Brain-Measurable Diffusion Decoding: Auditing Information Provenance in fMRI Reconstruction

Firstname Lastname
first.last@university.edu
University of Example

Firstname Lastname
first.last@company.com
Example AI Lab

Firstname Lastname
first.last@institute.edu
Institute of Example

Abstract

Diffusion models are increasingly used as scientific reconstruction engines, but realistic samples do not by themselves reveal where target-specific information entered the system. This ambiguity is acute in fMRI-to-image reconstruction: a frozen image generator can amplify weak neural conditioning into plausible scenes, while target captions, target-image features, retrieval, or candidate-pool selection can silently turn reconstruction into an oracle-assisted protocol. We formulate this issue as an information-provenance problem and define *brain-measurable diffusion decoding*: every inference-time control supplied to the generator must be a function of the measured fMRI response, fixed learned weights, fixed sampler settings, and target-independent randomness. We instantiate the criterion with a three-channel fMRI-conditioned decoder that predicts a Stable Diffusion VAE latent, learns a CLIP-aligned visual coordinate during training, and supplies brain-derived cross-attention tokens to a frozen SD1.5 denoiser. On 982 NSD shared-test images, the method reaches competitive standard-8 reconstruction quality under a clean train-only, single-rendering protocol. Provenance ablations, sampler-strength sweeps, learning-dynamics traces, calibration checks, and resource audits identify which controls carry stimulus-specific information and which evaluation choices can alter the scientific claim. The result is both a reconstruction system and an evaluation template for separating brain-derived generation from target-side shortcuts in scientific uses of deep generative models.

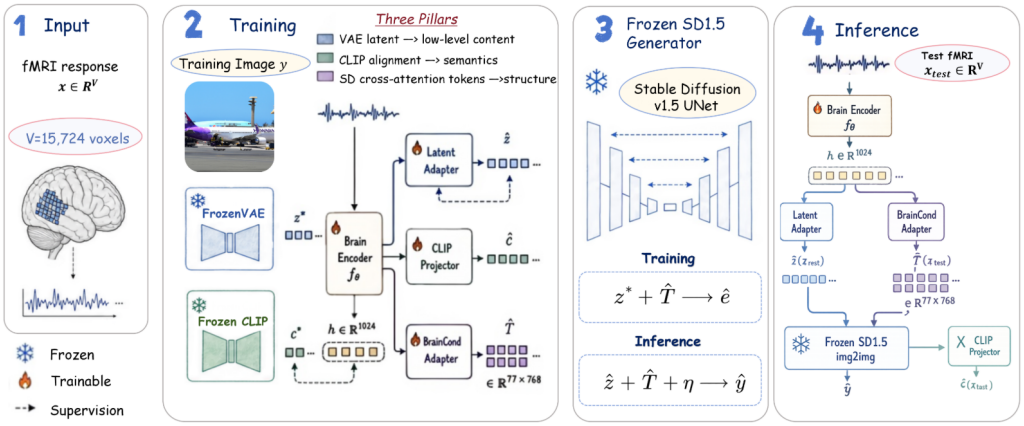

Figure 1. **Brain-measurable training and inference pipeline.** Training uses frozen visual teachers to supervise brain-predicted VAE latents, CLIP-aligned visual coordinates, and SD cross-attention tokens. Inference keeps the SD1.5 generator frozen and supplies only fMRI-derived latent initialization and fMRI-derived tokens, with no target caption, oracle retrieval image, target CLIP feature, or evaluation-pool candidate.

Accepted to FoGen 2026: Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning, an ICML 2026 workshop (non-archival).

1. Introduction

Deep generative models (DGMs) have moved from content creation into scientific workflows. Diffusion models now act as priors, decoders, simulators, and inverse-problem solvers when measurements are noisy, incomplete, or indirect (Ho et al., 2020; Song et al., 2021; Nichol & Dhariwal, 2021; Dhariwal & Nichol, 2021; Rombach et al., 2022). This creates a productive tension. A strong generator can convert weak scientific evidence into a structured sample, but the same generator can also conceal whether the decisive information came from the measurement, from memorized training structure, from retrieval, or from target-side metadata (Somepalli et al., 2023; Carlini et al., 2023).

We study this issue in fMRI-to-image reconstruction. The Natural Scenes Dataset (NSD) provides repeated human brain responses to natural images and has enabled a wave of diffusion-based reconstruction systems (Allen et al., 2022; Takagi & Nishimoto, 2023; Ozcelik & VanRullen, 2023; Lin et al., 2022; Lu et al., 2023; Chen et al., 2023; Scotti et al., 2023; 2024; Xia et al., 2024a;b; Wang et al., 2024; Huo et al., 2024; Gong et al., 2025; Beliy et al., 2025; Zhang et al., 2025). These methods are scientifically valuable because they suggest that visual information can be decoded from non-invasive brain measurements. They are also methodologically delicate: if a pipeline uses target captions, target-image CLIP features, evaluation-pool retrieval, or target-guided best-of-N selection, high visual quality no longer establishes that the reconstruction was determined by fMRI.

This paper places the information interface at the center of the reconstruction problem. The FoGen workshop asks when a generative model is memorizing, generalizing, reasoning, or exploiting evaluation artifacts. fMRI reconstruction is a concrete domain where that question becomes directly testable. A reconstruction can be semantically plausible yet scientifically uninformative if the generator was guided by a target-side oracle. Conversely, a stricter single-rendering protocol may produce less polished images but support a clearer claim: the target-specific evidence entered through the measured brain response.

We call a reconstruction rule *brain-measurable* if every deployed conditioning variable is measurable from the fMRI response, fixed learned weights, fixed sampler settings, and target-independent randomness. The definition allows frozen foundation models as priors and teachers. CLIP image embeddings (Radford et al., 2021), VAE latents (Rombach et al., 2022), or diffusion noise targets may supervise training, exactly as labels supervise ordinary prediction. The corresponding test-time controls, however, must be predicted from fMRI rather than copied from the held-out target image.

We instantiate the criterion with a three-channel fMRI-conditioned diffusion decoder. The model predicts (i) a Stable Diffusion VAE latent for low-level initialization, (ii) a CLIP-aligned visual coordinate for training-time semantic alignment, and (iii) a tensor of brain-derived SD1.5 cross-attention tokens that conditions a frozen UNet. At inference, the generator receives only the fMRI-predicted latent and fMRI-predicted tokens. It receives no target caption, target-image feature, target-derived prompt, retrieval result, candidate-pool reranking signal, or best-of-N selector.

Contributions.

- **Information-provenance criterion.** We formalize brain-measurable diffusion decoding and distinguish training-time visual teachers from inference-time oracle information.
- **A concrete brain-measurable decoder.** We implement the criterion with three fMRI-derived coordinates: a VAE latent, a CLIP-supervised visual coordinate, and brain-derived cross-attention tokens for a frozen SD1.5 denoiser.
- **An audited NSD evaluation.** We use a strict train-only split that removes all official shared-image identities before training, evaluate on 982 valid shared-test images, and report standard-8 metrics under a fixed single-rendering protocol.
- **Diagnostics for DGM behavior.** We report resource audits, full provenance ablations, sampler-strength sweeps, learning-dynamics traces, fixed-calibration checks, and representation-side probes in the main paper rather than hiding the evidence in the appendix.

