and perceptual fidelity.

Fig. 3 further shows that our reconstructions achieve consistently higher semantic correctness than prior methods. Across diverse stimuli, our results preserve the correct category and salient attributes, with notably better color fidelity (e.g., the red flower bouquet and piano scene, the characteristic color palettes in food images) and more accurate shape/geometry (object silhouettes, part proportions, and pose). We also retain more coherent scene layout and context, reducing common failure modes in baselines such as color drift, category confusion, or structural collapse. Overall, our reconstructions are closer to the intended content of the stimulus, indicating stronger high-level semantic alignment.

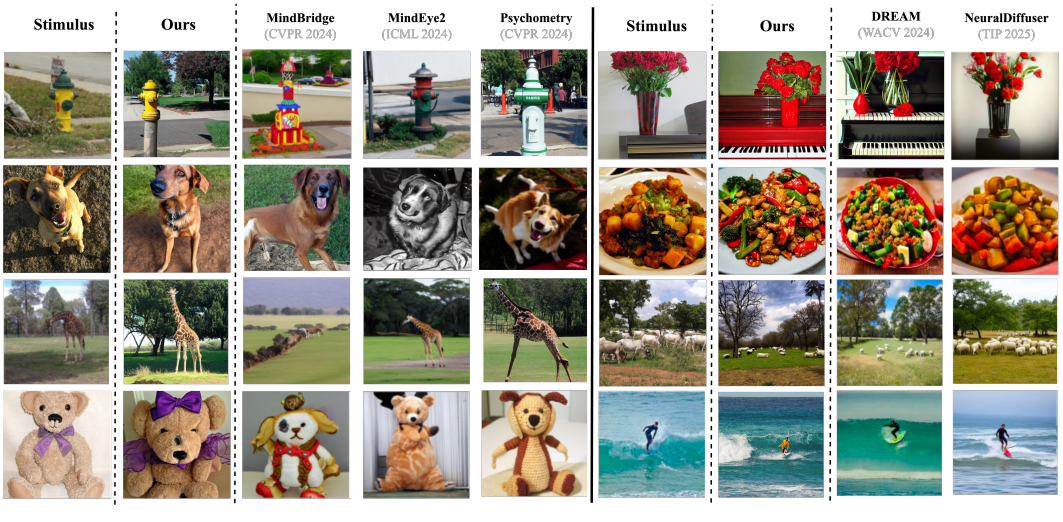

Figure 3: Qualitative comparison on the NSD test set (NSD Shared 1000). Each group shows the stimulus (left) and the corresponding reconstructions by HI-DREAM and prior methods.

Why HI-DREAM helps. Compared with prior methods, two design choices appear decisive: (i) a brain-inspired ROI adapter that separates early/mid/late groups and builds a multi-scale cortical pyramid. This aligns the information each group provides with the depth of the U-Net (layout-sensitive structure at shallow layers, parts/semantics at deeper layers), reducing representational frictions that arise when a single embedding is forced to serve all depths. (ii) a selective, depth-matched ControlNet that injects ROI evidence where it is most useful, with per-scale strength control. This avoids over-conditioning and prevents shallow detail from drowning out deeper semantic cues. Together, these choices yield stronger semantic metrics without substantially increasing com-

putation: the adapter and ControlNet operate on compact subject-specific maps rather than full 3D fMRI volumes, improving data throughput and simplifying optimization.

Relation to specific baselines. MindBridge Wang et al. (2024) focuses on cross-subject knowledge transfer; Psychometry Quan et al. (2024) optimizes perceptual and semantic objectives with powerful priors; MindEye2 Scotti et al. (2024) integrates fMRI encoders with diffusion via cross-attention. HI-DREAM is orthogonal and complementary: instead of seeking a single best fMRI embedding, we structure the conditioning by cortical hierarchy and scale, which preserves interpretability (each signal maps to an interpretable ROI stream) while driving gains on semantically grounded metrics. The consistent advantage on Inception/CLIP and feature distances, averaged over the four full-coverage NSD subjects, supports this interpretation.

