MIND-BIRTH: LEARNING HIERARCHICAL CANONICAL CORTICAL REPRESENTATIONS FOR ZERO-SHOT CROSS-SUBJECT VISUAL DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstructing perceived images from human fMRI offers a way to study neural representations and visual processing, but most existing approaches remain strongly subject-specific. Cross-subject decoding is particularly challenging because individuals differ in voxel coverage, spatial organization, and response distributions, making it difficult to learn representations that transfer reliably to unseen subjects. We propose MIND-BIRTH, a framework for learning hierarchical canonical cortical representations for cross-subject visual decoding. MIND-BIRTH organizes visual cortical responses into Early, Middle, and Late stages, converts heterogeneous ROI responses into coordinate-anchored tokens, and maps them into stage-wise canonical spaces using visual-semantic alignment. The learned representations are then decoded through hierarchy-matched diffusion conditioning, preserving the correspondence between cortical hierarchy and generator depth. This design supports zero-shot reconstruction of held-out subjects without target-subject training and lightweight few-shot personalization when limited target data are available. Experiments on NSD show that MIND-BIRTH achieves strong semantic reconstruction performance among cross-subject methods under both zero-shot and few-shot settings while remaining competitive in low-level fidelity. Controlled ablations further validate the importance of stage-wise cortical representations, canonical alignment, and hierarchy-matched routing. These results suggest that learning transferable, hierarchy-aware cortical representations provides an effective inductive bias for cross-subject visual decoding.

1 INTRODUCTION

Decoding visual experiences from human brain activity through fMRI (Brown et al., 2014; McRobbie et al., 2007) has long been a central goal in cognitive neuroscience, with implications that go beyond engineering: it offers a concrete way to test theories of neural representation, to probe what information is preserved across the visual hierarchy, and to develop new tools for studying perception and imagery (Ogawa et al., 1990; Logothetis et al., 2001; Glover, 1999). Recent progress has been accelerated by large-scale datasets and powerful generative priors, enabling increasingly photo-realistic reconstructions from fMRI signals using diffusion models and related generative pipelines (Takagi & Nishimoto, 2023; Scotti et al., 2024; Wang et al., 2024; Chen et al., 2023b;a; Gong et al., 2025; Lin et al., 2022).

Early diffusion-based (Ho et al., 2020; Dhariwal & Nichol, 2021; Ho & Salimans, 2022; Rombach et al., 2022) brain decoding systems typically relied on subject-specific training (Chen et al., 2023a; Shen et al., 2019), achieving strong reconstructions but requiring substantial data collection and optimization for each new participant. This subject-by-subject paradigm is costly and limits practical deployment and large-scale neuroscience studies. To address this, a growing line of work targets cross-subject generalization, aiming to transfer decoding models to unseen individuals with little or even no target-subject training (Wang et al., 2026; 2024; Huo et al., 2024). Despite impressive quality improvements, most of these methods remain primarily data driven and they focus on learning a shared latent space or a universal mapping across participants, often treating fMRI as a single flattened signal while relying on model capacity and alignment losses to close the subject gap.

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.

However, human vision is not processed as a monolithic stream. Neuroscience evidence supports a simple-feature-to-complex-shape progression from early visual cortex to higher-level regions (Bandettini, 2009; Baillet, 2017; Cohen, 2017), with distinct representational roles across stages (DiCarlo et al., 2012; Wandell & Winawer, 2015). Collapsing heterogeneous ROI signals into a single latent space can obscure this structure and promote shortcuts, yielding entangled features that overfit subject-specific idiosyncrasies and generalize poorly across subjects.

In this work we propose **MIND-BIRTH**, a brain-inspired diffusion decoding framework that explicitly mirrors the hierarchical organization of visual processing. Our key idea is to transform heterogeneous subject-specific cortical responses into stage-wise canonical representations that preserve both functional hierarchy and spatial organization. MIND-BIRTH achieves this through coordinate-anchored ROI tokenization and stage-wise visual-semantic alignment. The learned canonical representations are subsequently decoded through hierarchy-matched diffusion conditioning (Zhang et al., 2023), routing Early, Middle, and Late signals to corresponding generator depths.

Extensive experiments on NSD show that MIND-BIRTH achieves strong semantic reconstruction quality under both zero-shot and few-shot cross-subject protocols, outperforming recent baselines such as MindEye2 (Scotti et al., 2024), ZEBRA (Wang et al., 2026), MindAligner (Dai et al., 2025), NeuroPictor (Huo et al., 2024), Mindbridge (Wang et al., 2024) and MindTuner (Gong et al., 2025) on semantic metrics, while remaining competitive on low-level fidelity. The framework supports lightweight personalization with as few as $K = 250$ adaptation images, reducing the need for full per-subject retraining. Beyond performance, stage-wise analyses reveal complementary contributions across Early, Middle, and Late signals, while a capacity-matched Flat baseline isolates the benefit of explicit hierarchy-matched routing. Further component ablations validate coordinate anchoring, canonical alignment, ROI-stage integrity, and hierarchy correctness. These results suggest that the brain's visual hierarchy provides an effective inductive bias for cross-subject visual decoding.

Our contributions are summarized as follows:

- We introduce MIND-BIRTH, a cross-subject visual decoding framework that transforms heterogeneous subject-specific ROI responses into coordinate-anchored tokens and aligns them into shared stage-wise canonical cortical spaces, enabling zero-/few-shot transfer to unseen subjects. The learned representations are decoded through hierarchy-matched diffusion conditioning.
- MIND-BIRTH achieves the strongest semantic reconstruction quality among cross-subject methods on NSD under both zero-shot and few-shot protocols, while remaining competitive on low-level fidelity, with substantial gains achievable from limited adaptation images.
- We provide systematic stage-wise, capacity-matched, and component-wise ablations showing complementary cortical-stage contributions and demonstrating that explicit hierarchy-matched routing, coordinate anchoring, canonical alignment, and ROI-stage integrity each contribute to cross-subject decoding.

2 RELATED WORK

2.1 fMRI VISUAL DECODING.