2. Related Work and FoGen Positioning

Neural decoding before diffusion. Visual decoding has a long history in computational neuroscience. Early work identified images from neural responses or reconstructed images by combining brain likelihoods with image priors

(Kay et al., 2008; Naselaris et al., 2009; Nishimoto et al., 2011; Naselaris et al., 2011; Haxby et al., 2001; Horikawa & Kamitani, 2017). Later approaches used CNN representations, self-supervised objectives, GANs, and learned image priors to improve semantic and visual fidelity (Wen et al., 2018; Seeliger et al., 2018; Shen et al., 2019; Beliy et al., 2019; Gaziv et al., 2022; Goodfellow et al., 2014; Karras et al., 2019). These studies motivate fMRI reconstruction as a scientific inverse problem, but they predate the current evaluation challenge: a very strong frozen generator can make weak or ambiguous conditioning look convincing.

Diffusion-based fMRI reconstruction. Latent diffusion models provide a high-quality natural-image prior and a compact latent space (Rombach et al., 2022). Recent fMRI reconstruction methods connect brain responses to latents, CLIP embeddings, diffusion priors, captions, retrieval systems, or multi-level modulation (Takagi & Nishimoto, 2023; Ozcelik & VanRullen, 2023; Lu et al., 2023; Chen et al., 2023; Scotti et al., 2023; 2024; Xia et al., 2024a;b; Wang et al., 2024; Huo et al., 2024; Gong et al., 2025; Beliy et al., 2025; Zhang et al., 2025). This line has made reconstructions more realistic. Our contribution is complementary: we ask whether the inference-time controls supplied to the generator are brain-measurable, and we audit which control channel actually carries stimulus-specific information.

Foundation-model controls and leakage. Modern image generation often improves controllability through text conditioning, CLIP latents, adapters, ControlNet-like structural inputs, and multimodal supervision (Radford et al., 2021; Jia et al., 2021; Li et al., 2022; Liu et al., 2023; Nichol et al., 2022; Ramesh et al., 2022; Saharia et al., 2022; Podell et al., 2024; Zhang et al., 2023; Mou et al., 2024; Xu et al., 2023). These tools are powerful and appropriate in many generation tasks. In scientific reconstruction, however, the provenance of each control matters. A target-derived caption, CLIP feature, or retrieval result can turn a decoding claim into a target-assisted generation claim. This is why the same visual foundation model can be a legitimate training teacher and an invalid test-time oracle depending on when and how it is used.

FoGen relevance. FoGen emphasizes memorization, generalization, reasoning, optimization, and evaluation of DGMs. fMRI reconstruction connects these themes in one application. The frozen diffusion model contributes distributional knowledge; the brain encoder must generalize from noisy measurements; the sampler and selector define the decision rule; and the reported score is meaningful only if the information pathway is clear. Our experiments therefore treat reconstruction quality and provenance diagnostics as inseparable.

3. Brain-Measurable Diffusion Decoding

3.1. Information Available at Inference

Let (X, Y) denote a paired fMRI response and stimulus image. Let $\mathcal{G}_\phi$ be a frozen stochastic generator with fixed sampler settings ϕ, such as a latent diffusion model (Ho et al., 2020; Song et al., 2021; Rombach et al., 2022). A reconstruction has the form

$$\hat{Y} = \mathcal{G}_\phi(C, \eta), \tag{1}$$

where C is the conditioning object and η is generator noise. We say a decoder is *brain-measurable* if every deployed conditioning variable is measurable with respect to

$$\mathcal{I}_{\text{brain}}(X) = \sigma(X, \theta, \phi, \xi, \eta), \tag{2}$$

where θ are fixed learned weights and ξ, η are target-independent sources of randomness. Equivalently, there exists a fixed map a_θ such that

$$C = a_\theta(X, \xi), \qquad C, \hat{Y} \text{ are } \mathcal{I}_{\text{brain}}(X)\text{-measurable}. \tag{3}$$

Information computed from the held-out target image is outside this interface. This includes target captions, target-image CLIP features, depth or structural maps, VLM descriptions, retrieval labels, and candidate-pool scores derived from Y or from the evaluation set $\mathcal{E}_{\text{test}}$. Such signals may improve image quality, but they change the scientific claim from brain decoding to oracle-assisted generation.

3.2. Teachers Versus Oracles

The provenance constraint applies to variables used at inference, not to supervised targets used during training. A frozen teacher may define a target representation

$$T(Y), \qquad \theta^\star = \arg\min_\theta \mathcal{L}_{\text{teach}}(\theta),$$
$$\mathcal{L}_{\text{teach}}(\theta) = \mathbb{E}_{(X,Y)\sim\mathcal{D}_{\text{train}}}\big[d(a_\theta(X), T(Y))\big], \tag{4}$$

as in empirical risk minimization (Vapnik, 1998). The deployed system remains brain-measurable only if the test-time control is predicted from fMRI:

$$\begin{aligned} C_{\text{test}} &= a_{\theta^\star}(X_{\text{test}}) && \text{is allowed,} \\ C_{\text{test}} &= T(Y_{\text{test}}) && \text{is an oracle.} \end{aligned} \tag{5}$$

Thus CLIP image embeddings, VAE latents, VQ codes (van den Oord et al., 2017; Esser et al., 2021), and diffusion-noise targets can be legitimate teachers, provided their deployed counterparts are inferred from X_{test}.

3.3. Sampling, Selection, and Leakage

A stochastic decoder can also be viewed as a decision rule (Berger, 1985) $\delta = (g_\theta, q, N)$: it generates N candidates from a brain-derived representation $g_\theta(X)$ and selects one using a score q. If q depends on the target image, target caption, target CLIP feature, or evaluation-pool retrieval, the reported score estimates a different protocol. A best-of-N system can be scientifically meaningful if the selector is itself brain-measurable, but it is not equivalent to a single-rendering decoder.

Our main evaluation uses $N = 1$ with a vacuous selector. This deliberately gives up best-of-N gains, but it removes selection leakage by construction. Appendix A records the longer decision-rule decomposition into representation, sampling, and selection-alignment terms.

3.4. Relation to Memorization and Generalization

Brain-measurability does not claim that the frozen generator has no memory of its image pretraining distribution. In fact, the natural-image prior is the reason weak fMRI controls can produce coherent samples. The claim is narrower and more auditable: for a held-out stimulus, target-specific information must enter through the measured response rather than a target-side shortcut. This separates useful distributional prior knowledge from hidden leakage, and aligns fMRI reconstruction with FoGen's broader focus on memorization, generalization, reasoning, and evaluation.

4. Method

4.1. Problem Setup

Let $\mathbf{x} \in \mathbb{R}^V$ denote an fMRI response and $\mathbf{y}^\star$ the corresponding stimulus image. In the fully audited NSD setting reported here, $V = 15{,}724$ after applying the subject-specific `nsdgeneral` mask. The goal is to generate $\hat{\mathbf{y}}$ such that all deployed controls are functions of $\mathbf{x}$. We implement

$$\hat{\mathbf{y}} = \mathcal{G}_{\text{SD},\phi}\left(\hat{\mathbf{z}}(\mathbf{x}), \hat{\mathbf{T}}(\mathbf{x}), \eta\right), \tag{6}$$

where $\hat{\mathbf{z}}(\mathbf{x})$ is a brain-predicted latent initialization, $\hat{\mathbf{T}}(\mathbf{x})$ is a brain-predicted cross-attention tensor, and η is diffusion sampling noise. Figure 2 summarizes the training and inference paths and highlights the key constraint: frozen visual teachers supervise training, but only fMRI-predicted controls are deployed at test time.

