DECODING BEYOND HEMODYNAMIC DELAY: DELAY-COUPLED GRAPH DYNAMICS FOR FMRI-TO-IMAGE RECONSTRUCTION

Anonymous authors
Paper under double-blind review



Figure 1: Reconstruction examples on NSD and THINGS-fMRI.

ABSTRACT

Reconstructing visual stimuli from functional magnetic resonance imaging (fMRI) requires extracting stimulus-related information from spatially distributed blood-oxygen-level-dependent (BOLD) responses shaped by hemodynamic delay and temporal dispersion. We propose **DCG-CHI**, a framework that combines delay-coupled graph representation learning with hemodynamic consistency training. Its **Delay-Coupled Graph State-Space Model (DCG-SSM)** embeds graph diffusion within delayed state transitions, explicitly coupling propagation scales to temporal lags and combining propagation modes through input-dependent weights. **Counterfactual Hemodynamic Invariance (CHI)** synthesizes two BOLD views of a shared latent trajectory through a parameterized pair of hemodynamic response kernels. Cross-view agreement, stop-gradient anchoring, and synthetic BOLD re-synthesis encourage the shared encoder to recover consistent representations. Independent semantic and structural readouts condition a pretrained Stable Diffusion 2.1 backbone through cross-attention and a trainable ControlNet branch, respectively. CHI is used only during training. Experiments on NSD and THINGS-fMRI demonstrate competitive structural fidelity and semantic consistency, with SSIM scores of **0.468** and **0.366** and CLIP identification scores of **0.963** and **0.755**, respectively. Ablations and controlled comparisons support the contributions of both components, while additional tests show improved reconstruction under hemodynamic response shifts.

1 INTRODUCTION

Reconstructing viewed images from functional magnetic resonance imaging (fMRI) provides a means of investigating how visual information can be recovered from distributed brain responses. Pretrained latent diffusion models have improved reconstruction realism and semantic consistency (Takagi & Nishimoto, 2023; Ozcelik & VanRullen, 2023; Scotti et al., 2023). However, visually plausible outputs may contain details unsupported by the measured response. Reconstruction therefore requires

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.

both an expressive image prior and a brain representation that preserves stimulus-related information to guide generation.

The blood-oxygen-level-dependent (BOLD) signal reflects neural activity through a hemodynamic response function (HRF) that introduces delay and temporal dispersion. Variation in this response complicates the mapping from measured signals to latent activity (Friston et al., 1998; Lindquist et al., 2009). Trial-level activation vectors summarize these responses, whereas retaining BOLD sequences makes their temporal structure available to the encoder. Sequence modeling alone, however, does not explicitly constrain representations to remain consistent under changes in the hemodynamic observation process.

We therefore consider two related design questions. First, how should regional information be integrated across temporal lags? Applying the same graph filter directly to each retained state leaves its propagation scale independent of that state's lag. Second, how can training encourage consistent representations when the modeled HRF changes? Image-feature supervision provides visual targets but does not explicitly enforce agreement across different hemodynamic observations of the same latent trajectory.

We propose **DCG-CHI**, which combines a *Delay-Coupled Graph State-Space Model* (DCG-SSM) with *Counterfactual Hemodynamic Invariance* (CHI). DCG-SSM maintains states for individual regions of interest (ROIs) on a fixed functional graph and embeds graph diffusion within delayed state transitions. Lag-dependent diffusion scales and input-dependent mode weights determine how regional history contributes to each update. CHI synthesizes two BOLD views of a shared latent trajectory through a parameterized pair of HRFs. The shared encoder is regularized through cross-view agreement, stop-gradient anchoring, and synthetic BOLD re-synthesis. Together, the components structure temporal integration and encourage consistency across modeled hemodynamic observations (Figure 2). Independent semantic and structural readouts condition a pretrained Stable Diffusion 2.1 backbone through cross-attention and a trainable ControlNet branch, respectively (Huo et al., 2024). CHI is used only during training. Inference uses BOLD-derived conditions without access to stimulus images or captions.

