MINDLOOP: CLOSED-LOOP BRAIN SUPERVISION FOR FMRI-TO-IMAGE RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Decoding visual experience from brain activity offers a direct window into how the brain represents the visual world. Modern fMRI-to-image reconstruction decodes cortical activity into conditioning representations for large pretrained generators. However, the targets used to learn these representations are derived primarily from stimulus images and generator-specific objectives: trial-specific brain responses serve as inputs but provide little direct supervision.We refer to this imbalance as **supervision asymmetry**: it leaves the decoded representation only indirectly constrained by the measured cortical response, allowing a generator to produce plausible images that is weakly grounded in the underlying brain activity and may therefore fail to reflect the individual's own perceptual experience on a given trial. To address this, we introduce **MindLoop**, which makes the measured cortical response an explicit source of supervision for reconstruction, constraining decoding at three complementary levels: Brain Latent Back-Projection enforces consistency with the same-trial cortical state, Dual-Anchored Relational Alignment preserves stimulus geometry across visual and cortical representations, and Hierarchy-Aware Cortical Fields incorporate retinotopic organization into visual-token readout. On the Natural Scenes Dataset, MindLoop outperforms seven representative methods on seven of eight standard reconstruction metrics, averaged across four subjects. More importantly, brain-response substitution experiments show that reconstruction content tracks the measured cortical response: replacing the brain response with that elicited by a different image collapses identification accuracy to chance and shifts the reconstruction toward the image associated with the substituted response. Cortical fields learned without explicit retinotopic supervision independently recover retinotopic organization, and this organization's horizontal mapping degrades substantially when horizontal-flip augmentation breaks image–cortex correspondence. Together, these results show that balancing supervision from cortical responses and visual stimuli improves reconstruction fidelity, providing a more faithful route toward reconstructing individualized visual experience.

1 INTRODUCTION

Reconstructing natural images from brain functional magnetic resonance imaging (fMRI) activity is a fundamental problem in visual neuroscience and brain-computer interfaces. It turns latent visual experience into measurable image outputs. This capability provides a system-level tool for studying how the brain represents natural scenes and lays a foundation for recovering complex visual content from non-invasive signals (Miyawaki et al., 2008; Naselaris et al., 2011; Allen et al., 2022; Kamitani et al., 2025). Large pretrained generators have transformed this task from direct pixel regression into conditional generation (Guo et al., 2025). Modern systems decode fMRI into visual conditions that guide pretrained generative models to synthesize the final image (Beliy et al., 2026; Yang et al., 2026). MindEye series method predicts CLIP representations and low-level features (Scotti et al., 2023; 2024), NeuroPictor extracts semantic and structural conditions for diffusion generation (Huo et al., 2024), and PRISM converts brain activity into structured visual descriptions (Huang et al., 2026). Despite using different intermediate representations, these methods follow a common strategy: they transform brain activity into conditions that approximate image-derived targets. These conditions subsequently serve as inputs to a pretrained generator for image synthesis.

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.

This paradigm has substantially improved visual quality, but it also separates *plausible generation* from *stimulus-specific reconstruction*. A powerful generator can complete an incomplete category or scene cue with textures, objects, and layouts from its pretraining distribution. Therefore, the result may look convincing even when important details of the current stimulus were not preserved from the brain signal (Yang et al., 2026; Shirakawa et al., 2025). This issue arises from the structurally one-sided supervision of existing pipelines. Their principal objectives evaluate a decoded condition only by how well it matches image-derived target or generator's preferred distribution: CLIP alignment rewards semantic consistency (Scotti et al., 2023; 2024), token regression rewards similarity to visual embeddings (Beliy et al., 2026), and diffusion losses reward generator compatibility (Huo et al., 2024; Rombach et al., 2022). Once brain response has been mapped forward, however, it no longer constrains which information must survive in the condition used for generation. We call this imbalance **supervision asymmetry**: training specifies what can be generated from decoded condition, without equally specifying what evidence from current brain response should be preserved.

The consequence of this asymmetry is a weaker conditioning signal for reconstruction. Brain-to-image mapping is highly underdetermined, so information that is not rewarded by image-side objectives can be discarded and later guessed by the generator. This affects reconstruction in three ways. First, losing trial-specific cues enlarges the set of images compatible with the decoded condition. Second, collapsing the relative structure among brain states reduces the model's ability to distinguish visually similar stimuli and to generalize to unseen stimuli from noisy, limited fMRI data. Third, treating cortical features as exchangeable ignores the heterogeneous coding roles of different visual areas (Wandell et al., 2007; Dumoulin & Wandell, 2008), forcing decoder to learn an unnecessarily unconstrained readout from limited samples. These are not merely interpretability concerns: each increases uncertainty or noise in the condition supplied to generator and can therefore reduce stimulus-specific reconstruction accuracy.

We propose **MindLoop**, a framework that uses closed-loop brain-derived supervision to construct more informative and discriminative conditions for image generation. MindLoop introduces three complementary designs corresponding to the three limitations above. (i) **Brain Latent Back-Projection (Brain-LBP)** couples the forward brain-to-visual decoder with a reverse encoding path that projects stimulus representations into a subject-specific brain latent space defined by a momentum encoder. The predicted brain features are decoded back to the visual target through the shared forward decoder, encouraging visual representations to remain compatible with the brain state of the current trial. (ii) **Dual-anchored relational alignment** preserves pairwise distances and triplet directions using both CLIP representations and momentum-encoded brain states as relational anchors. It extends supervision from isolated sample pairs to population geometry, constraining category neighborhoods and continuous variations that per-sample regression does not directly preserve (Kriegeskorte et al., 2008; Kriegeskorte & Kievit, 2013). (iii) **Hierarchy-aware cortical fields** incorporates cortical organization into the decoder's information-routing mechanism. Image queries receive a learnable receptive-field bias when attending to retinotopic early and intermediate visual areas, while attention to higher-level category-selective regions remains content-driven and spatially unrestricted (Wandell et al., 2007; Dumoulin & Wandell, 2008).

Together, the three modules constrain trial state, population geometry, and cortical information source. On NSD, MindLoop gives the best low-level fidelity among all compared methods and achieves the best score on seven of eight standard metrics. These results suggest that MindLoop benefits from rebalancing the supervision asymmetry, supervising the condition with the current brain response rather than optimizing the same image target repeatedly. In summary, our contributions are threefold:

- We identify **supervision asymmetry** in generator-conditioned brain reconstruction, where image-side objectives do not explicitly preserve information from the current brain response.
- We propose **MindLoop** with three brain-derived constraints: Brain Latent Back-Projection, dual-anchored relational alignment, and hierarchy-aware cortical fields.
- MindLoop adapts to two image generators, Versatile Diffusion and SDXL; its reconstructions follow the measured cortical response, and its cortical fields recover retinotopic organization.

2 Related Work

Generator-conditioned fMRI-to-image reconstruction. Modern methods typically transform fMRI signals into conditions for pretrained generative models. Takagi & Nishimoto (2023) combine brain-decoded representations with Stable Diffusion, while MinD-Vis learns fMRI features through masked brain modeling to condition latent diffusion (Chen et al., 2023). MindEye aligns fMRI with the CLIP image space using contrastive learning and a diffusion prior (Scotti et al., 2023), and MindEye2 and MindBridge further improve shared-subject and cross-subject reconstruction (Scotti et al., 2024; Wang et al., 2024). NeuroPictor conditions diffusion on semantic and structural features decoded from cortical-surface maps (Huo et al., 2024), whereas PRISM maps brain signals into structured visual descriptions (Huang et al., 2026). Although some methods use fMRI reconstruction or self-supervision to improve brain encoding, their generator-facing conditions remain primarily optimized by image-side objectives. MindLoop instead introduces brain-derived constraints directly into the formation of these conditions.

Brain representation learning and bidirectional mapping. Neural encoding models predict stimulus-evoked brain responses from visual representations, providing a classical test of whether visual features contain neurally relevant information (Kay et al., 2008; Naselaris et al., 2011). Brain-LBP integrates this reverse direction with visual decoding: it predicts a subject-specific brain latent state from stimulus representations and passes the prediction through the shared visual decoder. Momentum target networks and latent prediction have been explored in BYOL, I-JEPA, and Brain-JEPA (Grill et al., 2020; Assran et al., 2023; Dong et al., 2024). In contrast to representation pretraining or raw-fMRI reconstruction, MindLoop uses this bidirectional mapping to improve the brain-derived condition for reconstruction of the current work.