4.2. Teacher Coordinates

For each training image, we precompute two frozen teacher coordinates. The SD1.5 VAE encodes the image into a scaled latent $\mathbf{z}^\star \in \mathbb{R}^{4\times32\times32}$ at 256×256 resolution. CLIP ViT-B/32 encodes the same image into a normalized visual embedding $\mathbf{c}^\star \in \mathbb{R}^{512}$ (Radford et al., 2021). Neither teacher is updated. More importantly, neither $\mathbf{z}^\star(Y_{\text{test}})$ nor $\mathbf{c}^\star(Y_{\text{test}})$ is available during test-time generation.

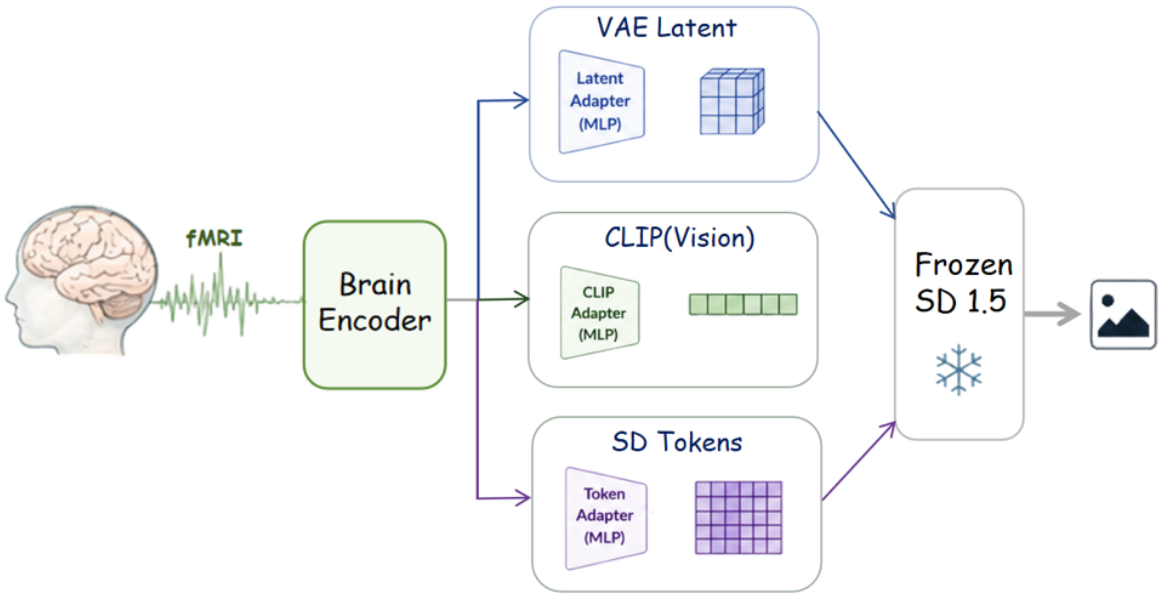

Figure 2. **Overview of the brain-measurable fMRI-conditioned diffusion framework.** The model learns three fMRI-derived pathways: a VAE latent, a CLIP-aligned visual coordinate, and SD1.5 cross-attention tokens. At inference time, CLIP serves strictly as a training-time teacher; the generator is conditioned only on the fMRI-predicted latent and tokens, without target captions, oracle retrieval images, or target-image features.

4.3. Phase I: Brain-to-Visual Coordinates

Phase I learns a brain representation $\mathbf{h} = f_\theta(\mathbf{x}) \in \mathbb{R}^{1024}$ using a lightweight MLP encoder with LayerNorm (Ba et al., 2016), GELU (Hendrycks & Gimpel, 2016), dropout (Srivastava et al., 2014), and low-rank adaptation on selected linear layers (Hu et al., 2022). Two heads predict a latent and a visual coordinate:

$$\hat{\mathbf{z}} = f_z(\mathbf{h}) \in \mathbb{R}^{4\times32\times32}, \qquad \hat{\mathbf{c}} = f_c(\mathbf{h}) \in \mathbb{R}^{512}. \tag{7}$$

The objective is

$$\begin{aligned} \mathcal{L}_{\mathrm{I}} &= \|\hat{\mathbf{z}} - \mathbf{z}^\star\|_2^2 + \lambda_{\mathrm{clip}}\mathcal{L}_{\mathrm{vis}}, \\ \mathcal{L}_{\mathrm{vis}} &= 1 - \cos(\hat{\mathbf{c}}, \mathbf{c}^\star), \qquad \lambda_{\mathrm{clip}} = 0.5. \end{aligned} \tag{8}$$

We train Phase I for 20 epochs with AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019), batch size 16, learning rate 10^{-4}, weight decay 0.01, mixed precision, and gradient clipping at 1.0.

4.4. Phase II: Brain-Derived Diffusion Tokens

Phase II freezes the Phase-I brain encoder and learns a diffusion-conditioning tensor. A BrainCondAdapter $A_\psi : \mathbb{R}^{1024} \to \mathbb{R}^{77\times768}$ maps the brain representation to cross-attention tokens $\mathbf{T} = A_\psi(\mathbf{h})$. These tokens replace the text-encoder hidden states normally consumed by the frozen SD1.5 UNet. A lightweight CLIPAlignHead H_ψ mean-pools the tokens and predicts a CLIP image embedding for an auxiliary alignment loss.

The tokens are not generated from text; they are learned by backpropagating through the frozen denoising objective:

$$\hat{\boldsymbol{\epsilon}} = U_{\mathrm{SD}}(\mathbf{z}_t, t, \mathbf{T}), \qquad \mathbf{z}_t = \mathrm{noise}(\mathbf{z}^\star, t), \tag{9}$$

$$\begin{aligned} \mathcal{L}_{\mathrm{II}} &= \|\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}\|_2^2 + \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{clip}}, \\ \mathcal{L}_{\mathrm{clip}} &= 1 - \cos(H_\psi(\mathbf{T}), \mathbf{c}^\star), \qquad \lambda_{\mathrm{align}} = 1.0. \end{aligned} \tag{10}$$

The SD1.5 VAE, UNet, tokenizer, and text encoder remain frozen. The method therefore learns a brain-to-conditioning interface rather than fine-tuning the image prior itself.

4.5. Inference and Auditability

At test time, the same fMRI vector produces $\hat{\mathbf{z}}(\mathbf{x})$ and $\hat{\mathbf{T}}(\mathbf{x})$. DDIM image-to-image sampling denoises from $\hat{\mathbf{z}}(\mathbf{x})$ under brain-token cross-attention. The empty prompt branch is used only for classifier-free guidance (Ho & Salimans, 2022);

it does not inject target-image content. We use a single rendering per test image ($N = 1$), so there is no target-guided candidate selection.