Early work on visual decoding from fMRI reconstructed stimuli using explicit priors or handcrafted feature models, showing that natural images and movies can be partially recovered from cortical activity (Naselaris et al., 2009; Nishimoto et al., 2011). With deep neural networks, later methods improved fidelity by predicting hierarchical visual features and inverting them back to image space (Horikawa & Kamitani, 2017; Shen et al., 2019). More recently, diffusion-based pipelines substantially advanced perceptual quality and semantic faithfulness by coupling brain-to-latent mappings with strong generative priors (Takagi & Nishimoto, 2023; Ozcelik & VanRullen, 2023; Scotti et al., 2023; Chen et al., 2023a; Li et al., 2025), and large-scale datasets such as NSD enable systematic evaluation at scale (Allen et al., 2022). Despite these advances, many systems remain within-subject, and cross-subject transfer is still unreliable due to inter-subject variability in functional topography and representational geometry. Recent efforts explicitly address cross-subject decoding (Wang et al., 2024; Scotti et al., 2024; Gong et al., 2025), while MindAligner (Dai et al., 2025) proposes an explicit transfer matrix to map a novel subject into a known subject space and

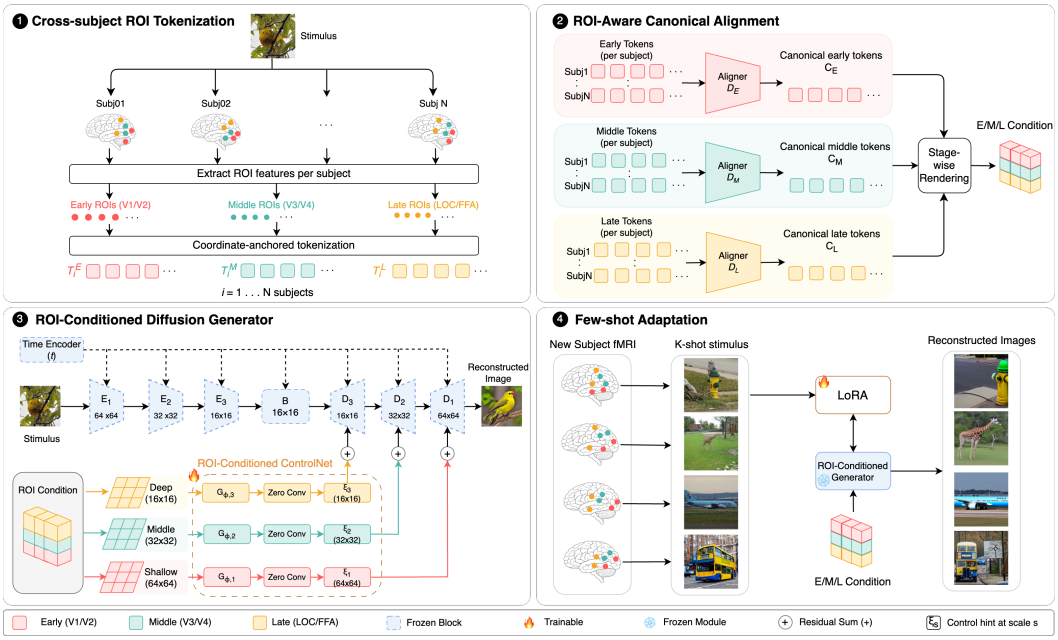


Figure 1: **Pipeline of MIND-BIRTH cross-subject fMRI-to-image model**. We tokenize early/mid/late ROI signals into coordinate-anchored tokens, align them into stage-wise canonical spaces, and render them into stage-specific ROI conditions for hierarchy-matched diffusion generation. For a new subject, we perform lightweight K-shot adaptation by updating only subject-specific calibration and LoRA parameters while keeping all shared components frozen.

reuse pretrained decoders under limited fMRI data. However, existing cross-subject approaches are often largely data-driven: they typically pursue a global subject-invariant latent space or a single transfer mapping without explicitly decomposing visual cortex into functionally distinct stages, which can blur ROI-specific contributions and limit interpretability; moreover, the learned alignment can be brittle when the held-out subject differs in ROI coverage or noise characteristics.

2.2 BRAIN VISUAL CORTICAL SYSTEM

The primate visual cortex is organized around two complementary principles: hierarchy and topography. Hierarchically, visual information is transformed through a cascade of cortical stages (Bandettini, 2009; Baillet, 2017; Cohen, 2017), where early areas (e.g., V1/V2) encode local, low-level structure such as edges and orientation, mid-level regions (e.g., V3/hV4) integrate larger-scale form and surface properties, and higher-level ventral-stream areas support increasingly abstract and category-selective representations (e.g., object- and face-related responses) (Felleman & Van Essen, 1991; DiCarlo et al., 2012; Grill-Spector & Malach, 2004). In parallel, early visual cortex is strongly topographically mapped: nearby cortical locations correspond to nearby locations in visual space (retinotopy), and population receptive field models provide a principled way to describe this spatial correspondence (Wandell & Winawer, 2015; Benson et al., 2018; Benson & Winawer, 2018). Motivated by this hierarchical and topographic organization, we encode neural responses as stage-wise canonical tokens (Early/Middle/Late) with coordinate anchors, enabling the generator to leverage both stage-specific content and spatial structure.

3 METHOD

3.1 MOTIVATION AND OVERVIEW

Motivated by limitations in previous works, we design a decoding framework (Figure 1) that explicitly respects the visual hierarchy and topographic organization. We inject stage-specific ROI signals into

different depths of the generator: Early ROIs (e.g., V1/V2) guide low-level structure, Middle ROIs (e.g., V3/hV4) guide intermediate structure and appearance, while Late ROIs (e.g., LOC/FFA) impose semantic constraints. This inductive bias improves cross-subject transfer by preserving stage-wise content and spatial organization, instead of overfitting to subject-specific patterns.

Let $s \in \{1, \ldots, S\}$ index subjects and $i \in \{1, \ldots, N\}$ index stimulus trials. For subject s, we denote the single-trial fMRI response (beta estimate) as $\mathbf{b}_i^{(s)} \in \mathbb{R}^{V_s}$, where V_s is the number of voxels/vertices used for that subject. The corresponding ground-truth stimulus image is $\mathbf{x}_i$. Our goal is cross-subject neural image generation: learn a shared model that reconstructs $\mathbf{x}_i$ from $\mathbf{b}_i^{(s)}$ and generalizes to unseen subjects with minimal calibration.

We exploit the hierarchical organization of the visual system by partitioning neural signals into three stages $r \in \{\mathrm{E}, \mathrm{M}, \mathrm{L}\}$ (Early/Mid/Late). The method consists of four components: (i) ROI tokenization with coordinate anchors (Sec. 3.2), (ii) structure-aware canonical alignment that maps subject-specific tokens into stage-wise canonical spaces and regularizes them with shared visual-semantic targets (Sec. 3.3), (iii) a shared diffusion generator conditioned on the canonical tokens (Sec. 3.4), and (iv) parameter-efficient calibration for new subjects with task-driven objectives (Sec. 3.5).

3.2 ROI to Coordinate-Anchored Tokens

For each subject s and trial i, we first construct three subject-specific 2D stage activation maps from the beta response $b_i^{(s)}$. These maps are obtained directly from the subject-native 1.8-mm isotropic functional space rather than from cortical-surface projection. Early contains V1/V2, Middle contains V3/hV4, and Late contains FFA/LOC.

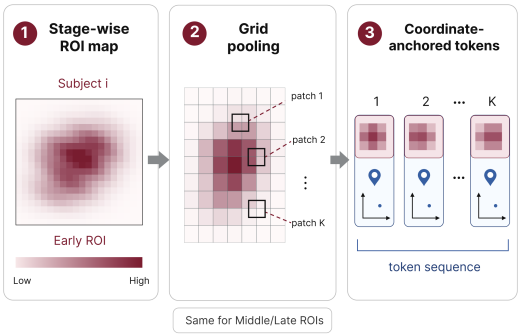


Figure 2: **Coordinate-anchored ROI tokenization.** Stage-wise ROI maps are spatially pooled into local response features, which are combined with 2D coordinate anchors to form ROI token sequences for alignment and decoding.