4.3 ABLATION STUDY

Composition of ROI streams. We first examine how each cortical stream contributes to reconstruction. The ROI adapter separates fMRI signals into Early, Middle, and Late streams, corresponding to retinotopic structure, intermediate shape/part information, and category-selective semantics. Rather than treating these streams as isolated predictors, HI-DREAM combines them through a shared diffusion decoder and MHLA, allowing their contributions to interact across denoising depths. Table 2 reports the reconstruction quality when using Early only, Early+Middle, and all three streams.

Table 2: ROI-stream ablation showing complementary contributions of Early, Middle, and Late streams.

ROI Streams	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
E	.267	**.421**	92.1%	94.0%	92.4%	92.1%	.628	.431
E+M	.294	.402	95.1%	96.5%	95.5%	96.6%	.572	.363
E+M+L	**.316**	.407	**96.8%**	**97.4%**	**97.2%**	**98.3%**	**.541**	**.342**

Early-only conditioning achieves strong structural fidelity, especially on SSIM, suggesting that early retinotopic ROIs provide useful spatial scaffolding. Adding Middle improves both low-level and semantic metrics, indicating that mid-level ROIs bridge local structure and part-level information.

The full Early + Middle + Late model obtains the best overall balance, with the highest PixCorr, AlexNet, Inception, and CLIP scores and the lowest feature distances. These results suggest that the streams contribute complementary interacting biases rather than independent linear effects.

To test whether the proposed hierarchy is an arbitrary grouping, we compare against two controls with the same model capacity in Table 3. 'Random' assigns ROIs randomly to the three streams, and for 'Reversed', we reverse the anatomical depth prior and ROI-to-depth initialization, so that late ROIs are favored at shallow scales and early ROIs at deep scales, while keeping the remaining architecture and parameter count unchanged. Both controls degrade performance, especially Reversed on high-level metrics, indicating that the depth–ROI correspondence is important for effective conditioning.

Table 3: Hierarchy-control experiments with the same model configuration. Random and reversed ROI-depth assignments degrade performance, indicating the importance of matching cortical hierarchy to U-Net depth.

Setting	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
Random	.281	.356	92.7%	95.1%	94.3%	95.6%	.589	.381
Reversed	.264	.331	90.8%	93.6%	91.7%	93.2%	.628	.414
Ours	**.316**	**.407**	**96.8%**	**97.4%**	**97.2%**	**98.3%**	**.541**	**.342**

Role of the ROI Adapter and MHLA. We next isolate the contribution of the hierarchical adapter and the MHLA module. The ROI adapter converts subject-specific cortical activity into multi-scale anatomy-aware priors, while MHLA coordinates interactions among Early, Middle, and Late

streams before depth-specific injection. This separation allows us to test whether performance gains come from spatial ROI conditioning alone, or from explicitly modeling cross-ROI cooperation.

Table 4: Ablation of hierarchy modules. The ROI Adapter provides anatomy-aware spatial priors, while MHLA further improves cross-stream coordination and semantic alignment.

Adapter	MHLA	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
×	×	.096	.191	76.8%	79.4%	84.0%	83.5%	.718	.543
✓	×	.205	.302	84.6%	88.8%	92.2%	92.1%	.626	.487
✓	✓	**.316**	**.407**	**96.8%**	**97.4%**	**97.2%**	**98.3%**	**.541**	**.342**

As shown in Table 4, without the ROI adapter, the ControlNet branch lacks structured cortical priors and performs poorly across both low- and high-level metrics. Adding the ROI adapter improves reconstruction by supplying depth-aligned spatial guidance.

Adding MHLA further improves both low-level fidelity and high-level semantic alignment, while also reducing feature distances, suggesting that cross-stream interaction benefits structural and semantic reconstruction simultaneously. In short, the adapter provides anatomical priors, and MHLA coordinates how these priors cooperate across U-Net depths.