The three-channel design makes information provenance testable. If the frozen diffusion prior alone determined the output, mismatching fMRI responses would not collapse identification metrics. If the VAE latent carried object identity by itself, latent-only decoding would remain strong. If the token sequence were decorative, token-only decoding would fail. These counterfactuals correspond directly to the ablations in Section 6.3.

Additional implementation details and the corresponding provenance rationale are provided in Appendix B and Appendix D.

5. Evaluation Protocol

Dataset and split. We evaluate on NSD (Allen et al., 2022) under a subject-specific `nsdgeneral` mask, using the canonical subject-01 split for the fully audited tables in this workshop version. The original pre-extracted tensor contains 29,997 fMRI–image pairs from 9,999 unique NSD image ids. The official shared-test set contains 1,000 unique image ids, each repeated three times. Our clean train-only tensor removes every row whose NSD id belongs to this shared set, leaving 26,997 rows from 8,999 unique image ids and zero identity overlap with evaluation images. The main evaluation uses 982 valid shared-test samples.

Metrics. We report the standard-8 reconstruction suite used by recent NSD decoding work: PixCorr, SSIM (Wang et al., 2004), AlexNet-2 and AlexNet-5 two-way identification (Krizhevsky et al., 2012), Inception two-way identification (Szegedy et al., 2016), CLIP two-way identification (Radford et al., 2021), EfficientNet distance (Tan & Le, 2019), and SwAV distance (Caron et al., 2020). Higher is better for PixCorr, SSIM, AlexNet, Inception, and CLIP; lower is better for EfficientNet and SwAV.

Fixed operating point. The headline benchmark uses DDIM image-to-image sampling with fMRI-predicted latent initialization and fMRI-predicted cross-attention tokens: 30 denoising steps, strength 0.46, guidance scale 5.0, seed 9, and one global target-independent blur/contrast/brightness calibration. These settings are applied identically to every reconstruction and do not depend on target images, target captions, retrieval outputs, candidate-pool reranking, or per-image metric feedback. Sampler and calibration sweeps are reported separately as diagnostics.

Appendix B records additional training, inference, and reporting details that are not essential to follow the main evidence.

6. Experiments

The experiments are organized around the information-provenance claim. Section 6.1 reports the clean benchmark. Section 6.2 audits resources used by comparable protocols. Section 6.3 tests whether outputs follow measured fMRI. Sections 6.4–6.6 analyze sampler behavior, learning dynamics, fixed calibration, and representation-side structure.

6.1. Main Benchmark

Table 1 shows that the brain-measurable decoder reaches competitive reconstruction quality under a stricter information contract. The method leads the comparison on PixCorr, CLIP, EfficientNet distance, and SwAV distance, and remains competitive on Inception and AlexNet identification. SSIM is not the strongest row, reflecting the known tension between exact local structure and diffusion-based perceptual realism. The important point is that our result is achieved by a single reconstruction per fMRI response under fixed settings; no target-guided sample selection is used. Qualitative examples in Figures 3 and 4 show that the same protocol preserves coarse object category and scene layout while leaving fine texture and geometry imperfect.

6.2. Resource and Provenance Audit

External baselines vary along multiple axes: subject regime, text supervision, retrieval, candidate selection, and metric implementation. Table 2 makes the main resource differences explicit. Cross-subject training and extra image pretraining are legitimate modeling choices, but they change the regime. Target-side captions, target-image features,

Table 1. **Quantitative comparison on the NSD shared-test set.** We report standard-8 reconstruction metrics. PixCorr and SSIM measure low-level similarity; AlexNet-2, AlexNet-5, Inception, and CLIP are two-way identification accuracies; EfficientNet and SwAV are correlation distances. External rows are quoted from original publications. Our row uses 982 valid shared-test images under a clean train-only, single-rendering, fixed-calibration protocol with no target captions, retrieval pool, candidate reranking, or target-image features at inference.

Method	Pix.↑	SSIM↑	Alex-2↑	Alex-5↑	Inc.↑	CLIP↑	Eff.↓	SwAV↓
Takagi–Nishimoto LDM (Takagi & Nishimoto, 2023)	0.246	0.410	0.789	0.856	0.838	0.821	0.811	0.504
Brain-Diffuser (Ozcelik & VanRullen, 2023)	0.273	0.365	0.944	0.966	0.913	0.909	0.728	0.421
MindEye (Scotti et al., 2023)	0.319	0.360	0.928	0.969	0.946	0.933	0.648	0.377
DREAM (Xia et al., 2024a)	0.274	0.328	0.939	0.967	0.934	0.941	0.645	0.418
UMBRAE (Xia et al., 2024b)	0.283	0.341	0.955	0.970	0.917	0.935	0.700	0.393
MindBridge (Wang et al., 2024)	0.151	0.263	0.877	0.955	0.924	0.947	0.712	0.418
NeuroPictor (Huo et al., 2024)	0.229	0.375	0.965	0.984	0.945	0.933	0.639	0.350
MindEye2 (Scotti et al., 2024)	0.322	0.431	0.961	0.986	0.954	0.930	0.619	0.344
MindTuner (Gong et al., 2025)	0.322	0.421	0.958	0.988	0.956	0.938	0.612	0.340
Brain-IT (Beliy et al., 2025)	0.386	**0.486**	**0.984**	**0.995**	0.973	0.964	0.564	0.320
Ours	**0.408**	0.427	0.982	**0.995**	**0.981**	**0.991**	**0.529**	**0.308**

Figure 3. **Qualitative reconstructions on the clean train-only split.** Each panel shows ground-truth images and the corresponding reconstructions generated with the same single-rendering protocol. The comparison illustrates the PixCorr–SSIM trade-off introduced by fixed, target-independent post-processing.

Table 2. **Resource audit for recent NSD reconstruction methods.** A check mark indicates reported use of the corresponding resource. The goal is not to penalize richer protocols, but to separate them from the brain-measurable claim made here.

Method	Other subjects	Caption/VLM training	Target-side caption/feature	Oracle rerank
MindEye (Scotti et al., 2023)	–	–	–	✓
MindEye2 (Scotti et al., 2024)	✓	–	–	–
NeuroPictor (Huo et al., 2024)	✓	✓	–	–
MindBridge (Wang et al., 2024)	✓	–	–	–
MindTuner (Gong et al., 2025)	✓	–	–	–
Ours	–	–	–	–

and oracle reranking are more directly relevant to the provenance claim because they can supply held-out stimulus information at inference.

Figure 4. **Paired reconstruction view.** Representative held-out viewed images and corresponding reconstructions are shown under the fixed protocol. This is not a best-of-N display; no target-guided candidate selection is used.

Table 3. **Full information-provenance ablation on the clean train-only split.** Values are mean±std over independent seeds at a fixed diagnostic operating point. Mismatch fMRI replaces the measured response with a different test response and collapses identification metrics to chance.