Relational geometry and cortical topography. Representational similarity analysis compares brain and model representations through inter-stimulus geometry (Kriegeskorte et al., 2008; Kriegeskorte & Kievit, 2013), while relational knowledge distillation transfers such structure through pairwise distances and triplet angles (Park et al., 2019). MindLoop adopts this formulation without model compression, using image embeddings and subject-specific brain states as two relational anchors. Separately, retinotopic and population-receptive-field studies show stable visual-field mappings in early visual cortex and increasingly category-selective representations in higher-level areas (Wandell et al., 2007; Dumoulin & Wandell, 2008). Previous work has used this organization to predict cortical responses or functional organization (St-Yves & Naselaris, 2018; Margalit et al., 2024). MindLoop instead incorporates it into decoder readout, applying a learnable receptive-field bias to early and intermediate visual areas while retaining unrestricted content-based attention for higher-level regions.

3 Method

We propose MindLoop, a brain-supervised framework for generator-conditioned fMRI reconstruction, as illustrated in Figure 1. Given an fMRI cortical flatmap, the system reconstructs the stimulus through two separately trained pathways. The inherited low-level pathway predicts an initial image, while the high-level pathway predicts the CLIP visual tokens used to condition the generator. MindLoop modifies the high-level pathway through brain-derived constraints while keeping the low-level branch fixed, so all compared configurations share the same initial image.

3.1 Forward Pathway: From Cortical Surface to Generator Condition

To preserve the spatial organization of cortical activity while interfacing with the pretrained generator, the forward pathway encodes the fMRI flatmap into cortical latents and then decodes them into the generator's visual-token space.

Cortical tokenization. The high-level pipeline of Fig. 1 starts from the fMRI input: let $x \in \mathbb{R}^{256\times256}$ denote a cortical flatmap response. A convolutional patch embedding $\mathcal{P}$ converts the flatmap into cortical tokens, after which a fixed visual-cortex ROI mask removes patches outside the modeled cortex. The resulting sequence contains $S = 168$ tokens (the Completed Tokens in Fig. 1), whose positions are determined by the flatmap geometry and anatomical mask. During training, a masking operator $\mathcal{M}$ retains a subset $\mathcal{S}_{\text{keep}}$ of size $S_{\text{keep}} = |\mathcal{S}_{\text{keep}}|$, providing the partial online

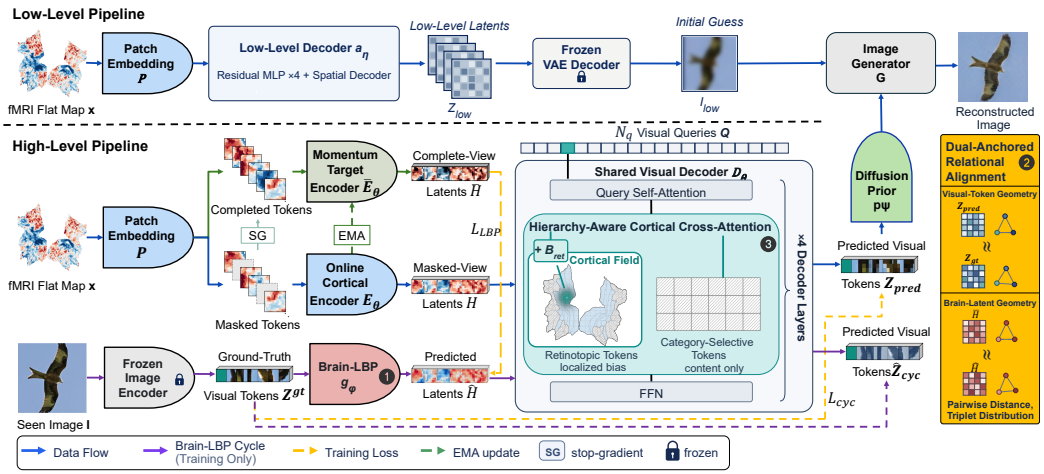

Figure 1: Overview of MindLoop. The same fMRI flatmap x supplies the inherited Low-Level Pipeline (top row) and the proposed High-Level Pipeline (second row), which maps Masked Tokens to Predicted Visual Tokens through the Online Cortical Encoder E_θ and the Shared Visual Decoder D_ϕ; a diffusion prior p_ψ then conditions the frozen generator G. Circled numbers mark the three proposed components: ❶ Brain Latent Back-Projection (Sec. 3.2), a training-only closed loop through the same-trial cortical state; ❷ Dual-Anchored Relational Alignment (Sec. 3.3), which preserves geometry in the visual-token and brain-latent spaces; ❸ Hierarchy-Aware Cortical Cross-Attention (Sec. 3.4), which applies a localized Cortical Field B_{ret} only to Retinotopic Tokens.

view paired with the complete-view cortical target introduced in Sec. 3.2. Masking is disabled at inference, when all cortical tokens are available.

Generator-facing visual decoding. An online cortical Transformer E_θ contextualizes the retained tokens, producing cortical latents H. A learned-query decoder D_θ then retrieves cortical evidence from H and organizes it into the fixed visual-token layout of the generator's CLIP image encoder (Radford et al., 2021):

$$H = E_\theta(\mathcal{M}(\mathcal{P}(x))) \in \mathbb{R}^{S_{\text{keep}} \times d_h}, \qquad Z_{\text{pred}} = D_\theta(Q, H; B_{\text{ret}}) \in \mathbb{R}^{N_q \times d_v}. \tag{1}$$

Here, $d_h = 768$, and Q contains one query per token of the target sequence: $N_q = 256$ and $d_v = 1664$ for the OpenCLIP ViT-bigG/14 patch tokens used by SDXL unCLIP, and $N_q = 257$ and $d_v = 768$ for the CLIP ViT-L/14 tokens used by Versatile Diffusion. Each query defines one output slot in the target CLIP sequence, while H supplies the cortical keys and values from which its content is retrieved. Each decoder block applies query self-attention, cortical cross-attention, and a feed-forward update. The bias B_{ret}, defined in Sec. 3.4, introduces cortical-field structure into this readout.

Generator interface and image-side supervision. As shown in Fig. 1 (bottom panel, starting from the seen image I), a conditional diffusion prior p_ψ adapts the predicted tokens to the generator's conditioning distribution. At inference, it samples $\widetilde{Z} \sim p_\psi(\cdot \mid Z_{\text{pred}})$, after which a frozen image generator produces $\widehat{I} = G(\widetilde{Z}, I_{\text{low}})$. The initial image I_{low} is obtained by decoding the predicted low-level VAE latents with a frozen decoder (Rombach et al., 2022).

Let Z^{gt} denote the CLIP visual tokens extracted from stimulus image. The image-side supervision:

$$\mathcal{L}_{\text{img}} = \mathcal{L}_{\text{align}}\big(Z_{\text{pred}}, Z^{\text{gt}}\big) + \lambda_{\text{prior}} \mathcal{L}_{\text{prior}}\big(Z_{\text{pred}}, Z^{\text{gt}}\big). \tag{2}$$

The alignment term matches the image-derived target and the prior term adapts the tokens to the generator; both are defined on the image side and do not specify which information from the measured cortical response should be preserved.

3.2 The Brain as a Prediction Target: Brain Latent Back-Projection

Because fMRI-to-image decoding is underdetermined, different mappings can reach comparable image-side losses while preserving different cortical evidence, and weakly constrained trial-specific

cues can be replaced by the generator prior. Brain Latent Back-Projection (Brain-LBP; ❶ in Fig. 1) treats the response as a same-trial prediction target: it maps the stimulus representation back to the corresponding cortical state and passes that state through the shared visual decoder D_θ, adding no inference-time cost.