4.4 Discussion

Complementary roles of ROI streams. The ablation results show that HI-DREAM benefits from integrating cortical streams in a depth-aware manner rather than simply adding more conditioning parameters. Early-only conditioning provides strong structural fidelity, adding Middle improves part-level and semantic consistency, and the full Early-Middle-Late configuration achieves the best overall balance. The Random and Reversed controls further confirm that the correspondence between ROI groups and U-Net depth is important for effective reconstruction.

Interpretability and practicality. HI-DREAM also provides an inspectable and efficient conditioning pathway. Grad-CAM visualizations and depth-wise contribution analysis show distinct ROI-dependent attention patterns and hierarchy-consistent contributions across U-Net depths, which should be interpreted as model-level rather than biological evidence. By operating on compact 2D cortical pyramids instead of full 3D fMRI volumes, the framework also keeps conditioning lightweight and leaves room for extensions to broader ROI sets and other brain decoding tasks.

4.5 Limitations

HI-DREAM still has several limitations. Fine-grained details such as small textures, object boundaries, and facial configurations remain challenging, especially when semantic guidance dominates local structure. The current ROI set is also relatively compact and does not cover the full diversity of visual cortex. Finally, the framework is subject-specific and relies on subject-specific ROI masks, which limits direct transfer across individuals. Future work will explore broader cortical coverage and test whether the generator-side hierarchy remains effective under greater subject variability.

5 Conclusion

In this paper, we introduced HI-DREAM, a brain-inspired diffusion framework that organizes subject-specific fMRI signals through a cortical hierarchy. By transforming early, middle, and late ROI streams into a multi-scale cortical pyramid and injecting them through a depth-matched ROI-ControlNet, HI-DREAM provides structured, anatomy-aware conditioning while retaining an inspectable generator-side pathway. On the NSD test set, HI-DREAM reaches SoTA performance on high-level semantic metrics while retaining strong low-level structure. Ablation and attribution analyses further demonstrate the benefit of hierarchy-aware conditioning and the complementary roles of different ROI streams. Future work will focus on improving fine-grained reconstruction and extending the framework to broader cortical coverage, and moving toward a decoder that more closely mirrors human visual processing.

AI USE STATEMENT

In this work, we used generative AI tools for no activities falling under required disclosure. We have not used generative AI tools for developing the method, designing or conducting experiments, processing experimental data, producing quantitative results, or drawing scientific interpretations, and the remaining required-disclosure activities are not applicable to this work. Additionally, we used generative AI tools for language polishing and manuscript editing, and for assisting with the review and organization of existing literature. We have reviewed all AI-assisted work. AI-assisted revisions were checked and edited by the authors, and references identified or organized with AI assistance were verified against the corresponding original publications. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We describe the key methodological and experimental settings needed to reproduce HI-DREAM in the main paper and supplementary appendix. Section 3 provides the complete model formulation, including ROI-stream construction, the hierarchical ROI adapter, depth-aware ControlNet conditioning, and the training and inference objectives. Section 4.1 specifies the NSD subjects, evaluation protocol, preprocessing resolution, and reconstruction metrics. The appendix further documents implementation settings, fMRI-to-ROI condition map construction, preprocessing safeguards, inference configuration, computational cost, and the scale–depth mapping used by the model. These details are provided to facilitate independent reproduction and verification of the reported results. Anonymous source code is also provided as supplementary material to support implementation-level reproducibility.

REFERENCES

Himanshu Aggarwal, Liza Al-Shikhley, and Bertrand Thirion. Across-subject ensemble-learning alleviates the need for large samples for fmri decoding. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 35–45. Springer, 2024.

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.

John Ashburner and Karl J Friston. Voxel-based morphometry—the methods. *Neuroimage*, 11(6): 805–821, 2000.

Sylvain Baillet. Magnetoencephalography for brain electrophysiology and imaging. *Nature neuroscience*, 20(3):327–339, 2017.

Peter A Bandettini. What's new in neuroimaging methods? *Annals of the New York Academy of Sciences*, 1156(1):260–293, 2009.