Condition	Pix.↑	SSIM↑	Alex-2↑	Alex-5↑	Inc.↑	CLIP↑	Eff.↓	SwAV↓
Combined latent+tokens	0.536 ± 0.000	0.311 ± 0.001	0.996 ± 0.001	0.999 ± 0.000	0.997 ± 0.001	0.999 ± 0.000	0.308 ± 0.001	0.143 ± 0.001
Token only	0.631 ± 0.002	0.345 ± 0.001	0.998 ± 0.000	0.999 ± 0.000	0.998 ± 0.001	0.999 ± 0.000	0.280 ± 0.001	0.130 ± 0.000
Latent only	0.050 ± 0.002	0.062 ± 0.001	0.517 ± 0.003	0.516 ± 0.004	0.502 ± 0.005	0.501 ± 0.004	0.998 ± 0.000	0.732 ± 0.001
Mismatch fMRI	0.037 ± 0.002	0.174 ± 0.001	0.497 ± 0.002	0.493 ± 0.001	0.498 ± 0.002	0.497 ± 0.002	0.981 ± 0.001	0.646 ± 0.001

Table 4. **Full sampler-strength audit for the combined latent-plus-token pathway.** The same trained model and evaluator are used for all rows. Increasing strength monotonically improves the combined pathway and brings it close to the token-only operating point.

Condition	Strength	Seeds	Pix.↑	SSIM↑	Eff.↓	SwAV↓
Combined	0.45	3	0.4022 ± 0.0018	0.2870 ± 0.0010	0.403	0.199
Combined	0.55	3	0.4742 ± 0.0016	0.2958 ± 0.0006	0.338	0.159
Combined	0.65	3	0.5361 ± 0.0003	0.3109 ± 0.0006	0.308	0.143
Combined	0.75	3	0.5814 ± 0.0012	0.3256 ± 0.0003	0.292	0.136
Combined	0.85	3	0.6112 ± 0.0001	0.3379 ± 0.0003	0.285	0.133
Combined	0.90	3	0.6244 ± 0.0008	0.3431 ± 0.0007	0.282	0.131
Combined	0.95	3	0.6293 ± 0.0017	0.3450 ± 0.0009	0.282	0.130
Combined	1.00	2	0.6332 ± 0.0011	0.3469 ± 0.0000	0.280	0.130
Token only	0.65	3	0.6306 ± 0.0020	0.3454 ± 0.0011	0.280	0.130

6.3. Information-Provenance Ablation

The central audit is whether generated images follow the measured fMRI response. Table 3 provides the full eight-metric ablation in the main text, and Figure 5 visualizes the same footprint across normalized metrics. This diagnostic uses a fixed ablation operating point for channel comparisons, so the absolute values should be read separately from the headline row in Table 1. Mismatching the fMRI response collapses two-way identification to chance, confirming that the frozen generator and sampler do not determine the viewed stimulus alone. Latent-only conditioning also collapses, indicating that the fMRI-predicted VAE latent provides a coarse anchor but not object identity. Token-only conditioning recovers most of the reconstruction signal, identifying brain-derived cross-attention tokens as the dominant generative channel.

6.4. Sampler-Strength Audit

The combined latent-plus-token pathway depends on the sampler operating point. DDIM image-to-image strength controls how much the denoising trajectory can move away from the initial latent. At low strength, the predicted latent can overconstrain the generation and suppress token information. At higher strength, the sampler relies more strongly on brain-derived tokens. Table 4 and Figure 6 show a monotonic improvement from strength 0.45 to 1.00, converging toward the token-only operating point. This suggests that the latent pathway acts as a structural initialization whose measured value depends on sampler calibration.

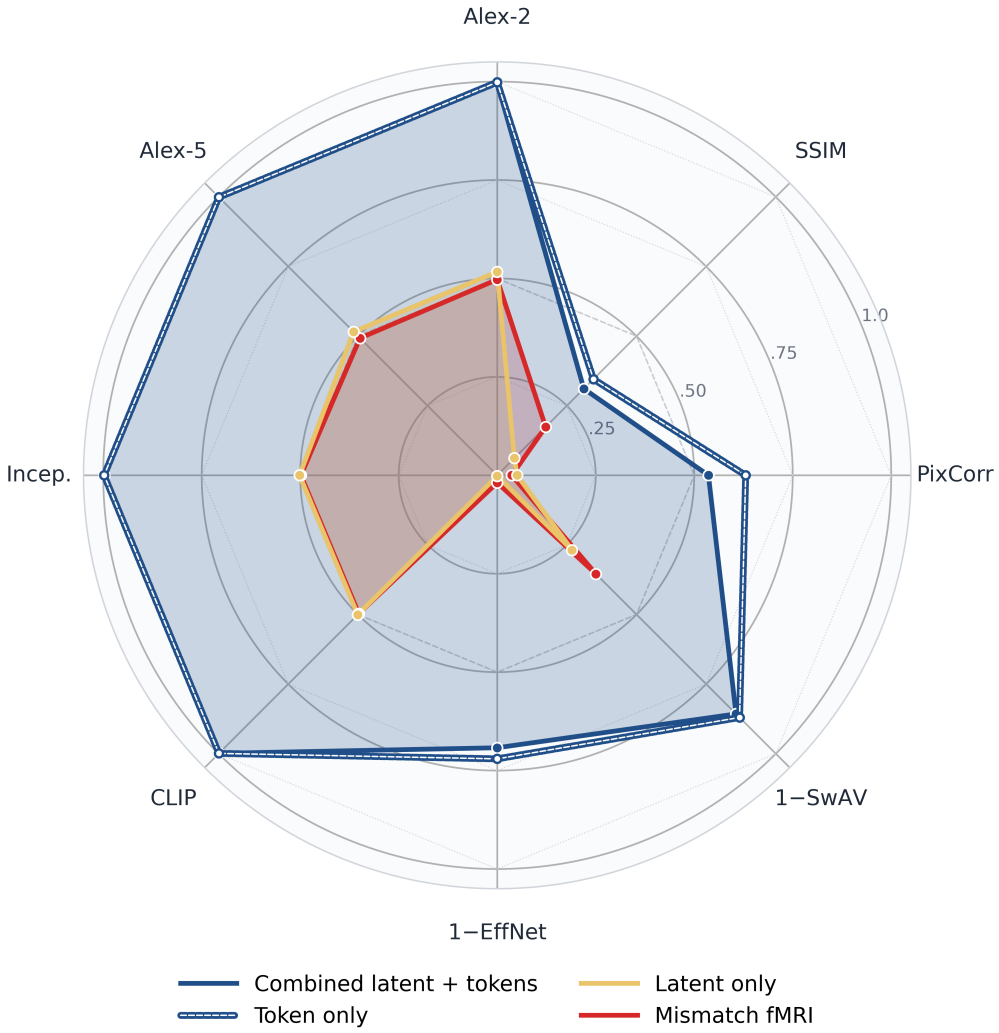

Figure 5. **Provenance-ablation footprint.** Each axis is a standard reconstruction metric; perceptual distances are inverted for visualization. Mismatch fMRI and latent-only collapse, while token conditioning preserves the main target-specific signal.