Stable same-trial cortical target. The online representation H is derived from a masked input and evolves rapidly during training, making it unsuitable as a target. We instead encode the complete, unmasked response using a momentum copy of the cortical encoder, following momentum-target representation learning (Grill et al., 2020; Assran et al., 2023):

$$\bar{H} = \mathrm{sg}[E_{\bar{\theta}}(\mathcal{P}(x))] \in \mathbb{R}^{S\times d_h}, \qquad \bar{\theta} \leftarrow m\bar{\theta} + (1-m)\theta. \tag{3}$$

Here, $\mathrm{sg}[\cdot]$ denotes stop-gradient and $m = 0.996$.

Brain-latent prediction and shared-decoder closure. The reverse pathway takes the fixed ground-truth visual tokens Z^{gt} as input, avoiding dependence on the current accuracy of Z_{pred}. A set of S learned cortical queries Q^B, indexed by the valid cortical patches, retrieves information from these tokens and directly predicts the cortical state:

$$\widehat{H} = g_\phi\big(Q^B; Z^{\mathrm{gt}}\big) \in \mathbb{R}^{S\times d_h}, \qquad Z_{\mathrm{cyc}} = D_\theta\Big(Q, \widehat{H}; 0\Big). \tag{4}$$

In this auxiliary pass, $\widehat{H}$ supplies all S cortical keys and values, and the cortical-field bias is omitted. Because the same D_θ maps the online cortical state H to Z_{pred}, the cycle objective directly regularizes the reconstruction pathway rather than training a detached encoding model.

Trial-level objective. We jointly enforce latent agreement, trial discriminability, cortical grounding, and shared-decoder closure:

$$\begin{aligned}\mathcal{L}_{\mathrm{trial}} = \underbrace{\left\|\widehat{H} - \bar{H}\right\|_F^2 + \lambda_{\cos}\,\mathbb{E}_s\left[1 - \cos\Big(\widehat{h}_s, \bar{h}_s\Big)\right]}_{\text{latent agreement}} \\ + \underbrace{\lambda_c\,\mathcal{L}_{\mathrm{NCE}}\Big(\widehat{H}, \bar{H}\Big)}_{\text{trial discriminability}} + \underbrace{\lambda_p\left\|D_{\mathrm{pix}}(\widehat{H}) - x\right\|_\Omega^2}_{\text{cortical grounding}} + \underbrace{\lambda_{\mathrm{cyc}}\left\|Z_{\mathrm{cyc}} - Z^{\mathrm{gt}}\right\|_F^2}_{\text{shared-decoder closure}}.\end{aligned} \tag{5}$$

Ω denotes the cortical patches retained by the ROI mask. The symmetric InfoNCE loss (van den Oord et al., 2018) is computed across the batch using mean-pooled, ℓ_2-normalized states. The cortical target prevents the closure objective from being satisfied by an autoencoder over Z^{gt} alone, while decoder sharing ensures that the reverse pathway regularizes the forward reconstruction. Loss weights and the warm-up schedule are provided in Appendix A.

3.3 The Brain as a Geometric Reference: Dual-Anchored Relational Alignment

Brain-LBP constrains each trial independently, yet predictions with similar pointwise errors can still induce different neighborhood structures across stimuli. Dual-Anchored Relational Alignment (❷ in Fig. 1) therefore preserves batch-level geometry in both the visual-token and cortical-latent spaces: the visual anchor keeps stimulus-defined relations, and the cortical anchor adds response-dependent structure not captured by image embeddings.

For a batch of B trials, let $U = \{u_i\}_{i=1}^B$ denote predictions flattened over tokens and channels, and let $V = \{v_i\}_{i=1}^B$ denote the corresponding stop-gradient references. We characterize their geometry using normalized pairwise distances and triplet angles:

$$d_{ij}^U = \frac{\|u_i - u_j\|_2}{\frac{1}{B(B-1)}\sum_{a\neq b}\|u_a - u_b\|_2}, \qquad a_{ijk}^U = \frac{(u_j - u_i)^\top (u_k - u_i)}{\|u_j - u_i\|_2\,\|u_k - u_i\|_2}. \tag{6}$$

The statistics d_{ij}^V and a_{ijk}^V are defined analogously.

We define the relational discrepancy and its two applications as

$$\begin{aligned}\mathcal{R}(U,V) = \lambda_d\,\mathbb{E}_{i,j}\left[\left(d_{ij}^U - d_{ij}^V\right)^2\right] + \lambda_a\,\mathbb{E}_{i,j,k}\left[\left(a_{ijk}^U - a_{ijk}^V\right)^2\right], \\ \mathcal{L}_{\mathrm{rel}}^{\mathrm{vis}} = \mathcal{R}\big(Z_{\mathrm{pred}}, Z^{\mathrm{gt}}\big), \qquad \mathcal{L}_{\mathrm{rel}}^{\mathrm{brain}} = \mathcal{R}\Big(\widehat{H}, \bar{H}\Big).\end{aligned} \tag{7}$$

Each anchor is matched within its native space, requiring no shared coordinate system between the visual and cortical representations. The cortical term uses $\widehat{H}$ rather than the masked, variable-length H because it shares the complete-view token structure of $\bar{H}$ and, through the shared-decoder closure in Eq. equation 5, must remain readable by D_θ. Relational settings are given in Appendix A.

3.4 The Brain as an Architectural Blueprint: Hierarchy-Aware Cortical Fields

Standard cross-attention provides no explicit prior over cortical topology and must learn spatial routing entirely from paired brain-image data. We instead inject into the cross-attention a bias reflecting a known property of visual cortex (Hierarchy-Aware Cortical Cross-Attention, ❸ in Fig. 1, inside the Shared Visual Decoder): early and intermediate visual areas exhibit strong retinotopic organization, whereas higher-level category-selective areas are less well described by a localized spatial prior. Let $p_s \in [-1,1]^2$ denote the normalized flatmap coordinate of cortical token s. For each spatial query q_i we learn a cortical center $\mu_i = \tanh(\widetilde{\mu}_i)$ and a width $\sigma_i = \mathrm{softplus}(\widetilde{\sigma}_i) + \sigma_{\min}$, and add to the cross-attention logits

$$A_{is} = \mathrm{softmax}_s\left(\frac{q_i^\top k_s}{\sqrt{d_k}} + B_{is}^{\mathrm{ret}}\right), \qquad B_{is}^{\mathrm{ret}} = -\chi_{is}\,\frac{\|p_s - \mu_i\|_2^2}{2\sigma_i^2}, \tag{8}$$

where $\chi_{is} = 1$ only when token s lies in an early or intermediate retinotopic area and q_i is not the first query, and $\chi_{is} = 0$ otherwise. The first query (the class token for ViT-L/14) therefore reads out without a cortical field, as do queries over higher-level category-selective areas (Appendix A). The bias reweights rather than masks: the content-dependent dot product is retained for every query–token pair, so higher-level readout remains fully content-driven while spatial queries are encouraged to draw on localized retinotopic evidence. Each query's field (μ_i, σ_i) is shared across attention heads and decoder layers. We call them cortical fields because they are defined in cortical flatmap coordinates rather than visual-field coordinates. The bias therefore assumes a fixed correspondence between image and cortical coordinates, so we disable horizontal-flip augmentation during training; Sec. 4.4 turns this dependence into a training-data intervention.

3.5 Training Objective

The high-level objective combines the image-side supervision of equation 2 with the trial-state and relational constraints:

$$\mathcal{L}_{\mathrm{high}} = \mathcal{L}_{\mathrm{img}} + \lambda_{\mathrm{vis}}\mathcal{L}_{\mathrm{rel}}^{\mathrm{vis}} + w(t)\left[\lambda_{\mathrm{trial}}\mathcal{L}_{\mathrm{trial}} + \lambda_{\mathrm{brain}}\mathcal{L}_{\mathrm{rel}}^{\mathrm{brain}}\right]. \tag{9}$$

The cortical-field bias B^{ret} introduces no additional loss. Terms involving the momentum cortical target are weighted by $w(t)$, which is zero for 30 epochs and then ramps linearly to one over 20, letting the cortical-to-visual mapping stabilize before brain-derived supervision is fully applied.