Noah C Benson, Keith W Jamison, Michael J Arcaro, An T Vu, Matthew F Glasser, Timothy S Coalson, David C Van Essen, Essa Yacoub, Kamil Ugurbil, Jonathan Winawer, et al. The human connectome project 7 tesla retinotopy dataset: Description and population receptive field analysis. *Journal of vision*, 18(13):23–23, 2018.

Robert W Brown, Y-C Norman Cheng, E Mark Haacke, Michael R Thompson, and Ramesh Venkatesan. *Magnetic resonance imaging: physical principles and sequence design*. John Wiley & Sons, 2014.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22710–22720. IEEE, 2023a.

Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity. *Advances in Neural Information Processing Systems*, 36:24841–24858, 2023b.

Michael X Cohen. Where does eeg come from and what does it mean? *Trends in neurosciences*, 40 (4):208–218, 2017.

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

Emily S Finn, Xilin Shen, Dustin Scheinost, Monica D Rosenberg, Jessica Huang, Marvin M Chun, Xenophon Papademetris, and R Todd Constable. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature neuroscience*, 18(11):1664–1671, 2015.

Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

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.

Kalanit Grill-Spector and Rafael Malach. The human visual cortex. *Annu. Rev. Neurosci.*, 27(1): 649–677, 2004.

Kalanit Grill-Spector and Kevin S Weiner. The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15(8):536–548, 2014.

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

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

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

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

Zhoufan Jiang, Yanming Wang, ChenWei Shi, Yueyang Wu, Rongjie Hu, Shishuo Chen, Sheng Hu, Xiaoxiao Wang, and Bensheng Qiu. Attention module improves both performance and interpretability of 4d fmri decoding neural network. *arXiv preprint arXiv:2110.00920*, 2021.

Jaehoon Joo, Taejin Jeong, and Seongjae Hwang. Brain-streams: fmri-to-image reconstruction with multi-modal guidance. *arXiv preprint arXiv:2409.12099*, 2024.

Kendrick N Kay, Jonathan Winawer, Aviv Mezer, and Brian A Wandell. Compressive spatial summation in human visual cortex. *Journal of neurophysiology*, 110(2):481–494, 2013.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Chong Li, Xuelin Qian, Yun Wang, Jingyang Huo, Xiangyang Xue, Yanwei Fu, and Jianfeng Feng. Enhancing cross-subject fmri-to-video decoding with global-local functional alignment. In *European Conference on Computer Vision*, pp. 353–369. Springer, 2024.

Haoyu Li, Hao Wu, and Badong Chen. Neuraldiffuser: Neuroscience-inspired diffusion guidance for fmri visual reconstruction. *IEEE Transactions on Image Processing*, 34:552–565, 2025.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

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

Yizhuo Lu, Changde Du, Qiongyi Zhou, Dianpeng Wang, and Huiguang He. Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 5899–5908, 2023.

Yizhuo Lu, Changde Du, Chong Wang, Xuanliu Zhu, Liuyun Jiang, Xujin Li, and Huiguang He. Animate your thoughts: Reconstruction of dynamic natural vision from human brain activity. In *International Conference on Learning Representations*, volume 2025, pp. 67529–67567, 2025.

Na Luo, Weiyang Shi, Zhengyi Yang, Ming Song, and Tianzi Jiang. Multimodal fusion of brain imaging data: Methods and applications. *Machine Intelligence Research*, 21(1):136–152, 2024.

Donald W McRobbie, Elizabeth A Moore, Martin J Graves, and Martin R Prince. *MRI from Picture to Proton*. Cambridge university press, 2007.

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

Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19):1641–1646, 2011.

Youzhi Qu, Junfeng Xia, Xinyao Jian, Wendu Li, Kaining Peng, Zhichao Liang, Haiyan Wu, and Quanying Liu. Uncovering cognitive taskonomy through transfer learning in masked autoencoder-based fmri reconstruction. In *International Workshop on Human Brain and Artificial Intelligence*, pp. 35–50. 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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 233–243. IEEE, 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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

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.