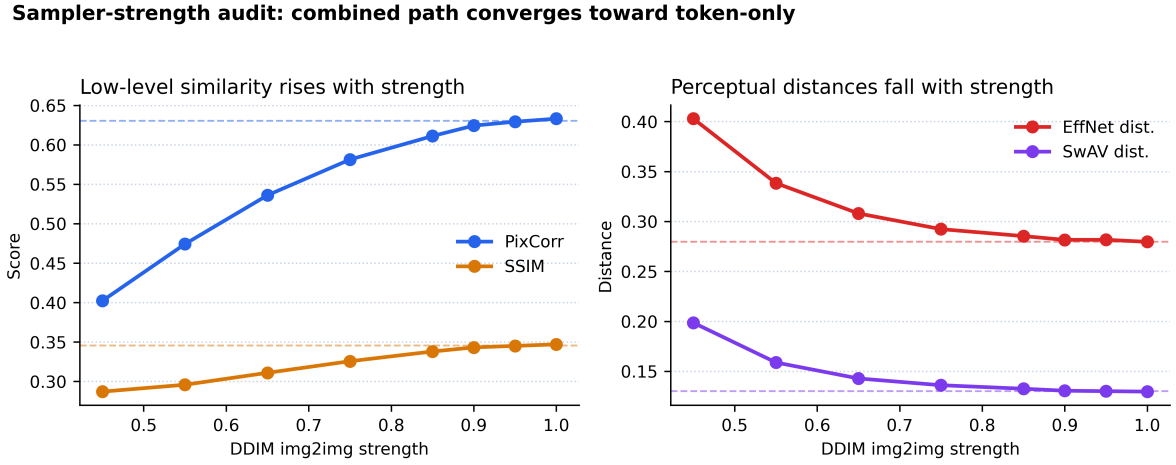

Figure 6. **Sampler-strength diagnostic.** Increasing image-to-image strength changes how much the frozen denoiser follows brain-derived tokens rather than the initial latent.

Table 5. **Extended learning-dynamics trace.** Each row is a fixed-sampler standard-8 evaluation at a different Phase-II checkpoint. Identification metrics approach ceiling early, while low-level and perceptual metrics continue improving.

Time	Pix.↑	SSIM↑	Alex-2↑	Alex-5↑	Inc.↑	CLIP↑	Eff.↓	SwAV↓
30m	0.156	0.279	0.630	0.705	0.702	0.773	0.886	0.544
1h	0.168	0.255	0.787	0.905	0.927	0.975	0.689	0.385
5h	0.203	0.251	0.888	0.964	0.977	0.993	0.560	0.292
10h	0.297	0.259	0.971	0.993	0.990	0.997	0.456	0.223
15h	0.431	0.280	0.992	0.998	0.996	0.998	0.369	0.173

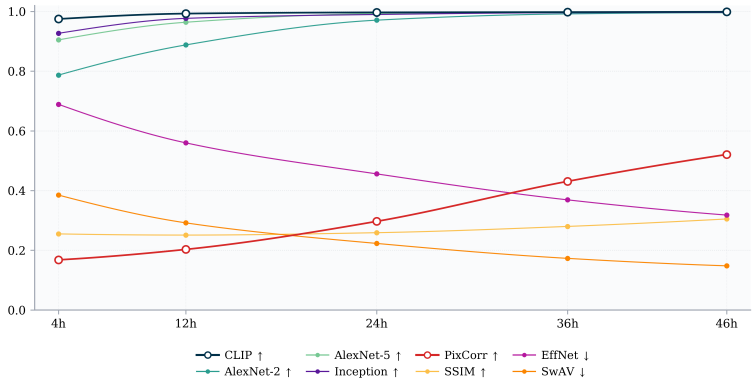

Figure 7. **Learning-dynamics diagnostic.** The larger two-column layout makes the epoch-level trends readable. Semantic identification metrics saturate early, while low-level and perceptual metrics continue improving under the same fixed sampler.

Table 6. **Fixed operating-point calibration and seed stability.** All rows use 982 shared-test images. Calibration parameters are global and target-independent.

Setting	Pix.↑	SSIM↑	Alex-2↑	Alex-5↑	Inc.↑	CLIP↑	Eff.↓	SwAV↓
Selected calibrated operating point	0.423	0.336	0.956	0.987	0.964	0.988	0.609	0.229
Balanced operating region, 3 seeds	0.431 ± 0.002	0.337 ± 0.001	0.955 ± 0.001	0.985 ± 0.000	0.961 ± 0.001	0.986 ± 0.001	0.619 ± 0.001	0.230 ± 0.001

Table 7. **Representation-side VAE/low-frequency probe.** Lower-resolution VAE/low-frequency targets become increasingly predictable from fMRI, supporting the role of the latent channel as a coarse structural anchor rather than a complete semantic generator.

Target	Hidden	Val MSE/dim ↓	Val cosine ↑
full 32×32	1024	0.637	0.128
pooled 16×16	4096	0.387	0.156
pooled 8×8	1024	0.290	0.178
pooled 4×4	4096	0.222	0.204
pooled 2×2	4096	0.156	0.233

6.5. Learning Dynamics

We examine whether the brain-to-diffusion interface improves with training time. Table 5 and the two-column Figure 7 show diagnostic traces under a fixed sampler; Appendix C provides a matched qualitative progression over training time. High-level two-way identification metrics approach ceiling early, while PixCorr, SSIM, EfficientNet, and SwAV continue improving. This pattern is inconsistent with a purely caption-like decoder: after semantic identification saturates, the system continues to refine structural visual information.

6.6. Fixed Calibration and Representation Probe

We also report two practical diagnostics in the main text. First, Table 6 shows that a nearby calibrated operating region is stable across seeds. Calibration parameters are global and target-independent; they are not selected per image. Second, Table 7 tests whether low-frequency latent targets become more predictable from fMRI. The trend supports the interpretation that the latent pathway is a coarse structural anchor rather than a complete semantic generator.

7. Analysis

The frozen prior is necessary but not sufficient. The generator supplies natural-image structure, but the mismatch-fMRI ablation shows that it cannot recover the viewed stimulus without the correct brain response. Replacing the measured response collapses identification metrics to chance. This provides central evidence that target specificity comes directly from fMRI-predicted controls rather than the frozen prior alone.

The token channel carries the main generative signal. The cross-attention sequence serves as the primary path by which visual identity reaches the denoiser. Combined with the structural initialization provided by the latent branch, the two pathways synergize to decode the brain signal, with their exact balance modulated by the sampler strength.

CLIP supervision is not a target oracle. CLIP is used strictly as a training teacher, not a test-time input. The teacher defines a loss, and the deployed predictor must infer the relevant coordinate exclusively from fMRI. Leakage would occur if the target image's CLIP feature, caption, or VLM description were supplied at inference; our deployed generator receives none of these.

Why this matters beyond fMRI. Many scientific DGM applications share a common structure: a weak measurement is fed into a powerful generator, and realism is often mistakenly treated as sufficient evidence of measurement informativeness. Our results emphasize that sample quality must be accompanied by a rigorous information audit. Verifying exactly which variables entered the generator—and whether they were measured, predicted, retrieved, or selected—is essential for any scientific discovery relying on generative models.

8. Discussion

This work reframes fMRI reconstruction as a provenance-controlled evaluation problem. Our brain-measurable decoder achieves competitive performance without using target-side shortcuts like captions, oracle retrieval, or best-of-N selection.

Our ablations confirm that the generated visual identity is strictly driven by the measured brain response, rather than the frozen image prior. Furthermore, diagnostics show that the learned cross-attention tokens and latent initialization synergize effectively to decode the visual signals.

Reproducibility. All experiments rely on the standardized NSD split described in Section 5, ensuring zero identity overlap between training and the 982 shared-test images. The complete training procedure, model architecture (frozen SD1.5 modules), and single-rendering inference settings are fully specified to ensure our provenance claims are transparent and verifiable.