The inherited low-level branch is trained separately with ℓ_1 losses on the predicted VAE latents and decoded image, and remains frozen during MindLoop training (Appendix A). All configurations use the same generator and low-level branch, isolating the proposed brain-derived constraints in our ablations.

4 Experiments

We ask whether brain-derived supervision improves reconstruction (Sec. 4.2), what each constraint contributes (Sec. 4.3), and whether reconstructions follow the measured brain response (Sec. 4.4).

4.1 Experimental Protocol

Data and reconstruction protocol. We evaluate on the Natural Scenes Dataset (NSD) (Allen et al., 2022) and, following prior work (Scotti et al., 2023; Huo et al., 2024), train subject-specific models for subjects 1, 2, 5, and 7 without data from other subjects, testing on the 982 shared images with responses averaged over three repetitions. The high-level pathway predicts OpenCLIP ViT-bigG/14 image tokens; images are generated with the SDXL unCLIP decoder of MindEye2 (Scotti et al.,

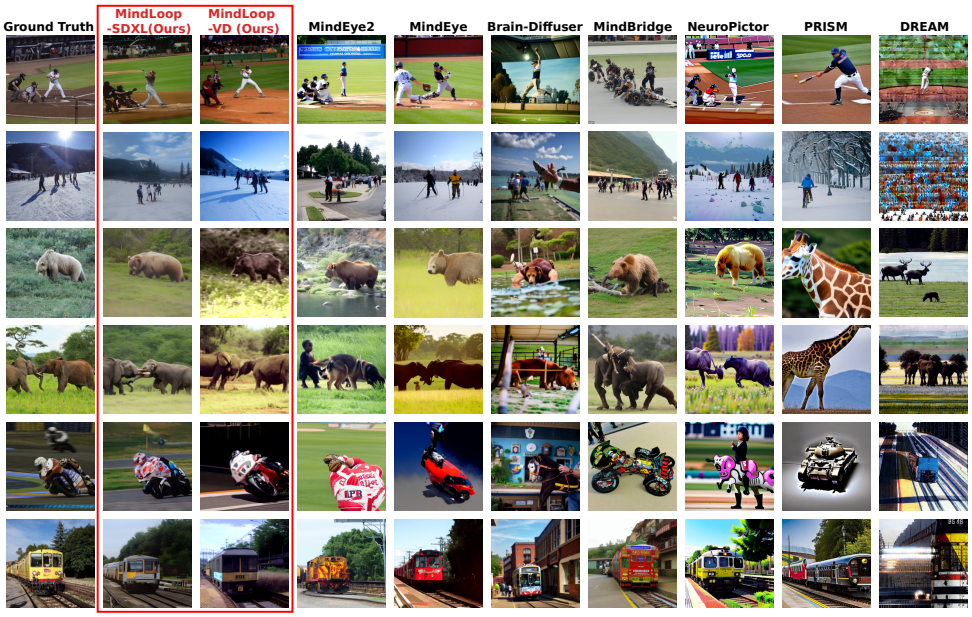

Figure 2: Qualitative comparison on six shared test images from subject 1. MindLoop more consistently preserves object identity, pose, relative position, orientation, and scene layout.

Table 1: MindLoop-SDXL scores best on seven of the eight metrics. Four-subject means on the 982-image NSD test set; identification accuracies in %. Best and second-best results are shown in **bold** and underlined.

		Low-Level				High-Level			
Model	Gen.	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep.↑	CLIP↑	EffNet-B↓	SwAV↓
DREAM (Xia et al., 2024)	SD	0.0949	0.2306	75.02	82.56	82.47	93.53	0.8614	0.5927
PRISM (Huang et al., 2026)	SDXL	0.1377	0.2981	63.53	74.49	78.91	82.99	0.8504	0.5426
NeuroPictor (Huo et al., 2024)	SD	0.1588	0.2920	94.42	96.81	92.60	90.15	0.7049	0.4258
MindBridge (Wang et al., 2024)	VD	0.1468	0.2976	88.31	94.49	88.15	91.77	0.7873	0.4533
Brain-Diffuser (Ozcelik & VanRullen, 2023)	VD	0.2287	0.2731	85.66	84.36	71.03	74.89	0.9091	0.5555
MindEye (Scotti et al., 2023)	VD	0.3168	0.3344	92.84	96.77	94.00	93.07	0.6605	0.3829
MindEye2 (Scotti et al., 2024)	SDXL	0.2705	0.3910	93.65	97.30	93.33	90.75	0.6858	0.3770
MindLoop-VD (ours)	VD	0.3358	0.3998	93.23	96.86	94.47	**94.61**	0.6577	0.3856
MindLoop-SDXL (ours)	SDXL	**0.3476**	**0.4427**	**96.67**	**98.64**	**95.39**	92.64	**0.6340**	**0.3705**

2024) under the standard best-of-16 protocol (Scotti et al., 2023), followed by caption-conditioned refinement and low-level blending as in MindEye2. We call this model MindLoop-SDXL; a Versatile Diffusion variant, MindLoop-VD, is given in Appendix B. MindLoop-SDXL and MindLoop-VD therefore use the same generative backbones as MindEye2 and MindEye, respectively. Training and reconstruction settings are given in Appendix A, and per-subject results and hyperparameter sensitivity in Appendices C and D.

Metrics. We report the eight metrics commonly used in fMRI reconstruction: PixCorr, SSIM (Wang et al., 2004), AlexNet(2), and AlexNet(5) (Krizhevsky et al., 2012) for low-level fidelity, and Inception-V3 (Szegedy et al., 2016), CLIP (Radford et al., 2021), EfficientNet-B1 (Tan & Le, 2019), and SwAV (Caron et al., 2020) for high-level fidelity. The last two are correlation distances (↓); the remaining feature metrics are two-way identification accuracies (↑). Cross-method results are averaged over four subjects, and controlled ablations use subject 1.

4.2 Reconstruction Quality

Table 1 compares MindLoop with representative fMRI-to-image decoders spanning different intermediate representations, training regimes, and generative backbones. Despite these differences, existing methods share the general strategy of predicting image-derived conditions for a pretrained generator. Across this diverse set of approaches, MindLoop-SDXL achieves the best performance on seven of eight standard metrics. The gains are not confined to semantic identification: MindLoop-

Table 2: Ablation of the brain-derived constraints on subject 1 under the final recipe. LBP denotes Brain-LBP (Sec. 3.2), $\mathcal{R}_{\text{vis}}$ and $\mathcal{R}_{\text{brain}}$ the visual and cortical relational anchors (Sec. 3.3), and CF the hierarchy-aware cortical fields (Sec. 3.4). The cortical anchor is defined on the Brain-LBP state and is therefore removed together with Brain-LBP. Identification accuracies are reported in %.

Brain-derived constraints				Low-Level				High-Level			
LBP	$\mathcal{R}_{\text{vis}}$	$\mathcal{R}_{\text{brain}}$	CF	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep.↑	CLIP↑	EffNet-B↓	SwAV↓
–	–	–	–	0.3948	0.4467	98.26	99.22	96.21	93.10	0.6239	0.3615
–	✓	–	✓	0.4007	0.4532	98.43	99.02	96.53	93.79	0.6103	0.3543
✓	–	–	✓	0.3968	0.4459	98.29	99.28	96.76	94.03	0.6198	0.3673
✓	✓	✓	–	0.3799	0.4505	97.72	98.96	95.15	92.31	0.6445	0.3769
✓	✓	✓	✓	0.4021	0.4550	98.59	99.46	96.81	94.12	0.6058	0.3477

SDXL also obtains the strongest PixCorr, SSIM, and AlexNet scores, indicating improved preservation of low-level and spatial image structure alongside high-level content. Relative to the strongest non-MindLoop baseline, MindLoop-SDXL improves PixCorr by 0.0308, SSIM by 0.0517, Inception identification by 1.39 points, and EffNet-B distance by 0.0265.