For each stage, the corresponding ROI responses are standardized within the stage mask and projected along the z-axis to obtain a 2D activation map. We retain the full projected x–y canvas without subject-specific bounding-box cropping or ROI rearrangement, and resize each stage map to 256×256. This standardizes map dimensions while preserving subject-specific spatial organization.

We then tokenize each stage map into a fixed number of coordinate-anchored tokens. For stage $r \in \{E, M, L\}$, the corresponding map is divided into M_r regular patches by adaptive average pooling. Each patch is summarized by an aggregator g_r and augmented with a coordinate anchor $c_{r,j}$ in the normalized 2D condition-map frame:

$$\tilde{t}_{i,r,j}^{(s)} = \mathrm{LN}\left(g_r(a_{i,r,j}^{(s)}) + W_c \gamma(c_{r,j})\right), \tag{1}$$

where $a_{i,r,j}^{(s)}$ is the pooled response of the j-th patch, W_c projects the coordinate feature to the token dimension, and $\gamma(\cdot)$ denotes Fourier positional features:

$$\gamma(c) = \left[x, y, \{\sin(2^k \pi x), \cos(2^k \pi x), \sin(2^k \pi y), \cos(2^k \pi y)\}_{k=0}^{B-1}\right], \tag{2}$$

with $c = (x, y) \in [-1, 1]^2$. The final stage-wise token set is

$$T_{i,r}^{(s)} = [\tilde{t}_{i,r,1}^{(s)}, \ldots, \tilde{t}_{i,r,M_r}^{(s)}] \in \mathbb{R}^{M_r \times d}. \tag{3}$$

In the current implementation, coordinate anchors are normalized 2D map coordinates rather than explicit pRF estimates, visual-field coordinates, or 3D MNI coordinates. They provide a lightweight

positional indexing structure, while the subsequent stage-wise canonical alignment maps subject-specific token representations into shared latent spaces. Fig. 2 illustrates this ROI-to-token conversion using the Early visual ROI group as an example; the same procedure is applied to Middle and Late ROIs. Implementation details, including token counts, grid sizes, and Fourier frequency bands, are provided in Appendix A.1.

3.3 ROI-Aware Canonical Alignment

Although Sec. 3.2 maps subject-specific ROI responses into a common token format, the resulting tokens may still contain subject-dependent response patterns. We therefore introduce a stage-specific canonical alignment module to map ROI tokens into a shared latent space before diffusion conditioning. For each visual stage $r \in \{E, M, L\}$, given the token set $T_{i,r}^{(s)} \in \mathbb{R}^{M_r \times d}$, we compute

$$C_{i,r}^{(s)} = A_r(T_{i,r}^{(s)}) \in \mathbb{R}^{M_r \times d}, \tag{4}$$

where A_r is a lightweight stage-specific token mapper that preserves both token number and token dimension.

To encourage the canonical tokens to capture visually meaningful information, we use the stimulus image paired with each fMRI trial as weak visual supervision. A frozen CLIP image encoder extracts a global image feature from the ground-truth image x_i, which is mapped by a stage-specific projection head:

$$z_{i,r} = P_r(f_{\mathrm{CLIP}}(x_i)) \in \mathbb{R}^d. \tag{5}$$

The CLIP encoder is fixed, while the projection heads and alignment modules are trained jointly. We use separate projection heads for Early, Middle, and Late stages, allowing each ROI group to learn a distinct visual target from the same frozen image representation.

We align the pooled canonical token representation with the corresponding CLIP-derived target:

$$\bar{C}_{i,r}^{(s)} = \frac{1}{M_r} \sum_{j=1}^{M_r} C_{i,r,j}^{(s)}, \qquad \mathcal{L}_{\mathrm{align}} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \left(1 - \cos(\bar{C}_{i,r}^{(s)}, z_{i,r})\right), \tag{6}$$

where $\mathcal{R} = \{E, M, L\}$. The final training objective combines the diffusion reconstruction loss with the alignment loss and a small token regularizer:

$$\mathcal{L}_{\mathrm{train}} = \mathcal{L}_{\mathrm{diff}} + \lambda_{\mathrm{align}} \mathcal{L}_{\mathrm{align}} + \lambda_{\mathrm{reg}} \mathcal{L}_{\mathrm{reg}}. \tag{7}$$

Implementation details of A_r, P_r, and $\mathcal{L}_{\mathrm{reg}}$ are provided in Appendix A.2.

3.4 ROI-Conditioned Diffusion Generation

After canonical alignment, the Early, Middle, and Late token sets are independently rendered into stage-specific spatial conditions. For each stage $r \in \{E, M, L\}$, a lightweight token-to-map readout ϕ_r projects each canonical token to a scalar conditioning value:

$$u_{i,r,j}^{(s)} = \phi_r\left(C_{i,r,j}^{(s)}\right). \tag{8}$$

The projected values are reshaped according to the corresponding token grid and resized to the spatial resolution of the matched denoiser layer:

$$X_{i,r}^{(s)} = \mathrm{Resize}_r\left(\mathrm{Reshape}\left(\{u_{i,r,j}^{(s)}\}_{j=1}^{M_r}\right)\right). \tag{9}$$

The three cortical stages remain separate throughout diffusion conditioning. Specifically, Early, Middle, and Late conditions are routed to decoder blocks D_1, D_2, and D_3 at resolutions 64×64, 32×32, and 16×16, respectively. Each stage is processed by an independent lightweight conditioning branch and injected only into its matched denoiser depth.

Thus, the learned canonical representations are decoded through fixed hierarchy-matched conditioning, preserving the correspondence between Early/Middle/Late cortical stages and shallow/middle/deep generator representations. The same routing is maintained throughout training and inference. Detailed conditioning and residual-modulation implementations are provided in Appendix A.3.

3.5 Few-Shot Subject Adaptation

For a held-out subject, we reuse the cross-subject model learned from the source subjects (Secs. 3.2-3.4) and adapt only a small set of subject-specific parameters. All shared components are frozen, including the diffusion backbone, ROI tokenizer, canonical alignment module, and token-to-map module.

For each stage $r \in \{E, M, L\}$, we introduce a lightweight affine calibrator on the canonical tokens:

$$C'_{i,r} = \gamma_r C_{i,r} + \beta_r, \tag{10}$$

where Early, Middle, and Late use separate calibration parameters. We additionally insert LoRA adapters into the linear projections of the stage-specific conditioning branches. During few-shot adaptation, only the calibration and LoRA parameters are updated.