Our contributions are threefold:

- **Delay-coupled graph encoding.** We introduce a recurrent encoder that explicitly couples graph diffusion scales to temporal lags, with input-dependent mode mixing and sparse polynomial evaluation.
- **Hemodynamic consistency training.** We develop a shared-encoder objective across synthetic hemodynamic views to regularize latent representations without adding an inference-time branch.
- **Reconstruction and robustness evaluation.** Experiments on NSD and THINGS-fMRI show competitive structural fidelity and semantic consistency. Controlled comparisons support the contributions of both components, while additional tests show improved reconstruction under the HRF shifts.

2 Related work

Generative brain decoding. MinD-Vis learns brain representations through masked modeling, while MindEye combines contrastive alignment with diffusion priors (Chen et al., 2023; Scotti et al., 2023). DREAM models complementary visual attributes, and NeuroPictor conditions generation through multilevel modulation (Xia et al., 2024; Huo et al., 2024). MindHier aligns hierarchical fMRI features with coarse-to-fine autoregressive generation (Zhang et al., 2026). Within this setting, DCG-CHI investigates delay-dependent regional propagation and consistency across synthetic hemodynamic observations as complementary aspects of BOLD representation learning.

Alignment and temporal representations. MindBridge, MindEye2, and related studies investigate cross-subject alignment, data efficiency, and generalization (Wang et al., 2024; Scotti et al., 2024; Kong et al., 2025; Dahan et al., 2025). CHI addresses a complementary source of variation by regularizing a shared encoder across synthetic hemodynamic observations of the same latent trajectory. Mind-Animator and BrainNRDS use temporal information for video reconstruction (Lu et al., 2025; Yeung et al., 2025). DCG-CHI instead models stimulus-centered BOLD histories for static-image reconstruction. EEG reconstruction methods further explore modality-specific representations (Bai et al., 2024; Li et al., 2024), with D^2-FOSA combining frequency-oriented encoding and dual-diffusion alignment (Yu et al., 2026).

Graph and state-space modeling. Graph convolutions and diffusion-based recurrence model relational dependencies (Kipf & Welling, 2017; Li et al., 2018), while S4, S5, and selective state-space models provide approaches to long-sequence modeling (Gu et al., 2022; Smith et al.,