We next control for the generative backbone to separate these improvements from advances in the image generator itself. MindLoop-VD uses the same Versatile Diffusion generator as MindEye, while MindLoop-SDXL uses the same SDXL unCLIP generator as MindEye2. Under these matched comparisons, MindLoop-VD improves over MindEye on seven of eight metrics and achieves the highest CLIP identification score in the table (94.61%). MindLoop-SDXL improves over MindEye2 on all eight metrics, although MindEye2 additionally benefits from cross-subject pretraining. MindLoop-SDXL also wins seven of eight metrics against the subject-specific MindEye2 model trained on all 40 sessions of subject 1 (Appendix E). Thus, the improvement persists when the generator is controlled and cannot be attributed solely to a stronger generative backbone.

Figure 2 further illustrates the type of information that distinguishes the reconstructions. Several baselines recover the correct broad semantic category but alter trial-specific properties such as pose, relative position, orientation, or scene layout. MindLoop more often preserves these properties jointly with object identity. This pattern is consistent with the motivation behind brain-derived supervision: constraining the decoded condition with the measured cortical response reduces the freedom of the generator to replace stimulus-specific evidence with plausible alternatives. We next test this interpretation through controlled ablations of the three proposed constraints.

4.3 How the Brain-Derived Constraints Contribute

Table 2 removes one constraint at a time from the full model, so that every row remains a valid configuration, and uses the same reconstruction recipe throughout.

Hierarchy-aware routing. Removing the cortical fields produces the largest change in the table: EffNet-B rises from 0.6058 to 0.6445, SwAV from 0.3477 to 0.3769, and Inception and CLIP identification fall by 1.7 and 1.8 percentage points. Routing visual queries through retinotopically organized fields is therefore the strongest individual contribution.

Cross-trial geometry. Removing both relational anchors while retaining Brain-LBP and the cortical fields degrades all eight metrics, including EffNet-B ($0.6058 \rightarrow 0.6198$) and SwAV ($0.3477 \rightarrow 0.3673$). Thus, preserving cross-trial geometry provides information complementary to the trial-level cortical supervision and hierarchy-aware routing constraints.

Same-trial cortical supervision. Removing Brain-LBP, and consequently the cortical relational anchor defined on its latent state, also degrades all eight metrics relative to the full model. EffNet-B increases from $0.6058 \rightarrow 0.6103$ and SwAV from $0.3477 \rightarrow 0.3543$, while CLIP identification decreases from $94.12\% \rightarrow 93.79\%$. Together, the ablations show that the full combination of trial-level cortical supervision, relational alignment, and hierarchy-aware routing yields the best performance across all eight metrics.

4.4 What MindLoop Takes from the Brain

We probe this with two interventions: replacing the brain response at inference, and breaking the image-to-cortex correspondence in training.

Reconstructions follow the measured response. On the 100 NSD *special100* test images of subject 1, we change only the brain input and keep every other stage and all random seeds fixed. When each image receives the response evoked by a different image, all four identification scores fall to chance (47.93–50.00%) and PixCorr to 0.0069, whereas the same reconstructions recover near-normal fidelity against the donor image (Appendix Table 6). Averaging more trials improves all eight metrics monotonically (Fig. 3a; Appendix Table 6), and reconstructions inherit each subject's representational geometry (Appendix F). The three constraints also bring reconstructions closer to the brain: an encoding model fit on training images only identifies the reconstruction made from each measured response among all 982 test reconstructions in 43.7% of cases, versus 37.7% for the plain decoder of Table 2, which uses none of them (paired bootstrap $p = 0.002$), and the gap is already present in the decoded condition (87.7% vs. 80.1%, $p < 0.001$; Fig. 3b).

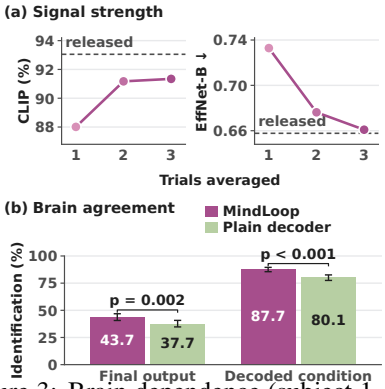

Figure 3: Brain dependence (subject 1, final recipe). (a) *special100* quality vs. averaged trials (re-derived; dashed: released average). (b) 982-way identification of the measured response; error bars: 95% bootstrap CIs; p: paired bootstrap, 2,000 resamples.

A topographic map emerges from paired data. No retinotopic labels or precomputed receptive fields are used. In MindLoop-SDXL the learned cortical-field centers map image position onto the flatmap in all four subjects: image-x onto μ_x ($r = -0.699, -0.673, -0.579, -0.744$ for subjects 1, 2, 5, 7) and image-y onto μ_y ($r = -0.738, -0.786, -0.584, -0.461$; permutation $p < 0.001$ in every subject; Fig. 7). Prior fine-tuning keeps the cortical encoder and decoder frozen, so these are the fields of the final model. In a pair of MindLoop models of subject 1 that differ only in horizontal-flip augmentation, flipping removes the x map ($-0.48 \rightarrow +0.09$; Fig. 4) but leaves the y map intact (-0.55 vs. -0.56), and in the same pair the unconstrained attention of the Brain-LBP predictor shows the x organization only without flips ($r = -0.59$ vs. -0.07; Fig. 8).

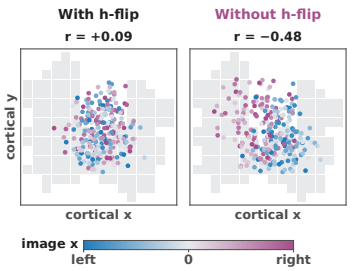

Figure 4: Decoder cortical-field centers in two MindLoop models of subject 1 that differ only in horizontal-flip augmentation, colored by the image-x position of their visual queries.

5 Conclusions

Generator-conditioned fMRI reconstruction learns its conditions mainly from image-derived targets, so nothing requires them to preserve information from the current brain response. We proposed MindLoop to close this loop by making the brain part of the supervision: Brain Latent Back-Projection treats the current response as a same-trial prediction target, dual-anchored relational alignment preserves cross-trial geometry in visual and cortical spaces, and hierarchy-aware cortical fields route spatial queries through retinotopic cortex. MindLoop adapts to both Versatile Diffusion and SDXL, and MindLoop-SDXL achieves the best score on seven of eight standard metrics. Its reconstructions follow the measured cortical response, and its cortical fields recover retinotopic organization from paired data alone in every subject.

More broadly, MindLoop treats the brain not only as an input to decode but also as a prediction target, a geometric reference, and an architectural blueprint for the decoder. Because these constraints act on the brain side of the decoder, they can be paired with new generators and larger fMRI datasets, and we expect the same principle to extend to other forms of neural decoding, such as video and cross-subject reconstruction.

AI USE STATEMENT

Large language model assistants were used solely to polish the language and improve the clarity and readability of the manuscript. All experimental designs, analyses, results, interpretations, and claims were developed and verified by the authors, who take full responsibility for the content.

ETHICS STATEMENT

This work uses only the publicly available Natural Scenes Dataset (Allen et al., 2022), which was collected with informed consent under institutional review board approval; no new human data were collected. Reconstructing images from brain activity raises questions of mental privacy. Our models are trained separately for each participant on many hours of 7T fMRI recorded with the participant's cooperation, and they cannot be applied to other individuals or without active participation in the scanner. Such technology should be used only with the informed consent of the person whose brain activity is decoded.

REPRODUCIBILITY STATEMENT

Section 4.1 and Appendix A specify the architecture, training schedule, and reconstruction settings, and Appendix B describes the MindLoop-VD variant. All data and pretrained components are publicly available: NSD (Allen et al., 2022), OpenCLIP ViT-bigG/14, the SDXL unCLIP decoder of MindEye2 (Scotti et al., 2024), GIT and zavychromaXL for refinement, and Versatile Diffusion (Xu et al., 2023). Reconstructions use fixed per-image random seeds, metrics follow the evaluation code of MindEye (Scotti et al., 2023), and all statistical tests state their resampling settings (Appendix F). We will release code, trained models, and reconstructions upon publication.

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, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, 2022. doi: 10.1038/s41593-021-00962-x.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023.

Roman Beliy, Amit Zalcher, Jonathan Kogman, Navve Wasserman, and Michal Irani. Brain-IT: Image reconstruction from fMRI via brain-interaction transformer. In *International Conference on Learning Representations (ICLR)*, 2026.