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

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

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

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PmLR, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Brian A Wandell and Jonathan Winawer. Computational neuroimaging and population receptive fields. *Trends in cognitive sciences*, 19(6):349–357, 2015.

Brian A Wandell, Serge O Dumoulin, and Alyssa A Brewer. Visual field maps in human cortex. *Neuron*, 56(2):366–383, 2007.

Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Mindbridge: A cross-subject brain decoding framework. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11333–11342. IEEE, 2024.

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

Weihao Xia, Raoul De Charette, Cengiz Oztireli, and Jing-Hao Xue. Dream: Visual decoding from reversing human visual system. In *2024 IEEE/CVF winter conference on applications of computer vision (WACV)*, pp. 8211–8220. IEEE, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824. IEEE, 2023.

Jianfei Zhu, Baichun Wei, Jiaru Tian, Feng Jiang, and Chunzhi Yi. An adaptively weighted averaging method for regional time series extraction of fmri-based brain decoding. *IEEE Journal of Biomedical and Health Informatics*, 28(10):5984–5995, 2024.

APPENDIX

A APPENDIX OVERVIEW

- **Implementation Details** (B): Reproducibility details for preprocessing, inference, and computational cost.
- **fMRI → ROI Condition Maps** (C): Construction of three-channel ROI condition maps from trial-wise betas and subject-specific ROI masks.
- **Model Details** (D): Additional condition-map statistics and scale–depth mappings used by the ROI-conditioned ControlNet.
- **Qualitative ROI-stream attribution.** (E): Visualizes stream-specific reconstruction biases using ROI ablations and Grad-CAM overlays.
- **Depth-wise ROI contribution.** (F): Examines normalized injected energy across U-Net depths to verify hierarchy-consistent ROI usage.

B IMPLEMENTATION DETAILS

This section provides implementation details that complement the method description in the main paper section 3. We focus on reproducibility, including data indexing, preprocessing, optimization, inference, and evaluation settings.

Implementation environment. We implement HI-DREAM in PyTorch 2.x with CUDA 12.x and use the diffusers implementation of the latent diffusion U-Net. Mixed precision is enabled during training, using bf16 when supported and otherwise fp16. We use fixed random seeds for data loading, initialization, and sampling; deterministic cuDNN options are enabled when available.

fMRI and ROI preprocessing. We use the preprocessed NSD trial-wise beta responses and subject-specific ROI annotations. Volumetric ROIs are mapped to the model plane using the NSD anatomical transforms and projected to 256×256 resolution. For each trial, ROI responses are normalized using training-set statistics and clipped to reduce the influence of outlier beta values. Spatial augmentations that would alter retinotopic correspondence, such as horizontal flipping or geometric warping, are not applied to ROI condition maps. Image inputs are normalized to $[0, 1]$; we use only light color jittering for image-side augmentation.

Inference. At test time, ROI condition maps are computed once and reused throughout sampling. Unless otherwise stated, classifier-free guidance is disabled. We use DDIM sampling with 50 steps and the same depth-wise guidance strengths as in training. The guidance schedule follows the cosine time schedule defined in the main paper section 3.4. All reconstructions are generated at 256×256 resolution.

Computational cost. On an A100 80GB GPU, the ROI adapter and ROI-conditioned ControlNet add less than 5% training step time relative to the latent diffusion backbone. Because the conditioning maps are compact 2D pyramids and are cached after preprocessing, the additional I/O and memory overhead is small compared with directly conditioning on full 3D fMRI volumes.

C FMRI → ROI CONDITION MAPS (3-CHANNEL)

We convert trial-wise GLM betas into a three-channel cortical prior aligned with the 2D model plane, with one channel for each ROI group: Early, Middle, and Late. For subject i and trial t, let $\mathcal{R}^e, \mathcal{R}^m, \mathcal{R}^l$ denote the ROI sets and let $b_v^{(t)}$ be the beta response of voxel v.