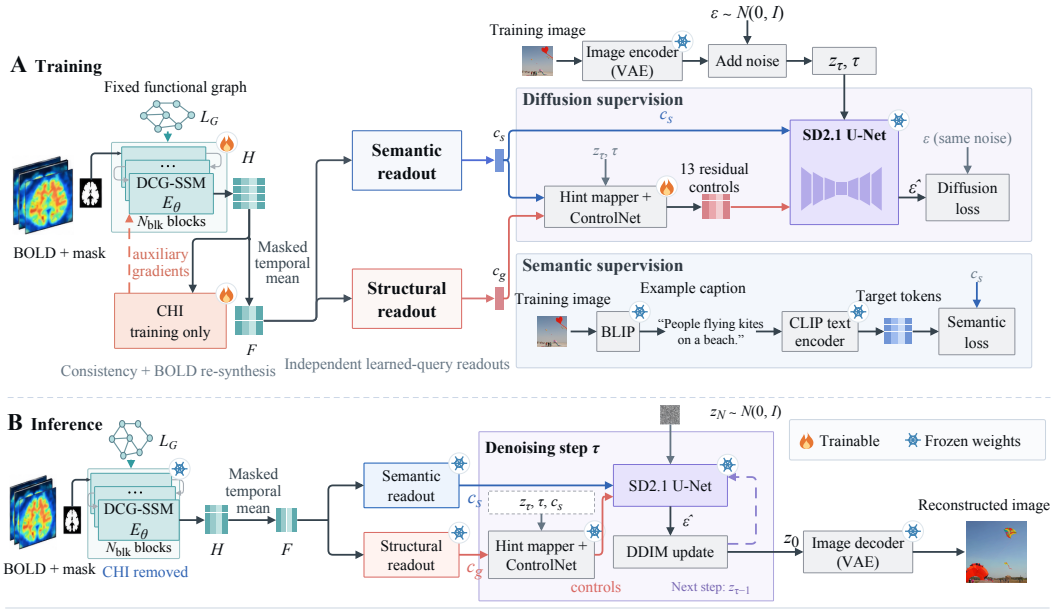

Figure 2: **Overview of DCG-CHI.** DCG-SSM encodes BOLD sequences with validity masks on a fixed functional graph. Semantic and structural readouts condition the diffusion generator. During training, CHI adds shared-encoder consistency across synthetic hemodynamic views and BOLD re-synthesis. At inference, the synthetic branches are removed, and reconstruction uses BOLD-derived conditions without stimulus images or captions.

2023; Gu & Dao, 2024; Dao & Gu, 2024). DCG-SSM uses a finite distributed-delay recurrence that explicitly couples graph diffusion scales to temporal lags, with input-dependent weights combining propagation modes. Chebyshev polynomial filtering enables sparse evaluation of these operators (Defferrard et al., 2016). Appendix O provides further comparisons, including semantic alignment and synthetic brain responses.

3 Method

3.1 Problem Formulation

Let $X \in \mathbb{R}^{T\times R}$ denote a trial-centered BOLD sequence, $M \in \{0,1\}^{T\times R}$ its validity mask, and I the corresponding training image. A fixed ROI graph is represented by a symmetric normalized Laplacian L_G with spectrum in $[0, 2]$. The graph is constructed from training data. Dataset-specific normalization is detailed in Appendix B. The encoder produces $H = E_\theta(X, M, L_G) \in \mathbb{R}^{T\times R\times d}$ for semantic and structural readout. Channel maps use row-vector notation, and the batch dimension is omitted. The observation model

$$X_{t,r} = \sum_{\ell\geq 0} \psi_r(\ell) n_{t-\ell,r} + \epsilon_{t,r} \tag{1}$$

relates unobserved neural drive n to BOLD through an HRF ψ_r and noise ϵ. This model motivates synthetic HRF observations of the task-trained trajectory H. Figure 3 summarizes the two core components.

3.2 Delay-Coupled Graph State-Space Model

Each BOLD observation and its validity indicator are mapped to ROI features $u_t \in \mathbb{R}^{R\times d}$ by a shared affine map, LayerNorm, and SiLU. Within each block, a masked ROI mean determines a mixture over M_q propagation modes:

$$\bar{u}_t = \frac{\sum_r M_{t,r} u_{t,r}}{\max(1, \sum_r M_{t,r})}, \qquad \pi_t = \mathrm{softmax}(\mathrm{LN}(\bar{u}_t) W_\pi + b_\pi)\,. \tag{2}$$

The weights $\pi_t \in \mathbb{R}^{M_q}$ vary with the input and are shared across ROIs at each time step. Each mode has a normalized delay profile over $k = 1, \ldots, K$:

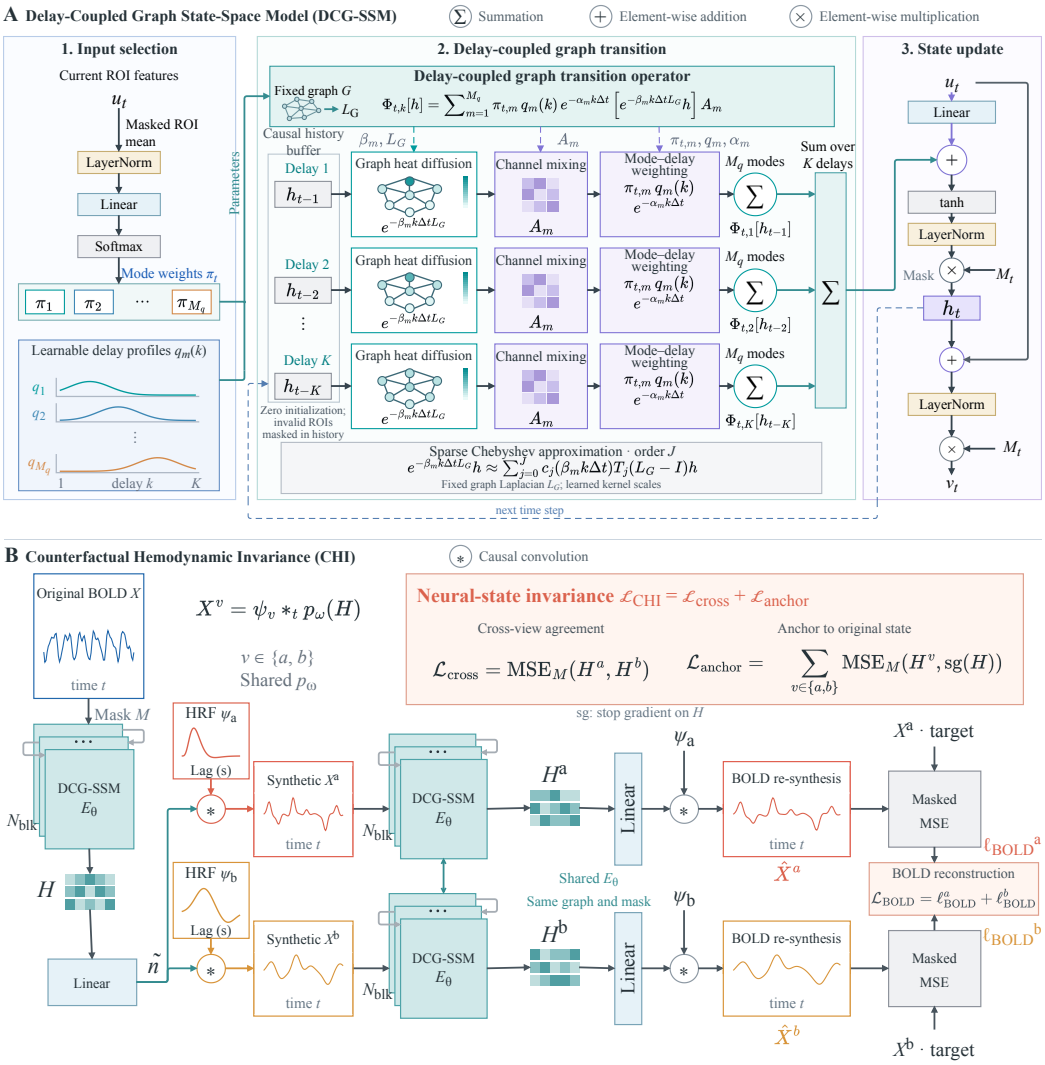

Figure 3: **Coupled graph dynamics and hemodynamic consistency.** DCG-SSM integrates masked historical states through lag-dependent graph filters. CHI reuses the encoder across HRF-generated views with cross-view agreement, stop-gradient anchoring, and synthetic BOLD re-synthesis. State dropout and $\mathcal{L}_{\text{reg}}$ are omitted for clarity.

$$q_m(k) = \frac{\exp[-(k-\mu_m)^2/(2\sigma_m^2)]}{\sum_{j=1}^{K}\exp[-(j-\mu_m)^2/(2\sigma_m^2)]}, \qquad \mu_m \in [1,K], \quad \sigma_m > 0. \tag{3}$$

These profiles parameterize recurrent history and are distinct from the HRF kernels used by CHI (delay parameterization in Appendix B.2). For a retained state h, the delay-coupled transition is

$$\Phi_{t,k}[h] = \sum_{m=1}^{M_q} \pi_{t,m} q_m(k) e^{-\alpha_m k\Delta t}\left[e^{-\beta_m k\Delta t L_G} h\right] A_m, \tag{4}$$

where $\alpha_m, \beta_m > 0$, $A_m \in \mathbb{R}^{d\times d}$ mixes channels, and $\Delta t = 1\,\text{s}$ is the model time step after preprocessing. The diffusion scale $\beta_m k\Delta t$ explicitly couples graph propagation to temporal lag. Within each mode, the exact heat kernel attenuates a graph-frequency component with eigenvalue λ by $e^{-\beta_m k\Delta t\lambda}$. Longer lags therefore induce stronger graph smoothing, while mode mixing combines different temporal-decay and spatial-diffusion scales.