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

Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Su Xian Chong, Fang Ji, Nathanael Ren Jie Tong, Christopher Li Hsian Chen, and Juan Helen Zhou. Brain-JEPA: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. *Advances in Neural Information Processing Systems*, 37:86048–86073, 2024.

Serge O. Dumoulin and Brian A. Wandell. Population receptive field estimates in human visual cortex. *NeuroImage*, 39(2):647–660, 2008. doi: 10.1016/j.neuroimage.2007.09.034.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

Weiyu Guo, Guoying Sun, Jianxiang He, Tong Shao, Shaoguang Wang, Ziyang Chen, Meisheng Hong, Ying Sun, and Hui Xiong. A survey of fMRI to image reconstruction. *arXiv preprint arXiv:2502.16861*, 2025.

Zheng Huang, Enpei Zhang, Weikang Qiu, Yinghao Cai, Carl Yang, Elynn Chen, Xiang Zhang, Rex Ying, Dawei Zhou, and Yujun Yan. Seeing through the brain: New insights from decoding visual stimuli with fMRI. In *International Conference on Learning Representations*, 2026.

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, Cham, 2024. Springer Nature Switzerland.

Yukiyasu Kamitani, Misato Tanaka, and Ken Shirakawa. Visual image reconstruction from brain activity via latent representation. *Annual Review of Vision Science*, 11(1):611–634, 2025. doi: 10.1146/annurev-vision-110423-023616.

Kendrick N. Kay, Thomas Naselaris, Ryan J. Prenger, and Jack L. Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008. doi: 10.1038/nature06713.

Nikolaus Kriegeskorte and Rogier A. Kievit. Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17(8):401–412, 2013. doi: 10.1016/j.tics.2013.06.007.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Eshed Margalit, Hyodong Lee, Dawn Finzi, James J. DiCarlo, Kalanit Grill-Spector, and Daniel L. K. Yamins. A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron*, 112(14):2435–2451.e7, 2024. doi: 10.1016/j.neuron.2024.04.018.

Yoichi Miyawaki, Hajime Uchida, Okito Yamashita, Masa-aki Sato, Yusuke Morito, Hiroki C. Tanabe, Norihiro Sadato, and Yukiyasu Kamitani. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5):915–929, 2008. doi: 10.1016/j.neuron.2008.11.004.

Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410, 2011. doi: 10.1016/j.neuroimage.2010.07.073.

Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023. doi: 10.1038/s41598-023-42891-8.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3967–3976, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Paul S. Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J. Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth A. Norman, and Tanishq Mathew Abraham. 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 Torrico, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq Mathew Abraham. MindEye2: Shared-subject models enable fMRI-to-image with 1 hour of data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44038–44059. PMLR, 2024.

Ken Shirakawa, Yoshihiro Nagano, Misato Tanaka, Shuntaro C. Aoki, Yusuke Muraki, Kei Majima, and Yukiyasu Kamitani. Spurious reconstruction from brain activity. *Neural Networks*, 190: 107515, 2025.

Ghislain St-Yves and Thomas Naselaris. The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. *NeuroImage*, 180:188–202, 2018. doi: 10.1016/j.neuroimage.2017.06.035.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14453–14463, 2023.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 6105–6114, 2019.

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

Brian A. Wandell, Serge O. Dumoulin, and Alyssa A. Brewer. Visual field maps in human cortex. *Neuron*, 56(2):366–383, 2007. doi: 10.1016/j.neuron.2007.10.012.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. GIT: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research*, 2022.

Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. MindBridge: A cross-subject brain decoding framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11333–11342, 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 Öztireli, and Jing-Hao Xue. DREAM: Visual decoding from reversing human visual system. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8211–8220, 2024.

Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7754–7765, 2023.

Lan Yang, Minghan Yang, Ke Li, Honggang Zhang, Kaiyue Pang, and Yi-Zhe Song. SynMind: Reducing semantic hallucination in fMRI-based image reconstruction. *arXiv preprint arXiv:2601.17857*, 2026.

A Implementation Details

The cortical response is represented as a 256×256 flatmap. A convolutional patch embedding with kernel size and stride 16 produces a 16×16 grid, from which patches outside the visual-cortex mask are removed. The cortical encoder contains six Transformer layers with 12 heads and width 768; the learned-query decoder contains four cross-attention layers with 12 heads.

Brain-LBP projects the stimulus tokens to width 512 and uses three blocks containing cross-attention, query self-attention, and feed-forward transformations, followed by a final self-attention block. It uses eight heads, dropout 0.15, stochastic depth up to 0.1, EMA coefficient $m = 0.996$, and contrastive temperature $\tau = 0.1$. Approximately 25% of online cortical tokens are masked by alternating random-token and hierarchy-group masking.

The current bilateral spatially biased ROI set is $\mathcal{V}_{\mathrm{ret}} = \{\mathrm{V1}, \mathrm{V2}, \mathrm{V3}, \mathrm{V4}, \mathrm{V8}, \mathrm{PIT}, \mathrm{VVC}\}$. We set $\sigma_{\min} = 0.05$ and initialize receptive-field widths to 1.15. The parameter table contains $256 \times 3 = 768$ scalars and is shared across decoder layers; the global-token row is ignored when constructing the bias. Horizontal-flip augmentation is disabled because it disrupts the fixed correspondence between image and cortical coordinates.

SDXL reconstruction. The diffusion prior is fine-tuned for 60 epochs (batch size 32, learning rate 10^{-4}) with the cortical encoder and token decoder frozen. Reconstruction starts from the low-level output via img2img at strength 0.65; for best-of-16 selection, 16 samples are drawn from the diffusion prior with 20 DDIM steps and one image is decoded per sample. All pretrained components remain frozen. The unCLIP decoder uses 38 sampling steps. For refinement, captions are generated by GIT-large-COCO (Wang et al., 2022) from the selected reconstruction, and the SDXL checkpoint zavychromaXL v3.0 denoises the re-noised latent over the last 4 steps of a 25-step schedule, with classifier-free guidance 5 and the negative prompt of MindEye2. The final image is $0.8\times$ the refined image plus $0.2\times$ the low-level reconstruction; the latter comes from a low-level pathway trained on single-trial responses, whereas the img2img initialization uses the pathway trained on trial-averaged responses.

Loss weights and schedules. With t the (0-indexed) epoch out of 400, $\mathcal{L}_{\text{align}}$ in Eq. (2) combines a contrastive term on the projected Z_{pred} (weight 0.3; MixCo InfoNCE with τ=0.006 for t<132, then SoftCLIP with τ cosine-annealed from 0.004 to 0.0075) with token-level regression to Z^{gt}: a token cosine loss (weight 1.0; 0.6 per-token plus 0.4 on token means), per-channel moment matching (weight rising from 1 to 4 over $t\leq 40$, held at 4, then decaying to 2 over $t\geq 320$), CORAL (weight rising from 0 to 1 over $t\leq 40$) and a token MSE weighted by $0.2\,\lambda_{\text{prior}}(t)$, where $\lambda_{\text{prior}}(t) = 8\min(1, t/50)$. Both relational anchors use λ_d=0.2 and λ_a=0.1 with λ_{vis}=λ_{brain}=1, and the trial-level objective uses λ_{trial}=1, λ_{cos}=0.5, λ_c=0.1, λ_p=0.5 and λ_{cyc}=0.5. For t<132, MixCo mixes the input flatmaps of half of each batch with $\beta \sim \mathrm{Beta}(0.15, 0.15)$, and the mixed flatmap is fed to all branches. Models are trained for 400 epochs. Training uses AdamW (peak learning rate 1.5×10^{-4}, one-cycle schedule with 0.5% warm-up), gradient clipping at 1.0 and bf16 mixed precision.

B MindLoop-VD

MindLoop-VD applies the same brain-derived constraints with the generator stack of MindEye (Scotti et al., 2023). The high-level pathway predicts CLIP ViT-L/14 image tokens (257×768), and reconstructions are generated with Versatile Diffusion (Xu et al., 2023), initialized by img2img from the low-level output at strength 0.75, under the standard best-of-16 protocol. The diffusion prior is used as trained, without separate fine-tuning, and no caption refinement or blending is applied. The image-only configuration gates all caption-branch losses to zero and sets the caption mixing coefficient to zero during Versatile Diffusion inference.