Given K target-subject training pairs outside the evaluation split, this lightweight adaptation adjusts the shared canonical representation to the held-out subject without full-model fine-tuning. In our main few-shot experiments, we use $K = 250$. Detailed LoRA configuration and initialization are provided in Appendix A.4.

4 Experiment

4.1 Experimental Setup

Table 1: **Quantitative comparison on NSD shared1000.** MIND-BIRTH is evaluated under zero-shot, $K = 250$ few-shot adaptation, and fully fine-tuned settings. All MIND-BIRTH results are averaged over four held-out subjects (1,2,5,7), with per-subject results and standard deviations in Appendix B.

Methods	Low-Level				High-Level			
	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
Few-shot (K = 250)								
MindEye2	.195	.419	84.2%	90.6%	81.2%	79.2%	.810	.468
MindAligner	.206	.414	85.6%	91.6%	83.0%	81.2%	.802	.463
MindBridge	.112	.229	79.6%	89.0%	82.3%	86.7%	.840	.521
MindTuner	.224	**.420**	87.8%	93.6%	84.8%	83.5%	.780	.440
MIND-BIRTH	**.266**	.387	**91.4%**	**95.3%**	**91.6%**	**92.1%**	**.616**	**.382**
Zero-shot (K = 0)								
NeuroPictor	.057	.297	71.4%	74.7%	62.5%	66.0%	.939	.607
ZEBRA	**.131**	**.375**	74.6%	81.2%	72.2%	71.5%	.837	.506
MIND-BIRTH	.118	.336	**78.2%**	**85.6%**	**80.4%**	**79.5%**	**.717**	**.429**
Fully Fine-tuned								
MindEye1	.319	.360	92.8%	96.9%	94.6%	93.3%	.648	.377
MindBridge	.151	.263	87.7%	95.5%	92.4%	94.7%	.712	.413
MindEye2	**.322**	**.431**	**96.1%**	98.6%	95.4%	94.5%	.619	.344
MindTuner	**.322**	.421	95.8%	**98.8%**	95.6%	93.8%	.612	.340
MIND-BIRTH	.315	.418	94.8%	96.9%	**98.4%**	**98.6%**	**.526**	**.324**

Dataset. We use the Natural Scenes Dataset (NSD) (Allen et al., 2022) for training and evaluation. Following prior cross-subject decoding studies (Wang et al., 2024), we report results on subjects 1, 2, 5, and 7 and use the NSD shared1000 images as the common evaluation set.

For each held-out subject, we adopt a leave-one-subject-out (LOSO) protocol in which the model is trained on the remaining source subjects and evaluated on the unseen target subject. To prevent stimulus-level leakage, the shared1000 evaluation image identities and all corresponding fMRI trials are excluded from every source subject during training. Thus, in the zero-shot setting, the model observes neither target-subject responses nor source-subject responses associated with the evaluation images.

Evaluation Protocol and Metrics. We evaluate reconstructions using eight metrics spanning low- and high-level similarity. Low-level metrics include PixCorr, SSIM (Wang et al., 2004), and Alex(2)/Alex(5) (Krizhevsky et al., 2012) feature similarity from a pretrained AlexNet, while high-level metrics include Incep (Szegedy et al., 2016) and CLIP (Radford et al., 2021) similarity as well as Eff (Tan & Le, 2019) and SwAV (Caron et al., 2020) feature-space distances.

We evaluate three settings. In **zero-shot**, no target-subject samples are used for training or adaptation. In **few-shot**, we initialize from the LOSO checkpoint and update only the calibration and LoRA parameters using $K = 250$ target-subject training images outside the shared1000 evaluation split. In **fully fine-tuned**, we initialize from the same LOSO checkpoint and adapt the model using the complete target-subject training split, while evaluation remains on the held-out target test set.

4.2 Main Results

Quantitative Results. Table 1 reports results under zero-shot, $K = 250$ few-shot adaptation, and fully fine-tuned settings. In the zero-shot regime, MIND-BIRTH achieves the strongest semantic performance among compared cross-subject baselines, improving CLIP from 71.5% to 79.5% and Inception from 72.2% to 80.4% over ZEBRA, while also improving Alex(2)/Alex(5). With only $K = 250$ target-subject samples, MIND-BIRTH further improves high-level reconstruction quality and reaches 91.6% Inception and 92.1% CLIP, outperforming prior few-shot methods on semantic metrics. Fully fine-tuned results are included as an upper-reference setting, where MIND-BIRTH remains highly competitive and achieves the best high-level scores, supporting that hierarchical ROI conditioning mainly strengthens semantic alignment and cross-subject generalization.

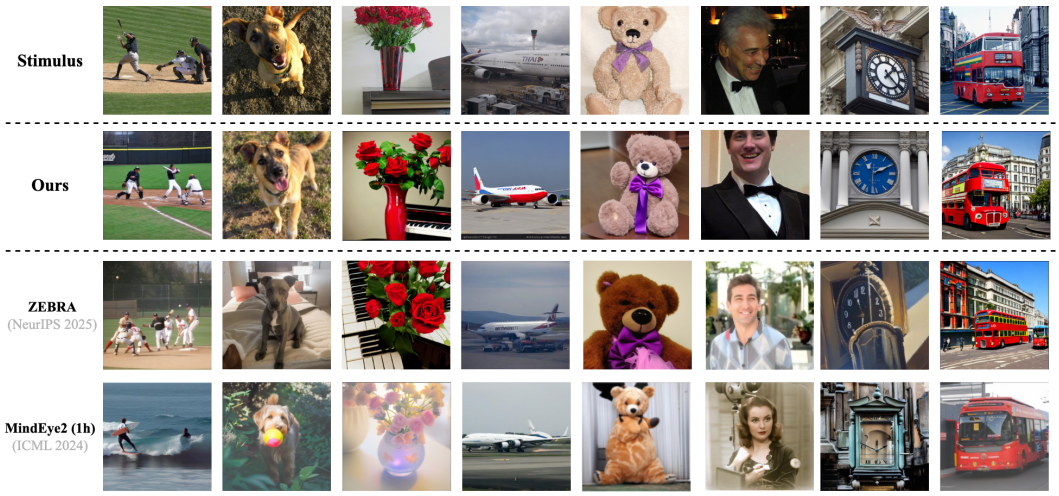


Figure 3: **Qualitative comparison on NSD shared1000.** MIND-BIRTH performs zero-shot reconstruction and is compared with ZEBRA in the zero-shot setting and MindEye2 with 1-hour subject adaptation.

Qualitative Results. Figure 3 compares our zero-shot reconstructions with ZEBRA under the zero-shot setting and MindEye2 with 1-hour subject adaptation. Despite using no target-subject samples, MIND-BIRTH more consistently preserves the semantic identity and global layout of the stimuli across diverse categories, including animals, objects, faces, clocks, and vehicles. Compared with baselines, our reconstructions exhibit fewer category-level errors and better maintain salient attributes such as color, shape, and scene structure.