(1). Robust ROI activity. For each ROI r, voxel betas are normalized using training-set statistics. We winsorize the top and bottom 1% of responses and summarize the ROI activity using a Huber

mean to obtain $\tilde{a}_r^{(t)}$. The response is then squashed to $[0, 1]$ with a sigmoid using a per-ROI scale κ_r estimated on the training set:

$$a_r^{(t)} = \sigma(\tilde{a}_r^{(t)}/\kappa_r). \tag{1}$$

Trials with missing betas or motion outliers are excluded.

(2). 2D ROI projection. Subject-specific ROI masks $M_r \in \{0,1\}^{H_f \times W_f \times D_f}$ from FreeSurfer/retinotopy annotations are mapped from functional space to a 2D model plane using NSD anatomical transforms and rasterized at the model resolution $H \times W$ (here 256×256). This produces a projected ROI mask $m_r \in \{0,1\}^{H\times W}$. We optionally erode the mask by one voxel to reduce partial-volume effects and apply a small Gaussian blur (σ=1 pixel) to obtain soft spatial boundaries.

(3). Group-wise condition maps. For each group $g \in \{e, m, l\}$, we construct an area-normalized weighted mask accumulation:

$$\begin{aligned} C_g^{(t)}(x,y) &= \frac{1}{Z_g} \sum_{r\in\mathcal{R}^g} a_r^{(t)}\, m_r(x,y), \\ Z_g &= \sum_{r\in\mathcal{R}^g} \max\left\{1,\ \sum_{x,y} m_r(x,y)\right\}. \end{aligned} \tag{2}$$

This normalization prevents large ROIs from dominating the condition map. We then min–max normalize each $C_g^{(t)}$ to $[0, 1]$ per subject and apply instance normalization at load time to stabilize statistics across sessions.

(4). Stacking and caching. The three group maps are stacked channel-wise into $C^{(t)} \in \mathbb{R}^{3\times H\times W}$ in the fixed order [Early, Middle, Late]. The resulting tensors are stored as float16 with metadata including subject ID, session, 73KID, and ROI version, and are cached for reuse during training and inference.

(5). Safeguards. To preserve retinotopic correspondence, we do not apply spatial flips, geometric warps, or brightness/contrast jitter to the condition maps. During training, $C^{(t)}$ is used to construct the scale-specific ROI tensors for the adapter. For qualitative attribution, the same maps are normalized as heatmaps and overlaid on the stimulus. Overall, this pipeline provides a lightweight and reproducible three-channel cortical prior while preserving subject-specific spatial locality.

D Model Details

Condition map statistics. To verify that the ROI condition maps have stable subject-wise statistics, we summarize the three-channel cortical priors used by Hi-DREAM across the four evaluated NSD subjects. For each trial, GLM betas are summarized within each ROI, projected to the 2D model plane, and accumulated into Early, Middle, and Late condition-map channels. Table A1 reports the median across trials for each subject and channel, including the mean intensity, robust percentile range, and sparsity. Across subjects, Early maps are spatially broader, Middle maps have intermediate support, and Late maps are more compact, consistent with their intended roles in the hierarchical conditioning design.

Scale–depth mapping. The ROI-conditioned ControlNet injects the cortical pyramid into three matched U-Net depths, following the shallow/middle/deep structure used in the main method. Table A2 summarizes the mapping between condition-map scale, U-Net depth, fusion type, and primary role. The shallow 64×64 scale is aligned with early visual ROIs and mainly supports layout and structural detail; the middle 32×32 scale bridges object parts and intermediate structure; and the deep 16×16 scale is aligned with late visual ROIs and supports semantic refinement.

All three scales use the residual + FiLM fusion described in the main paper section 3.4. The injection sites are fixed by design, while the ROI gates, group mixing weights, and depth-wise strengths are learned or calibrated from data. This allows the model to preserve the intended cortical-to-generative correspondence without forcing a single ROI group to dominate every denoising layer.