C Per-Subject Results

Table 3 reports the eight metrics of both MindLoop variants for each subject; the mean rows are the four-subject averages in Table 1. MindLoop-SDXL scores better than MindLoop-VD on SSIM,

Alex(2), Alex(5), Inception, EffNet-B and SwAV in every subject. Figures 5 and 6 show, for each variant, reconstructions of shared test images from all four subjects.

Table 3: Per-subject results of MindLoop-SDXL and MindLoop-VD on the 982-image NSD test set. Identification accuracies are reported in %.

		Low-Level				High-Level			
Model	Subject	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep.↑	CLIP↑	EffNet-B↓	SwAV↓
MindLoop-SDXL	1	0.4021	0.4550	98.59	99.46	96.81	94.12	0.6058	0.3477
	2	0.3653	0.4488	98.02	99.41	95.86	93.27	0.6278	0.3621
	5	0.3191	0.4332	95.98	98.42	95.76	93.19	0.6247	0.3700
	7	0.3038	0.4338	94.10	97.27	93.12	89.98	0.6778	0.4022
	Mean	0.3476	0.4427	96.67	98.64	95.39	92.64	0.6340	0.3705
MindLoop-VD	1	0.4035	0.4071	96.34	98.23	95.42	95.49	0.6356	0.3667
	2	0.3549	0.4070	94.71	97.61	94.26	94.68	0.6581	0.3838
	5	0.2991	0.3992	91.68	96.47	95.46	95.25	0.6503	0.3848
	7	0.2857	0.3860	90.20	95.13	92.72	93.01	0.6866	0.4071
	Mean	0.3358	0.3998	93.23	96.86	94.47	94.61	0.6577	0.3856



Figure 5: Reconstructions of MindLoop-SDXL from each of the four subjects. GT is the viewed image and Subj n the reconstruction from the brain activity of subject n; each row shows two test images.

D Hyperparameter Sensitivity

Table 4 varies the two hyperparameters of the SDXL reconstruction stage on all four subjects: the number of final denoising steps t_p (out of 25) re-run by the caption-conditioned refinement, and the weight α of the low-level reconstruction in the final blend. The results are stable across the six settings: PixCorr and SSIM vary by at most 0.013, the two-way identification accuracies by at most 0.3 points except CLIP, and EffNet-B and SwAV by less than 0.008. CLIP favors shallow refinement (92.64 at $t_p = 4$ vs. 90.80 at $t_p = 9$), and the smaller blending weight keeps EffNet-B and SwAV lower; we use $t_p = 4$ and $\alpha = 0.20$ throughout.

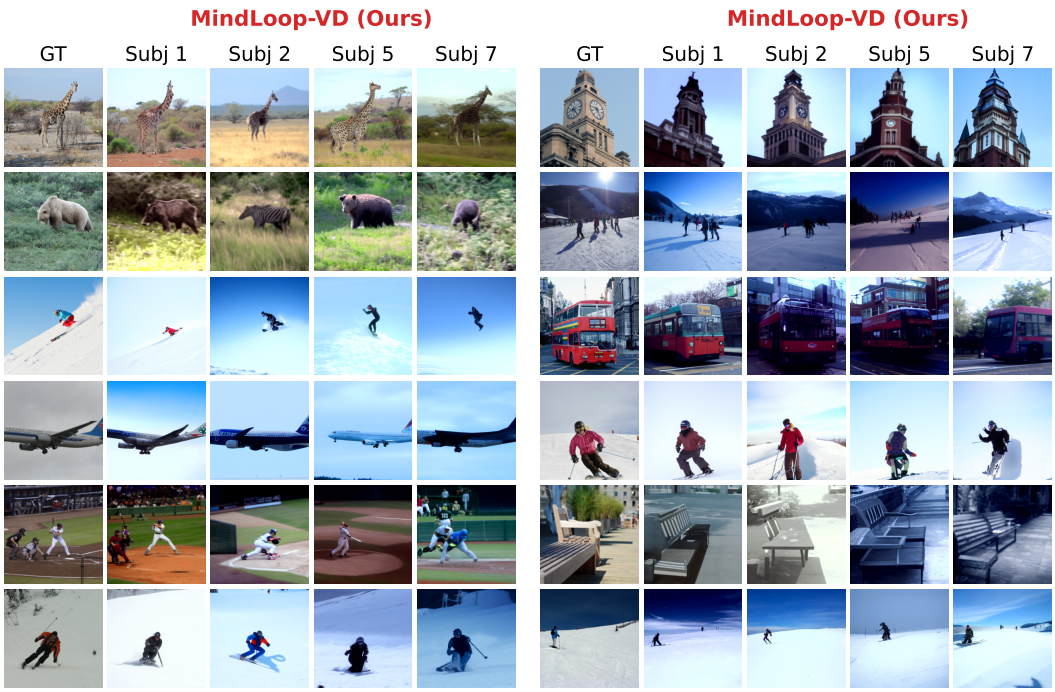

Figure 6: Reconstructions of MindLoop-VD from each of the four subjects, arranged as in Figure 5.

Table 4: Sensitivity of MindLoop-SDXL to the refinement depth t_p and the blending weight α (four-subject means on the 982-image NSD test set). The shaded row is the setting used throughout the paper. Identification accuracies are reported in %.

		Low-Level				High-Level			
t_p	α	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep.↑	CLIP↑	EffNet-B↓	SwAV↓
4	0.20	0.3476	0.4427	96.67	98.64	95.39	92.64	0.6340	0.3705
4	0.25	0.3561	0.4527	96.73	98.65	95.39	92.66	0.6366	0.3759
6	0.20	0.3481	0.4437	96.65	98.62	95.24	92.15	0.6319	0.3687
6	0.25	0.3566	0.4537	96.70	98.63	95.24	92.21	0.6344	0.3741
9	0.20	0.3489	0.4453	96.57	98.56	95.13	90.80	0.6310	0.3683
9	0.25	0.3574	0.4553	96.62	98.56	95.13	90.89	0.6334	0.3735

E Same-Setting Comparison with MindEye2

Table 5 compares MindLoop-SDXL with the subject-1 result reported by Scotti et al. (2024) for MindEye2 trained from scratch on subject 1 alone, the same single-subject setting as ours. MindLoop-SDXL obtains the best score on seven of the eight metrics.

Table 5: Subject-1 comparison with MindEye2. MindEye2 values are those reported by Scotti et al. (2024) for the model trained from scratch on all 40 sessions of subject 1 (their Table 6). MindLoop-SDXL uses subject-1 data only. Identification accuracies are reported in %; best results are shown in **bold**.

Model	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep.↑	CLIP↑	EffNet-B↓	SwAV↓
MindEye2 (from scratch)	0.376	0.440	97.5	99.1	95.4	92.6	0.612	**0.341**
MindLoop-SDXL (ours)	**0.4021**	**0.4550**	**98.59**	**99.46**	**96.81**	**94.12**	**0.6058**	0.3477

F Brain-Dependence Analyses

The interventions and the agreement analysis use subject 1; the geometry analysis and the cortical-field readout use all four subjects. All reconstructions follow the final recipe of Section 4.1.

Interventions. The NSD *special100* set consists of 100 of the shared test images (Allen et al., 2022). For swapping, each image receives the response to another *special100* image under a fixed derangement, so that no image keeps its own response. For degrading, we re-derive single-trial cortical flatmaps from the NSD single-trial responses with the same surface projection as the released averages (median correlation with the released average $r = 0.9986$) and average the first one, two, or three presentations of each image. Every other stage of the pipeline and all per-image random seeds are unchanged.

Table 6: Brain-input interventions on subject 1 (*special100*; two-way identification among the 100 images). Rows marked re-derived use flatmaps rebuilt from single-trial responses.