4.3 Ablation Study

Stage-Wise Contributions. To isolate the contribution of each cortical stage, we evaluate Early (E), Middle (M), and Late (L) individually, in all pairwise combinations, and jointly under the same zero-shot LOSO setting, varying only the available cortical stages. Results are summarized in Table 2.

The three stages exhibit distinct and complementary roles: Early contributes most strongly to spatial layout and low-level structure, Middle provides an intermediate structural–semantic balance, and Late contributes stronger semantic cues. Pairwise combinations generally improve this balance over single-stage variants. Notably, Late-only remains below the full hierarchy, showing that Early and Middle provide complementary information despite the shared frozen CLIP supervision. The full model performs best on seven of eight metrics, while Early-only achieves a marginally higher SSIM, consistent with its stronger emphasis on local structure.

Table 2: Stage-wise contribution analysis under zero-shot LOSO.

ROI	PixCorr↑	SSIM↑	Incep↑	CLIP↑
E	.094	**.342**	70.8%	69.6%
M	.077	.301	74.9%	74.1%
L	.059	.270	76.3%	75.7%
E+M	.112	.339	77.0%	76.3%
E+L	.103	.321	77.8%	77.1%
M+L	.090	.304	79.0%	78.3%
E+M+L	**.118**	.336	**80.4%**	**79.5%**

Capacity-matched non-hierarchical baseline. The stage-wise analysis above shows that Early, Middle, and Late provide complementary information. We next ask whether the gain comes from explicit stage-to-depth correspondence, rather than simply providing all three stages to every denoiser depth. Results can be found in Table 3.

Table 3: Capacity-matched comparison between Flat and hierarchical routing under zero-shot LOSO.

Routing	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
Flat	.096	.313	73.2%	83.1%	77.6%	77.2%	.738	.475
Hierarchical	**.118**	**.336**	**78.2%**	**85.6%**	**80.4%**	**79.5%**	**.717**	**.429**

We construct a capacity-matched **Flat** baseline using the same E/M/L inputs, tokenizer, canonical alignment, conditioning heads, diffusion backbone, losses, and trainable parameter count. The only difference is routing: instead of Early$\rightarrow D_1$, Middle$\rightarrow D_2$, and Late$\rightarrow D_3$, Flat averages the three resized stage conditions and supplies the same fused condition to all three depths. Both models are independently trained under identical zero-shot LOSO settings.

Despite using identical neural information and model capacity, the hierarchical model outperforms Flat across all eight metrics. This isolates the benefit of explicit stage-to-depth correspondence from additional conditioning information or parameter count.

Table 4: **Unified ablation studies on MIND-BIRTH components.** We systematically verify four core design choices, each corresponding to one component of our method: ROI signal integrity (**C1**, Sec. 3.2), coordinate-anchored tokenization (**C2**, Sec. 3.2), canonical alignment loss (**C3**, Sec. 3.3), and hierarchical depth-aware injection (**C4**, Sec. 3.4). All variants are evaluated under LOSO zero-shot protocol on NSD shared1000, averaged over held-out subjects. **Full (Ours)** is shown as reference; degradation under each variant indicates the contribution of the corresponding component.

		Low-Level				High-Level			
Component	**Variant**	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
Full (Ours)		**.118**	**.336**	**78.2%**	**85.6%**	**80.4%**	**79.5%**	**.717**	**.429**
ROI Signal Integrity (C1)	Voxel shuffle	.085	.291	70.6%	76.8%	72.2%	71.6%	.821	.574
	Cross-ROI swap	.065	.258	62.8%	66.2%	65.8%	63.9%	.910	.664
Coordinate Anchor (C2)	w/o PE	.108	.322	76.6%	84.2%	78.5%	77.0%	.754	.463
	Random coords	.098	.310	75.1%	82.7%	76.8%	75.4%	.777	.483
Alignment Loss (C3)	w/o $\mathcal{L}_{\text{align}}$	.087	.298	71.3%	78.4%	73.9%	73.2%	.805	.555
	Shuffled target	.081	.285	69.0%	75.8%	70.8%	70.1%	.839	.589
Hierarchy Mapping (C4)	Random	.094	.317	73.5%	81.7%	77.4%	76.6%	.747	.484
	Reversed	.073	.274	65.9%	69.6%	68.2%	66.0%	.881	.638

Core Component Ablations. Table 4 evaluates the core design choices of MIND-BIRTH under the same zero-shot LOSO protocol. 'Voxel shuffle' and 'Cross-ROI swap' perturb the spatial organization and ROI-stage identity of the cortical inputs, respectively, and both reduce reconstruction quality, with Cross-ROI swap producing the largest drop. Removing positional encoding ('w/o PE') or replacing anatomical coordinates with 'Random coords' also degrades performance, confirming the importance of consistent spatial anchoring. Similarly, removing the alignment objective ('w/o L_{align}') or using a 'Shuffled target' weakens both low- and high-level reconstruction, showing that cross-subject canonical alignment depends on meaningful visual-semantic supervision.

Finally, 'Random' and 'Reversed' hierarchy controls test whether the stage-to-depth correspondence itself matters. Random routing reduces performance by disrupting the systematic association between cortical stages and denoiser depths, while Reversed routing causes a larger degradation by injecting otherwise correctly encoded stage representations into hierarchically mismatched depths. Together, these results show that MIND-BIRTH relies on a coherent processing chain-from ROI-specific signals and spatial anchoring, through stage-wise canonical alignment, to hierarchy-matched diffusion conditioning. The gains therefore arise from the complementary contribution of these components rather than from any single module or additional model capacity.

4.4 DISCUSSION

Our results show that cross-subject decoding benefits from preserving cortical hierarchy rather than collapsing neural responses into a single representation(Scotti et al., 2024; Wang et al., 2024; Dai et al., 2025). Under strict LOSO evaluation, MIND-BIRTH generalizes to unseen subjects and supports lightweight few-shot personalization. Stage-wise ablations reveal complementary roles across Early, Middle, and Late signals, while the capacity-matched Flat baseline confirms that explicit stage-to-depth routing provides gains beyond model capacity or conditioning information. Together with the core component ablations, these results support the importance of maintaining consistent ROI identity, spatial anchoring, canonical alignment, and hierarchy-matched conditioning throughout the decoding pipeline. More broadly, they suggest that structuring cross-subject representation learning around known functional organization can provide a stronger inductive bias than relying on a single globally aligned latent space.

Limitations. A key limitation is that our coordinate anchors are defined in a normalized 2D condition-map frame rather than explicit cortical-surface coordinates, visual-field coordinates, or pRF estimates, and therefore only approximate the topographic organization of visual cortex. Future work could incorporate surface-based anchors, retinotopic maps, or pRF-derived coordinates to provide a more biologically grounded token space, particularly for early visual areas. In addition, our ROI grouping is predefined and relatively coarse; learned, soft, or subject-adaptive ROI grouping may better capture individual functional organization. Although the full hierarchy achieves the strongest overall performance, Early-only conditioning yields slightly higher SSIM, suggesting that improving low-level structural fidelity without sacrificing semantic accuracy remains an open challenge. Finally, our current evaluation is limited to four NSD subjects acquired under a common experimental protocol, and broader validation across subjects, scanners, acquisition settings, and independent datasets is needed to assess the robustness and generality of the learned cross-subject representations.