Table A1: Summary statistics of three-channel ROI condition maps. Values are medians across trials for each subject.

Subject	Channel	Mean±SD	P10–P90	Sparsity
Subj01	Early	0.13±0.05	0.02–0.30	8.1%
	Middle	0.09±0.04	0.01–0.20	5.6%
	Late	0.06±0.03	0.00–0.12	2.4%
Subj02	Early	0.11±0.04	0.01–0.26	7.3%
	Middle	0.08±0.03	0.01–0.18	5.1%
	Late	0.05±0.02	0.00–0.11	2.1%
Subj05	Early	0.14±0.06	0.02–0.32	8.6%
	Middle	0.10±0.04	0.01–0.22	5.9%
	Late	0.07±0.03	0.00–0.13	2.6%
Subj07	Early	0.12±0.05	0.02–0.28	7.9%
	Middle	0.09±0.03	0.01–0.19	5.3%
	Late	0.05±0.03	0.00–0.12	2.3%

Table A2: Scale–depth mapping used by the ROI-conditioned ControlNet.

Scale	U-Net depth	Dominant ROI stream	Primary role
64×64	Shallow blocks	Early	Layout / structural detail
32×32	Middle blocks	Middle	Parts / structure bridge
16×16	Deep blocks	Late	Category / semantic refinement

E QUALITATIVE ROI-STREAM ATTRIBUTION.

To complement the quantitative ablations, we visualize how different ROI streams affect the reconstruction process using Grad-CAM overlays. Fig. A1 shows reconstructions from the same stimulus when conditioning on Early-only, Middle-only, Late-only, and all ROI streams.

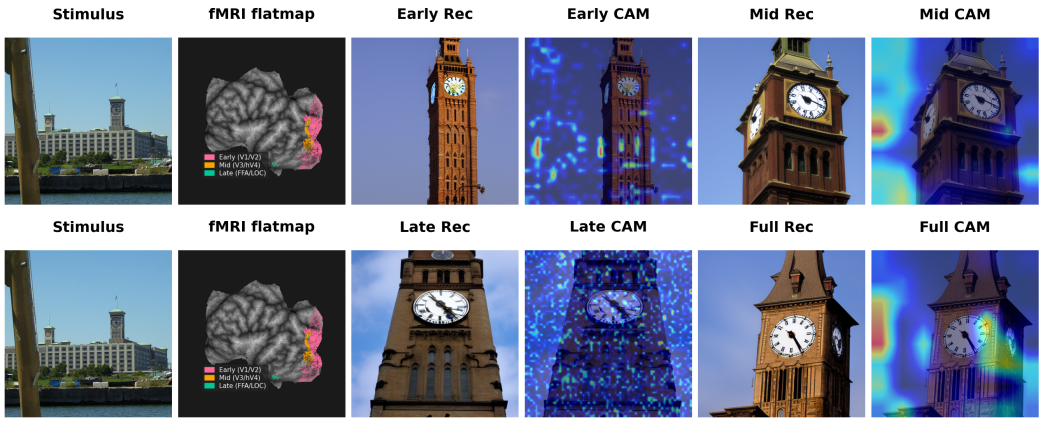

Figure A1: Qualitative ROI-stream ablation and Grad-CAM visualization. For the same stimulus and projected fMRI flatmap, we compare reconstructions generated from Early-only, Middle-only, Late-only, and Full ROI conditioning. Grad-CAM overlays show how different ROI streams induce distinct spatial attention patterns during reconstruction.

The ROI-specific reconstructions reveal distinct visual biases. With Early-only conditioning, the model roughly preserves the global scene layout, including the tall tower-like structure and the surrounding building silhouette, but the object identity and fine architectural details remain unstable. With Middle-only conditioning, the reconstruction better captures part-level structure, such as the

clock face, roof shape, and local building components, although the overall geometry is still distorted. With Late-only conditioning, the model emphasizes category-level cues of a clock tower, but the reconstruction becomes less spatially grounded and loses much of the original scene context.