Brain input	Scored against	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep.↑	CLIP↑	EffNet-B↓	SwAV↓
Measured (3-trial average)	viewed image	0.3745	0.4194	98.36	99.84	94.27	93.05	0.6577	0.3647
Swapped	viewed image	0.0069	0.3302	49.97	48.99	50.00	47.93	0.9716	0.6282
Swapped	donor image	0.3702	0.4156	98.73	99.43	93.21	89.57	0.6747	0.3788
Single trial (re-derived)	viewed image	0.3157	0.4094	96.39	97.93	91.35	88.01	0.7329	0.4152
2-trial average (re-derived)	viewed image	0.3591	0.4150	98.66	98.76	93.65	91.17	0.6762	0.3870
3-trial average (re-derived)	viewed image	0.3716	0.4180	98.75	99.64	94.54	91.34	0.6609	0.3705

Representational geometry. For each subject we compute 100×100 representational dissimilarity matrices (RDMs) over the *special100* images for the reconstructions (CLIP ViT-L/14 embeddings), for the viewed images (the same CLIP space and pixel space), and for the fMRI responses (correlation distance over the visual-cortex flatmap). After partialling out both image RDMs, the reconstruction RDM correlates with the subject's own fMRI RDM in all four subjects (partial Spearman $\rho = 0.087$–0.108, image-label permutation $p \leq 0.0002$; with pixel-space reconstruction RDMs $\rho = 0.051$–0.072, $p \leq 0.004$), and more strongly than with the other three subjects' fMRI RDMs ($\Delta\rho = +0.023$, 95% bootstrap CI $[+0.006, +0.042]$; the true subject assignment ranks first among all 24). Before refinement the values are similar ($\rho = 0.099$–0.109; $\Delta\rho = +0.021$, CI $[+0.004, +0.037]$).

Agreement with the measured response. The encoding model is a ridge regression from CLIP ViT-L/14 image features (class token, mean patch token, and 4×4 spatially pooled patch tokens) to the visual-cortex flatmap response, fit on the subject's training images with the penalty chosen on a held-out 10% of them; it is independent of every reconstruction model. For each measured test response we rank the predictions from all 982 reconstructions by Pearson correlation (response → reconstruction), and for each reconstruction we rank all 982 measured responses (reconstruction → response); two-way identification is the fraction of the 981 competitors that are beaten. Differences are tested with a paired bootstrap over test images (2,000 resamples, two-sided); Table 7 shows that the advantage holds at every stage of the pipeline.

Table 7: MindLoop minus the plain decoder without the three constraints (subject 1, 982 test images), in percentage points with 95% bootstrap CIs.

	982-way identification		Two-way identification	
Stage	response → recon.	recon. → response	response → recon.	recon. → response
Before refinement	$+7.43$ [$+3.87$, $+10.90$]	$+2.85$ [$+0.61$, $+5.09$]	$+0.37$ [$+0.09$, $+0.67$]	$+0.60$ [$+0.26$, $+0.93$]
After refinement	$+7.64$ [$+3.97$, $+11.20$]	$+3.36$ [$+1.12$, $+5.70$]	$+0.49$ [$+0.19$, $+0.81$]	$+0.63$ [$+0.30$, $+0.98$]
Final output	$+6.01$ [$+2.34$, $+9.57$]	$+2.55$ [$+0.31$, $+4.79$]	$+0.44$ [$+0.15$, $+0.77$]	$+0.60$ [$+0.25$, $+0.95$]

Decoded condition. We repeat the analysis on the decoded condition itself, before any generation. Here the encoding model is a kernel ridge regression from the generator's conditioning features (OpenCLIP ViT-bigG/14 image tokens; mean token and 4×4 spatially pooled tokens) to the visual-cortex response, fit on the training images; each model's decoded condition for the 982 test responses is projected with it. The full model identifies the matching condition for 87.7% of measured responses versus 80.1% for the plain decoder ($p < 0.001$), and the matching response for 40.3% of conditions versus 36.2% ($p = 0.001$); two-way identification improves in both directions ($p \leq 0.011$). With coarser 2×2 pooling the corresponding values are 84.2% vs. 71.1% and 34.7% vs. 31.0% ($p \leq 0.003$).

Cortical fields in all four subjects. Figure 7 shows the decoder cortical-field centers of MindLoop-SDXL for every subject. In each subject, queries from the left and right halves of the image occupy opposite sides of the flatmap, and queries from the upper and lower halves separate along the other

axis; the correlations are those reported in Section 4.4. For the main model of each subject, we read the learned field centers (μ_x, μ_y) of the 255 queries that carry a cortical field (the 16×16 image patches except the first, which carries no field) and correlate them with the image coordinates of the corresponding patches; significance is assessed with a permutation test (2,000 permutations).

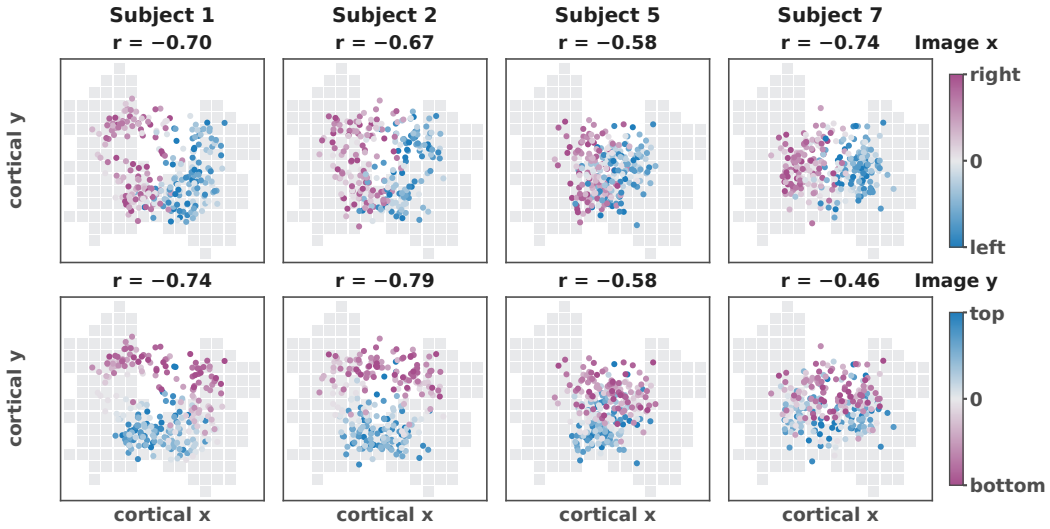

Figure 7: Decoder cortical-field centers of MindLoop-SDXL in all four subjects. Top: queries colored by their horizontal image position, with the correlation between image x and cortical μ_x. Bottom: colored by vertical image position, with the correlation between image y and cortical μ_y. Each panel shows the 255 queries that carry a cortical field; gray squares mark the visual-cortex flatmap. Permutation $p < 0.001$ in every subject (2,000 permutations).

Flip control for both components. Figure 8 quantifies the flip control of Fig. 4: the query image-x of the decoder fields against their cortical μ_x, and the same switch in the effective fields read from the unconstrained attention of the Brain-LBP predictor.

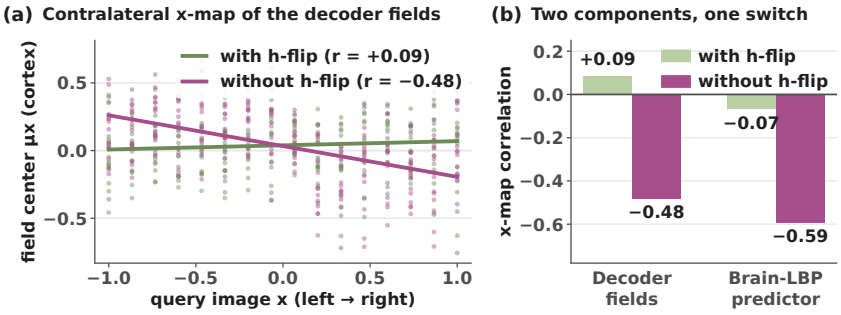

Figure 8: Quantification of Fig. 4 (subject 1, VD-line models differing only in horizontal-flip augmentation). (a) Query image-x versus learned cortical μ_x, with linear fits. (b) x-map correlation of the decoder fields and of the Brain-LBP predictor's effective fields, with and without flips.