5 CONCLUSION

We presented MIND-BIRTH, a framework for learning hierarchical canonical cortical representations for cross-subject fMRI-to-image reconstruction. By combining coordinate-anchored ROI tokenization, stage-wise canonical alignment, and hierarchy-matched diffusion conditioning, MIND-BIRTH transfers structured cortical representations across subjects and supports both zero-shot decoding and lightweight few-shot personalization. Experiments on NSD demonstrate strong cross-subject reconstruction performance, while stage-wise analyses reveal complementary structural and semantic roles across Early, Middle, and Late cortical signals. Capacity-matched and component-wise ablations further show that the gains arise from preserving cortical-stage structure, spatial anchoring, cross-subject alignment, and matched-depth conditioning rather than from additional model capacity alone. Overall, our results suggest that explicitly modeling hierarchical cortical organization provides an effective inductive bias for transferable visual decoding from human brain activity.

AI USE STATEMENT

In this work, we used generative AI tools for no tasks requiring mandatory disclosure. We have not used generative AI tools for methodological design, experimental procedures, data processing, quantitative result generation, or interpretation of experimental findings, and the remaining tasks requiring disclosure are not applicable to this work. Additionally, we used generative AI tools for writing and editing to improve clarity, grammar, and presentation, as well as for reviewing and organizing previously published literature. We have reviewed all AI-assisted work. All AI-assisted text was manually checked and revised by the authors, and all cited literature was verified against the original sources. 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 provide the methodological and experimental details required to reproduce our results throughout the main paper and appendix. In addition, we provide anonymous source code as supplementary material to support implementation-level 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.

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 and Jonathan Winawer. Bayesian analysis of retinotopic maps. *elife*, 7:e40224, 2018.

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.

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.

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

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

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.

Gary H Glover. Deconvolution of impulse response in event-related bold fmri1. *Neuroimage*, 9(4): 416–429, 1999.

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.

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.

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

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

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.

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.

Nikos K Logothetis, Jon Pauls, Mark Augath, Torsten Trinath, and Axel Oeltermann. Neurophysiological investigation of the basis of the fmri signal. *nature*, 412(6843):150–157, 2001.

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.

Seiji Ogawa, Tso-Ming Lee, Asha S Nayak, and Paul Glynn. Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magnetic resonance in medicine*, 14(1):68–78, 1990.

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

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 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.

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.

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

Haonan Wang, Jingyu Lu, Hongrui Li, and Xiaomeng Li. Zebra: Towards zero-shot cross-subject generalization for universal brain visual decoding. *Advances in Neural Information Processing Systems*, 38:1030–1051, 2026.

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.

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 (ICCV)*, pp. 3836–3847, October 2023.

APPENDIX

A IMPLEMENTATION DETAILS

We implement MIND-BIRTH in PyTorch and train on 2× NVIDIA A100 GPUs (80GB). We use NSD GLMdenoise beta estimates in the func1pt8mm space. Stage-wise ROI signals are converted into coordinate-anchored token sets and mapped into canonical token spaces. The resulting Early, Middle, and Late canonical representations are independently rendered into stage-specific spatial conditions and processed by separate conditioning branches before injection into their matched denoiser depths.

Our diffusion backbone is Stable Diffusion v1.5, with the VAE, U-Net, and CLIP text encoder frozen. The stage-specific conditioning branches are trained jointly with the ROI representation modules. Optimization is performed with AdamW under mixed precision.

Unless otherwise stated, each LOSO model is trained for 18,000 steps with batch size 4. During few-shot adaptation, all shared components remain frozen and only the subject-specific calibration and LoRA parameters are updated; full details are provided in Appendix A.4. At inference time, we use 50 DDIM denoising steps, guidance scale 7.5, and ControlNet conditioning scale 1.0.

A.1 IMPLEMENTATION DETAILS OF ROI MAP CONSTRUCTION AND TOKENIZATION

Although $b_i^{(s)}$ is written as a vector in the main text for notational convenience, the implementation places the voxel responses back into their native func1pt8mm 3D volumetric grid using the corresponding voxel indices and ROI masks.

For each stage $r \in \{E, M, L\}$, let $M_r^{(s)}(v)$ denote the union mask of its constituent ROIs. We compute the mean $\mu_{i,r}^{(s)}$ and standard deviation $\sigma_{i,r}^{(s)}$ over voxels inside the stage mask and form the standardized activation volume:

$$z_{i,r}^{(s)}(v) = M_r^{(s)}(v)\frac{b_i^{(s)}(v) - \mu_{i,r}^{(s)}}{\sigma_{i,r}^{(s)} + 10^{-6}}. \tag{1}$$

Responses outside the stage mask are set to zero. We then obtain the 2D stage activation map by maximum-intensity projection along the z-axis:

$$p_{i,r}^{(s)}(x, y) = \max_z z_{i,r}^{(s)}(x, y, z). \tag{2}$$

Within the projected ROI support, we scale the map using the 1st and 99th percentiles, clip values to $[0, 1]$, set values outside the support to zero, and apply light Gaussian smoothing. We do not perform subject-specific bounding-box cropping or rearrange individual ROIs; therefore, spatially separated ROIs remain disconnected components at their relative locations in the projected map. Each stage map is finally resized to 256×256.

For tokenization (Sec. 3.2), we apply adaptive average pooling over fixed regular grids. Early, Middle, and Late use grids of 16×8, 8×8, and 8×4, producing $M_E = 128$, $M_M = 64$, and $M_L = 32$ tokens, respectively. Each pooled scalar is mapped to a $d = 64$ dimensional feature using a stage-specific two-layer MLP with GELU activation. Each coordinate anchor is encoded using Fourier positional features with $B = 8$ frequency bands.

A.2 IMPLEMENTATION DETAILS OF ROI-AWARE CANONICAL ALIGNMENT.

In Sec. 3.3, for each ROI stage $r \in \{E, M, L\}$, the canonical mapper A_r is implemented as a residual MLP with layer normalization:

$$A_r(T) = \mathrm{LN}\left(T + \mathrm{MLP}_r(T)\right). \tag{3}$$

It preserves the token shape, mapping $T_{i,r}^{(s)} \in \mathbb{R}^{M_r \times 64}$ to $C_{i,r}^{(s)} \in \mathbb{R}^{M_r \times 64}$, with $M_E = 128$, $M_M = 64$, and $M_L = 32$.

For visual supervision, we use a frozen CLIP image encoder f_{CLIP} to extract one global image representation for each ground-truth stimulus. Three trainable stage-specific projection heads P_E,

P_M, and P_L map this frozen CLIP feature into 64-dimensional targets for Early, Middle, and Late stages:

$$z_{i,r} = P_r(f_{\text{CLIP}}(x_i)) \in \mathbb{R}^{64}. \tag{4}$$

The same CLIP encoder and projection-head design are used across all experiments. We do not use different CLIP layers for different ROI stages; stage specificity is introduced through the three projection heads.