The full model combines these complementary signals and produces a more balanced reconstruction, preserving both the clock-tower identity and its spatial arrangement. Its Grad-CAM overlay is also more spatially organized, with strong responses around visually salient regions such as the foreground railing/occlusion and the clock face. These regions are also naturally informative for human recognition of the scene, suggesting that the full model attends to semantically meaningful and structurally important parts of the image. This qualitative analysis is consistent with the quantitative ablations: individual ROI streams induce different reconstruction biases, whereas their combination provides a more coherent structural-semantic reconstruction.

F Depth-wise ROI contribution.

We further examine whether Hi-DREAM uses each ROI stream at the intended U-Net depth. For ROI group $g \in \{\text{Early}, \text{Middle}, \text{Late}\}$ and depth s, we measure the normalized injected energy

$$\mathcal{I}_{g,s} = \mathbb{E}_{x,t}\left[\|\alpha_{g,s}\,\Xi_{g,s}(x,t)\|_1\right], \tag{3}$$

where $\Xi_{g,s}$ is the group-specific map before aggregation and $\alpha_{g,s}$ is the learned group weight. We row-normalize $\mathcal{I}_{g,s}$ across ROI groups at each depth and average over the test set, so larger values indicate stronger reliance on that ROI stream.

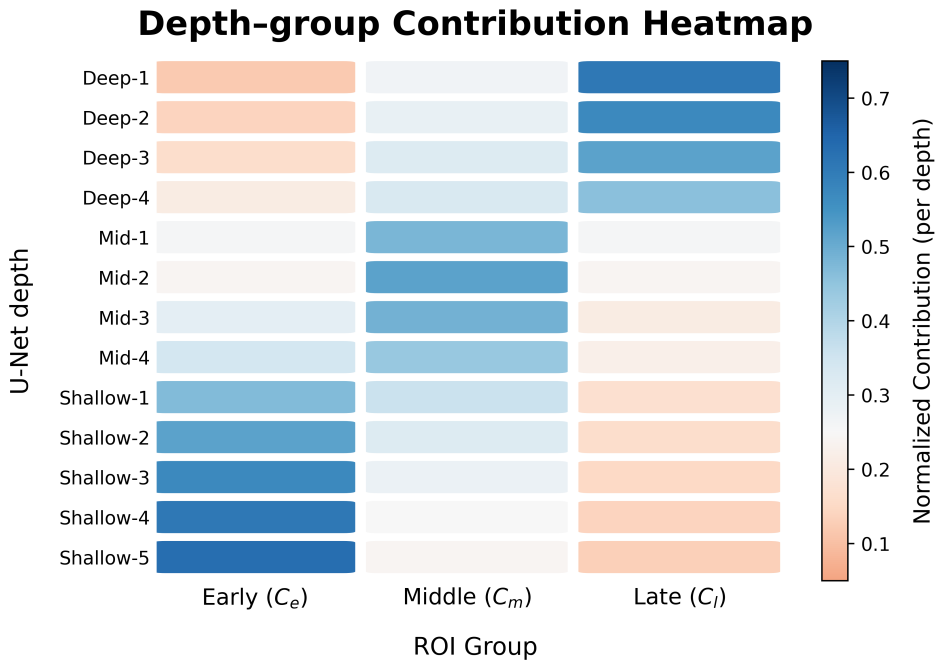

Figure A2: Depth–ROI alignment measured by normalized injected energy. Rows are ordered from deep to shallow U-Net blocks; darker blue indicates higher relative contribution.

As shown in Fig. A2, the learned contribution pattern follows the intended hierarchy: Late, Middle, and Early streams show stronger relative contributions at deep, intermediate, and shallow blocks, respectively. Since the injection sites are fixed but the group weights and gates are learned from data, this pattern indicates that the model does not simply use ROI maps as generic conditions. Instead, Hi-DREAM learns a depth-consistent use of cortical streams, supporting the design that early ROIs provide spatial scaffolding, middle ROIs bridge part-level structure, and late ROIs refine semantic content.