Limitations. To establish a strict baseline, we currently focus on NSD subj01 and forgo best-of-N selection. Natural next steps include expanding to a multi-subject benchmark, incorporating human preference metrics, and further interpreting the internal representations of the frozen UNet.

Conclusion. Brain-measurable diffusion decoding provides a rigorous framework for applying powerful foundation models to scientific reconstruction tasks. By demanding that all target-specific controls be predicted directly from measurements, it yields competitive fMRI reconstructions without confounding evaluation with target-side shortcuts. Ultimately, we demonstrate that generative evaluation should measure not just the visual quality of a sample, but also the true source of its information.

Impact Statement

This work advances the evaluation and interpretability of diffusion models in scientific discovery by making fMRI decoders more auditable. While improved neural decoding can raise mental-privacy concerns, our system requires subject-specific fMRI calibration and a controlled stimulus protocol, and is not a deployable mind-reading tool.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., and Kay, K. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 2022.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Beliy, R., Gaziv, G., Hoogi, A., Strappini, F., Golan, T., and Irani, M. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fMRI. In *Advances in Neural Information Processing Systems*, 2019.

Beliy, R., Zalcher, A., Kogman, J., Wasserman, N., and Irani, M. Brain-IT: Image reconstruction from fMRI via brain-interaction transformer. *arXiv preprint arXiv:2510.25976*, 2025.

Berger, J. O. *Statistical Decision Theory and Bayesian Analysis*. Springer, 1985.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. In *USENIX Security Symposium*, 2023.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, 2020.

Chen, Z., Qing, J., Xiang, T., Yue, W. L., and Zhou, J. H. 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*, 2023.

Dhariwal, P. and Nichol, A. Q. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.

Esser, P., Rombach, R., and Ommer, B. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Gaziv, G., Beliy, R., Granot, N., Hoogi, A., Strappini, F., Golan, T., and Irani, M. Self-supervised natural image reconstruction and large-scale semantic classification from brain activity. *NeuroImage*, 2022.

Gong, Z., Zhang, Q., Bao, G., Zhu, L., Xu, R., Liu, K., Hu, L., and Miao, D. MindTuner: Cross-subject visual decoding with visual fingerprint and semantic correction. In *AAAI Conference on Artificial Intelligence*, 2025.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 2001.

Hendrycks, D. and Gimpel, K. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. In *NeurIPS Workshop*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Horikawa, T. and Kamitani, Y. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 2017.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Huo, J., Wang, Y., Qian, X., Wang, Y., Li, C., Feng, J., and Fu, Y. NeuroPictor: Refining fMRI-to-image reconstruction via multi-individual pretraining and multi-level modulation. In *European Conference on Computer Vision*, 2024.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Kay, K. N., Naselaris, T., Prenger, R. J., and Gallant, J. L. Identifying natural images from human brain activity. *Nature*, 2008.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

Li, J., Li, D., Xiong, C., and Hoi, S. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022.

Lin, S., Sprague, T., and Singh, A. K. Mind reader: Reconstructing complex images from brain activities. *arXiv preprint arXiv:2210.01769*, 2022.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Lu, Y., Du, C., Wang, D., and He, H. MindDiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion. *arXiv preprint arXiv:2303.14139*, 2023.

Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., and Shan, Y. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI Conference on Artificial Intelligence*, 2024.

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., and Gallant, J. L. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 2009.

Naselaris, T., Kay, K. N., Nishimoto, S., and Gallant, J. L. Encoding and decoding in fMRI. *NeuroImage*, 2011.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 2021.

Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, 2022.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J. L. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 2011.

Ozcelik, F. and VanRullen, R. Brain-Diffuser: Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, 2023.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Lopes, R. G., Ayan, B. K., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022.

Scotti, P. S., Banerjee, A., Goode, J., Shabalin, S., Nguyen, A., Cohen, E., Dempster, A. J., Verlinde, N., Yundler, E., Weisberg, D., and Norman, K. A. Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. *arXiv preprint arXiv:2305.18274*, 2023.

Scotti, P. S., Tripathy, M., Torrico, C., Kneeland, R., Chen, T., Narang, A., Santhirasegaran, C., Xu, J., Naselaris, T., Norman, K. A., and Abraham, T. M. MindEye2: Shared-subject models enable fMRI-to-image with one hour of data. In *International Conference on Machine Learning*, 2024.

Seeliger, K., Güçlü, U., Ambrogioni, L., Güçlütürk, Y., and van Gerven, M. A. J. Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, 2018.

Shen, G., Horikawa, T., Majima, K., and Kamitani, Y. Deep image reconstruction from human brain activity. *PLOS Computational Biology*, 2019.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 2014.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

Takagi, Y. and Nishimoto, S. 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*, 2023.

Tan, M. and Le, Q. V. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, 2019.

van den Oord, A., Vinyals, O., and Kavukcuoglu, K. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, 2017.

Vapnik, V. N. *Statistical Learning Theory*. Wiley, 1998.

Wang, S., Liu, S., Tan, Z., and Wang, X. MindBridge: A cross-subject brain decoding framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004.

Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., and Liu, Z. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, 2018.

Xia, W., de Charette, R., Oztireli, C., and Xue, J.-H. DREAM: Visual decoding from reversing human visual system. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024a.

Xia, W., de Charette, R., Oztireli, C., and Xue, J.-H. UMBRAE: Unified multimodal brain decoding. In *European Conference on Computer Vision*, 2024b.

Xu, X., Wang, Z., Zhang, E., Wang, K., and Shi, H. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Zhang, X., Quan, R., Wang, W., and Yang, Y. Moving beyond diffusion: Hierarchy-to-hierarchy autoregression for fMRI-to-image reconstruction. *arXiv preprint arXiv:2510.22335*, 2025.

A. Decision-Rule View of Selection

The main paper uses a single-rendering protocol. For completeness, this appendix records the more general decision-rule notation. Let a brain representation $g_\theta(X)$ induce a candidate distribution

$$Y_i \overset{\text{i.i.d.}}{\sim} P_\theta(\cdot \mid g_\theta(X)), \qquad \mathcal{C}_N(X) = \{Y_1, \dots, Y_N\}. \tag{11}$$

A selector q returns

$$\hat{Y}_{q,N} = Y_{\arg\max_{1 \le i \le N} q(Y_i, g_\theta(X))}. \tag{12}$$

Thus a stochastic decoder is the triple $\delta = (g_\theta, q, N)$. If q uses the target image or target-derived features, the protocol is no longer equivalent to a brain-measurable single-rendering decoder.

For fixed (X, Y) and utility $u(y, Y) \in [0, 1]$, the gap between the best possible image and the selected candidate can be decomposed into representation, sampling, and selection-alignment terms. Our main protocol sets $N = 1$, so the selection-alignment term is absent by construction.

B. Additional Experimental Details

Data handling. The reported audit uses subject-01 NSD beta responses under the `nsdgeneral` voxel mask. We form the training set only from non-shared NSD images and remove every official shared-image identity from the training tensor before fitting the decoder. The evaluation set contains 982 valid shared-test examples after intersecting available fMRI trials, images, reconstructions, and metric files. No test image, caption, CLIP image embedding, VAE latent, or retrieval index is used to choose a reconstruction.