The regularization term is a token-level ℓ_2 penalty on the canonical tokens:

$$\mathcal{L}_{\text{reg}} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \frac{1}{M_r} \sum_{j=1}^{M_r} \|C_{i,r,j}^{(s)}\|_2^2. \tag{5}$$

Unless otherwise specified, we use $\lambda_{\text{align}} = 0.1$ and $\lambda_{\text{reg}} = 0.01$ (in Equation. 7) in all experiments. We use a cosine alignment loss rather than a temperature-scaled contrastive objective; therefore no temperature parameter τ is introduced.

A.3 Token-to-Map Rendering and Stage-Specific Diffusion Conditioning

The token-to-map rendering module follows the fixed token grids defined in Sec. 3.2. Early, Middle, and Late canonical token sets have shapes $C_E \in \mathbb{R}^{128\times64}$, $C_M \in \mathbb{R}^{64\times64}$, and $C_L \in \mathbb{R}^{32\times64}$, corresponding to grids of size 16×8, 8×8, and 8×4, respectively.

For each stage $r \in \{E, M, L\}$, a stage-specific token-wise readout ϕ_r projects each 64-dimensional canonical token to a scalar conditioning value:

$$u_{i,r,j}^{(s)} = \phi_r\left(C_{i,r,j}^{(s)}\right). \tag{6}$$

These values are reshaped according to the original stage-specific token grid and resized to the resolution of the matched U-Net decoder block:

$$X_{i,r}^{(s)} = \text{Resize}_r\left(\text{Reshape}\left(\{u_{i,r,j}^{(s)}\}_{j=1}^{M_r}\right)\right). \tag{7}$$

The routing is fixed as follows:

Cortical stage	Condition resolution	Matched denoiser layer
Early	64×64	Decoder block D_1
Middle	32×32	Decoder block D_2
Late	16×16	Decoder block D_3

Importantly, the three stages remain separate conditioning streams after token-to-map rendering. Each stage is processed by an independent conditioning head followed by a zero-initialized convolution. For stage r, the resulting control feature is

$$c_{i,r}^{(s)} = G_r\left(X_{i,r}^{(s)}\right), \tag{8}$$

where G_r denotes the stage-specific conditioning head.

At the corresponding denoiser depth, residual and FiLM conditioning are jointly applied as

$$h_r' = h_r + \lambda_r \left[A_r(c_{i,r}^{(s)}) + \gamma_r(c_{i,r}^{(s)}) \odot h_r + \beta_r(c_{i,r}^{(s)})\right], \tag{9}$$

where h_r and h_r' denote the original and conditioned U-Net features, respectively. A_r is a 1×1 residual projection, γ_r and β_r are channel-wise FiLM scale and bias terms, $\odot$ denotes element-wise multiplication, and λ_r controls the conditioning strength.

The stage-to-depth assignment is fixed throughout both training and inference. Therefore, Early information only conditions D_1, Middle information only conditions D_2, and Late information only conditions D_3. This prevents premature mixing of the three cortical stages before depth-wise injection while allowing each conditioning branch to adapt the shared canonical representation to the feature statistics of its matched denoiser layer.

A.4 Few-Shot Adaptation Details

In Sec. 3.5, few-shot adaptation starts from the source-subject LOSO checkpoint. All shared model components are frozen, including the diffusion backbone, ROI tokenizer, canonical alignment module, token-to-map readout, and the original conditioning weights.

For each stage $r \in \{E, M, L\}$, the canonical tokens are calibrated by

$$C'_{i,r} = \gamma_r C_{i,r} + \beta_r, \tag{10}$$

where γ_r and β_r are stage-specific learnable parameters. We initialize $\gamma_r = 1$ and $\beta_r = 0$, so that adaptation initially preserves the source-trained representation.

LoRA adapters are inserted into the linear projections of the stage-specific conditioning branches before injection into the matched denoiser layers. We use rank 4, scaling coefficient 4, and no dropout. Only the calibration and LoRA parameters are optimized during adaptation.

For the main few-shot setting, we use $K = 250$ unique target-subject training images. All shared1000 evaluation images are excluded from adaptation. The same training objective used for source training is retained, while all remaining model parameters stay frozen.

B Per Subject Result

Tables A1 and A2 provide subject-level results for low-level and high-level metrics. The zero-shot setting shows larger inter-subject variability, reflecting the difficulty of transferring to unseen subjects without target-subject calibration. With $K = 250$ few-shot adaptation, performance improves consistently across all four subjects, especially on semantic metrics such as Inception and CLIP. Fully fine-tuned results achieve the strongest overall performance, while the remaining variation across subjects suggests that individual differences in fMRI signal quality and ROI coverage still affect reconstruction quality.

Table A1: Low level results for each test subject.

Subject	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑
Few-shot (K = 250)				
Subject 1	.247	.366	89.2%	93.9%
Subject 2	.271	.392	91.8%	95.6%
Subject 5	.258	.381	90.6%	94.7%
Subject 7	.288	.409	94.0%	97.0%
Mean ± Std	.266 ± .018	.387 ± .018	91.4% ± 2.0	95.3% ± 1.3
Zero-shot (K = 0)				
Subject 1	.101	.319	75.1%	82.7%
Subject 2	.121	.341	78.6%	85.9%
Subject 5	.112	.331	77.0%	84.5%
Subject 7	.138	.353	82.1%	89.3%
Mean ± Std	.118 ± .016	.336 ± .014	78.2% ± 3.0	85.6% ± 2.8
Fully Fine-tuned				
Subject 1	.294	.397	93.0%	95.7%
Subject 2	.319	.423	95.0%	97.0%
Subject 5	.309	.414	94.4%	96.4%
Subject 7	.338	.438	96.8%	98.5%
Mean ± Std	.315 ± .018	.418 ± .017	94.8% ± 1.6	96.9% ± 1.2

Table A2: High level results for each test subject.

Subject	Incep↑	CLIP↑	Eff↓	SwAV↓
Few-shot (K = 250)				
Subject 1	88.9%	89.4%	.672	.451
Subject 2	92.4%	92.6%	.601	.369
Subject 5	90.7%	91.5%	.630	.397
Subject 7	94.4%	94.9%	.561	.311
Mean ± Std	91.6% ± 2.4	92.1% ± 2.3	.616 ± .047	.382 ± .058
Zero-shot (K = 0)				
Subject 1	75.8%	74.6%	.792	.515
Subject 2	81.2%	79.8%	.712	.422
Subject 5	78.7%	77.0%	.750	.458
Subject 7	85.9%	86.6%	.614	.321
Mean ± Std	80.4% ± 4.3	79.5% ± 5.2	.717 ± .076	.429 ± .082
Fully Fine-tuned				
Subject 1	97.6%	97.5%	.558	.362
Subject 2	98.1%	98.8%	.521	.315
Subject 5	98.5%	99.1%	.539	.333
Subject 7	99.4%	99.0%	.486	.286
Mean ± Std	98.4% ± 0.76	98.6% ± 0.74	.526 ± .031	.324 ± .032

C Full Stage-Wise Contribution Results

Table A3 reports the complete eight-metric results for the stage-wise analysis summarized in Table 2 of the main paper. Early (E), Middle (M), and Late (L) cortical stages are evaluated individually, in all pairwise combinations, and jointly under the same zero-shot LOSO protocol.

Table A3: Full stage-wise contribution analysis under the zero-shot LOSO protocol. Early (E), Middle (M), and Late (L) cortical stages are evaluated individually, in pairwise combinations, and jointly.

ROI	**PixCorr↑**	**SSIM↑**	**Alex(2)↑**	**Alex(5)↑**	**Incep↑**	**CLIP↑**	**Eff↓**	**SwAV↓**
E	.094	**.342**	73.6%	80.8%	70.8%	69.6%	.824	.548
M	.077	.301	71.4%	80.2%	74.9%	74.1%	.783	.498
L	.059	.270	66.8%	75.9%	76.3%	75.7%	.759	.472
E+M	.112	.339	77.0%	84.5%	77.0%	76.3%	.754	.458
E+L	.103	.321	75.4%	82.6%	77.8%	77.1%	.746	.450
M+L	.090	.304	73.3%	82.1%	79.0%	78.3%	.731	.440
E+M+L	**.118**	.336	**78.2%**	**85.6%**	**80.4%**	**79.5%**	**.717**	**.429**

D Depth-Wise Probing Analysis

To further examine whether hierarchy-matched routing induces meaningful depth-wise specialization, we probe the residual features extracted from different denoiser depths under the Correct and Reversed hierarchy settings. For each frozen model, we train linear probes to predict three visual targets corresponding to progressively higher-level information: edge structure, part-level representations, and global CLIP image embeddings.

As shown in Figure A1, the Correct hierarchy exhibits a clear diagonal specialization pattern: shallow features are most predictive of edge structure, middle features better capture part-level information, and deep features are most predictive of CLIP semantic representations. In contrast, reversing the Early/Middle/Late-to-depth correspondence disrupts this specialization and reduces



Figure A1: **Depth-wise probing under Correct and Reversed hierarchy.** Linear probes are trained on frozen residual features from shallow, middle, and deep denoiser stages to predict edge representations, part-level visual features, and CLIP image embeddings. Under the Correct hierarchy, the probes exhibit a clear depth-dependent specialization, whereas reversing the cortical-stage-to-depth assignment weakens this structure and reduces overall predictability.

probe performance across the three targets. These results provide complementary mechanistic evidence that the benefit of hierarchical routing is not solely reflected in reconstruction metrics, but also in the organization of representations across denoiser depths.

E Additional Zero-Shot Reconstruction Examples

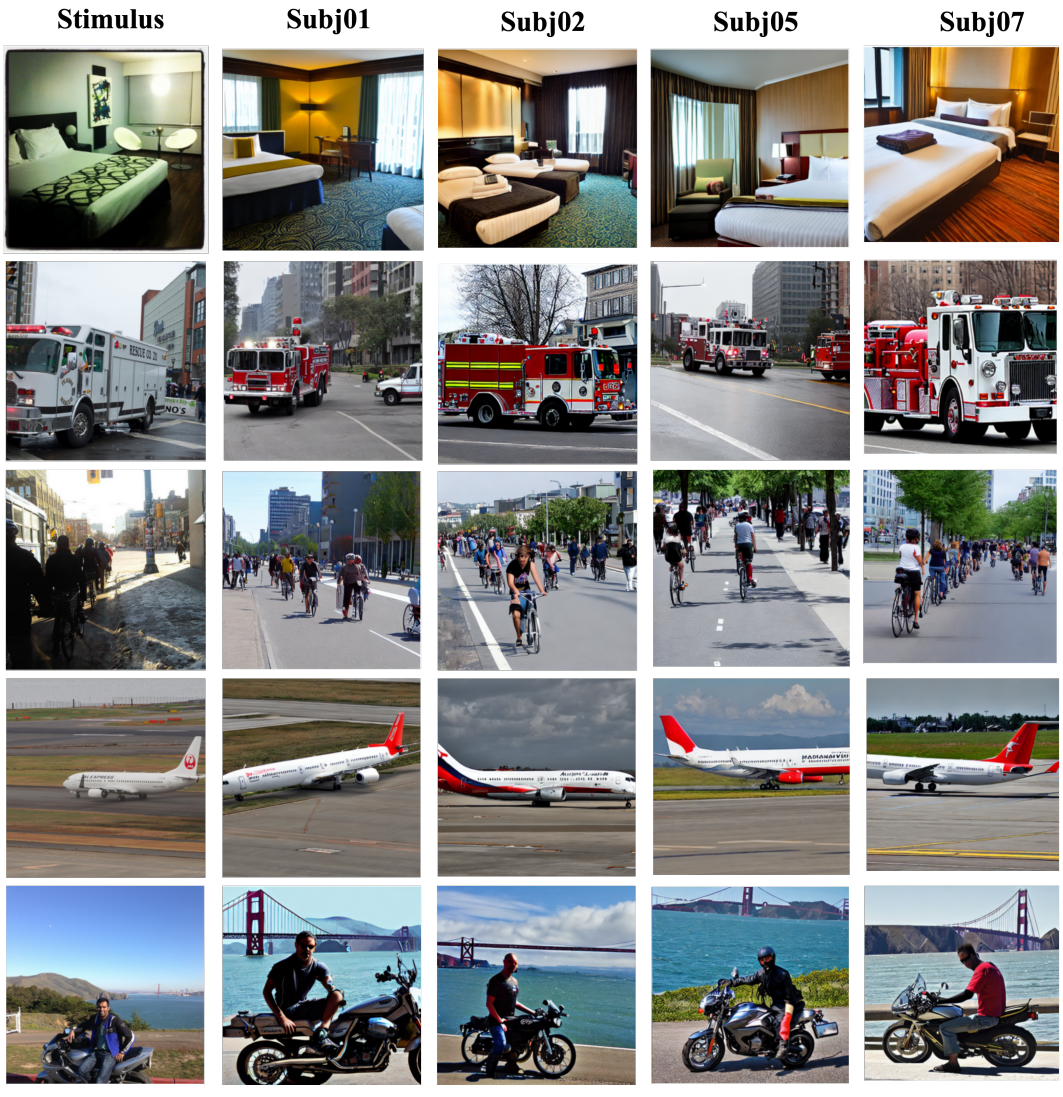


Figure A2: More reconstruction images for different subjects (subj01/02/05/07). Each block shows the stimulus (left) and our reconstructions for each subject (right columns).