Phase-I targets. The visual teachers are computed once from the training images and then frozen. The VAE target is the SD1.5 autoencoder latent; the CLIP target is the frozen CLIP image embedding. These teachers supervise the mapping from fMRI to visual coordinates, but at test time the model only receives X_{test}. In symbols, the deployed variables are

$$\begin{aligned} \hat{\mathbf{z}}_{\text{test}} &= f_z(f_\theta(X_{\text{test}})), \\ \hat{\mathbf{c}}_{\text{test}} &= f_c(f_\theta(X_{\text{test}})), \end{aligned} \tag{13}$$

not oracle targets computed from Y_{test}.

Phase-II conditioning. Phase II freezes the Phase-I fMRI encoder and learns a BrainCondAdapter that maps the brain representation into cross-attention tokens for the frozen SD1.5 denoiser. The denoising loss samples timestep t, noise $\boldsymbol{\epsilon}$, and noised latent $\mathbf{z}_t$ in the usual latent-diffusion form. The alignment term encourages the learned tokens to preserve the CLIP-supervised visual coordinate without supplying target text at inference.

Inference and calibration. Each test example is reconstructed once with fixed DDIM parameters: 30 steps, strength 0.46, guidance scale 5.0, and seed 9. The reported clean benchmark applies one global, target-independent calibration setting to all reconstructions. This calibration is allowed because it is selected without seeing target identities or per-image metric feedback; it should be read as display normalization rather than image-specific post-processing.

Reporting discipline. We separate three evidence types. The clean benchmark is the primary row because it uses the strict single-rendering information contract. Diagnostic tables vary one factor at a time, such as sampler strength or checkpoint epoch, to explain behavior. Development evidence is not used as the headline claim unless it satisfies the same no-oracle, no-retrieval, no-best-of-N rule.

C. Qualitative Learning Progression

Figure 8 shows matched held-out stimuli reconstructed at several Phase-II training checkpoints. The panel is diagnostic rather than a separate benchmark: the stimuli, sampler family, and display layout are held fixed while the brain-to-diffusion interface improves over training time. The qualitative trend mirrors Section 6.5: coarse semantic structure emerges early, while object shape, viewpoint, and local visual detail become more stable at later checkpoints.

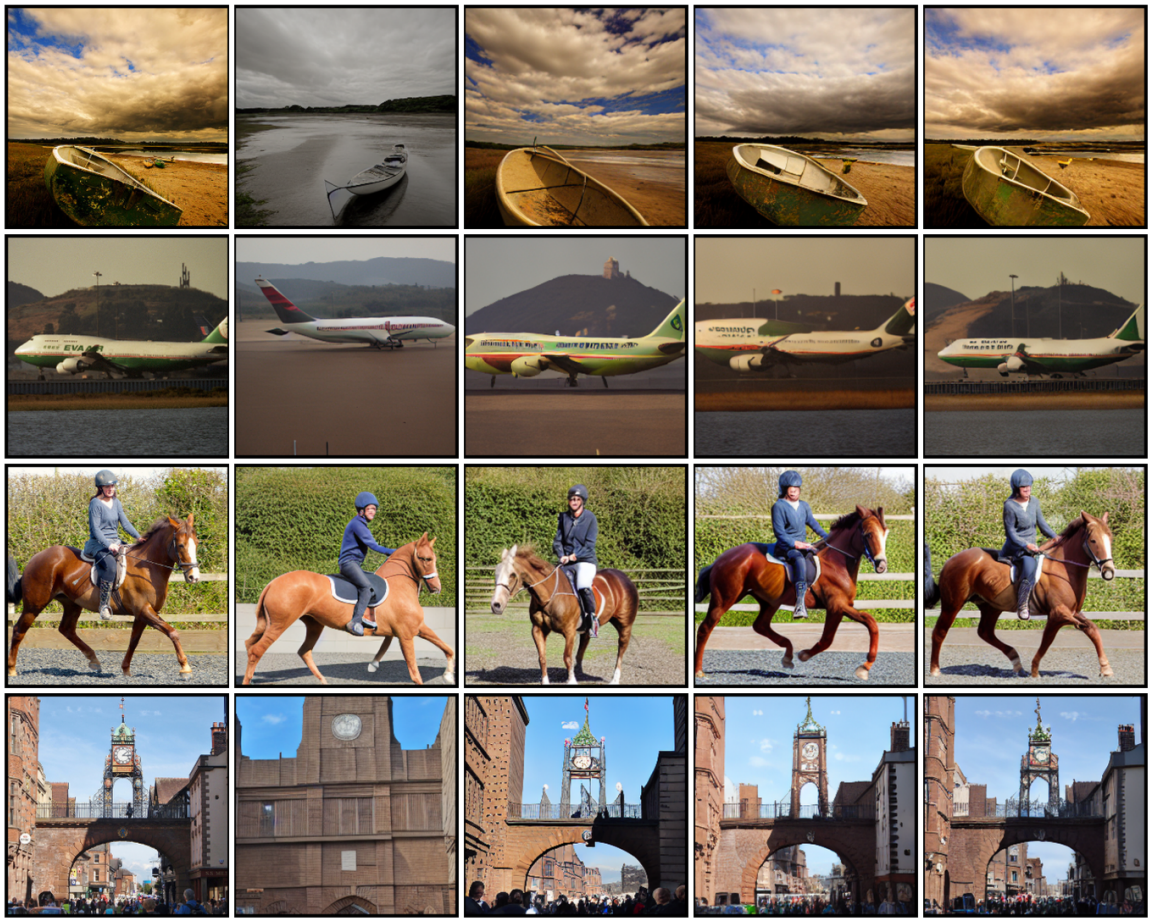

Figure 8. **Qualitative reconstruction progression across training time.** Fixed held-out NSD samples are shown across multiple Phase-II checkpoints. The progression is included as a diagnostic visualization of the learning-dynamics trace, not as an additional selection or reranking protocol.

D. Principle: Brain-Measurable Diffusion Decoding

The main principle is to distinguish *where a representation was learned* from *what information enters at inference*. A target-side foundation model can be a useful teacher during training, but the deployed reconstruction is brain-measurable only when every target-specific control is predicted from fMRI. This gives a simple audit rule:

$$\mathrm{Info}(\hat{Y}) \subseteq \mathrm{Info}(X_{\mathrm{test}}, \theta^\star, \eta), \tag{14}$$

where $\theta^\star$ denotes fixed learned parameters and η denotes fixed sampler randomness. The set should not include Y_{test}, target captions, target CLIP features, retrieval neighbors, or evaluation-pool labels.

This criterion is stricter than ordinary sample-quality evaluation. A powerful diffusion prior may fill in plausible texture, but plausibility alone does not show that the missing information came from the brain measurement. The ablations in Section 6.3 are designed around this principle: mismatched fMRI tests whether outputs follow the measurement, latent-only decoding tests whether the VAE initialization is sufficient, and token-only decoding tests whether learned cross-attention controls carry the main semantic signal.

The principle also explains why we emphasize sampler diagnostics. DDIM strength changes the relative influence of the fMRI-predicted latent and the fMRI-predicted tokens. A fixed sampler is therefore part of the scientific protocol, not a cosmetic rendering choice. Reporting the sampler-strength curve makes the operating point auditable and prevents a hidden hyperparameter from being mistaken for a new source of brain information.