Let $D_t = \text{diag}(M_{t,:})$. Before each update, retained buffer entries are overwritten as $h_{t-k} \leftarrow D_t h_{t-k}$, clearing currently invalid ROI entries before graph propagation. The recurrent state and residual output are

$$\widetilde{h}_t = \tanh\left(u_t B + \mathbf{1}_R b^\top + \sum_{k=1}^{K} \Phi_{t,k}[h_{t-k}]\right),$$
$$h_t = D_t \,\mathrm{Dropout}\Big(\mathrm{LN}(\widetilde{h}_t)\Big), \qquad v_t = D_t \,\mathrm{LN}(u_t + h_t). \tag{5}$$

Here $B \in \mathbb{R}^{d\times d}$ and $b \in \mathbb{R}^d$ are shared across ROIs. History is initialized to zero. h_t enters the recurrent buffer, while v_t feeds the next block.

Sparse graph evaluation. We evaluate the heat operator using a Chebyshev approximation:

$$e^{-sL_G}h \approx \mathcal{P}_J(s; L_G)h = \sum_{j=0}^{J} c_j(s) T_j(L_G - I_R)h, \tag{6}$$

where I_R is the identity matrix, T_j is the jth Chebyshev polynomial, and $c_j(s)$ is obtained by differentiable cosine interpolation (Appendix B.3). Polynomial bases are shared across modes for each retained state, avoiding eigendecomposition in the forward pass. For $|E|$ graph nonzeros, the per-block cost is $O\big(TK[J|E|d + M_q(JRd + Rd^2)] + TRd^2\big)$, excluding selector and coefficient-generation overhead. For fixed architectural budgets, computation is linear in T, and the inference history requires $O(KRd)$ memory per block. Training additionally stores activations, and the HRF cycle adds two encoder passes.

3.3 Counterfactual hemodynamic invariance

CHI constructs counterfactual views by applying different synthetic HRF operators to the same latent trajectory. A shared linear projection $p_\omega : \mathbb{R}^d \to \mathbb{R}$ produces the scalar trajectory $\widetilde{n} = p_\omega(H)$. The peak–undershoot HRF pair shares trainable base parameters, with fixed view-specific peak shifts and width scales, and is shared across ROIs (Appendix B). Causal, zero-padded convolution defines

$$X^v = \mathcal{O}_v(H) = \psi_v *_{\text{causal}} \widetilde{n}, \qquad H^v = E_\theta(X^v, M, L_G), \qquad v \in \{a, b\}, \tag{7}$$

where $\mathcal{O}_v(U) = \psi_v *_{\text{causal}} p_\omega(U)$. All views use the same encoder parameters, graph, and mask. Let MSE_M average squared error over valid entries and feature channels. The consistency objective is

$$\mathcal{L}_{\text{CHI}} = \underbrace{\mathrm{MSE}_M(H^a, H^b)}_{\mathcal{L}_{\text{cross}}} + \underbrace{\sum_{v\in\{a,b\}} \mathrm{MSE}_M(H^v, \mathrm{sg}(H))}_{\mathcal{L}_{\text{anchor}}}, \tag{8}$$

where sg denotes the stop-gradient operation. The recovered states are further constrained to reproduce their corresponding synthetic observations:

$$\mathcal{L}_{\text{BOLD}} = \sum_{v\in\{a,b\}} \mathrm{MSE}_M(\mathcal{O}_v(H^v), X^v)\,. \tag{9}$$

Only the anchor target is detached. View synthesis, including its dependence on H, and shared re-encoding remain differentiable. These losses regularize the synthetic observation cycle, while diffusion and semantic supervision connect the representation to image reconstruction.

3.4 Semantic and structural conditioning

Masked temporal pooling produces ROI descriptors $F_r = \sum_t M_{t,r}H_{t,r} / \max(1, \sum_t M_{t,r})$. Two independent multihead attention readouts with learned queries generate conditioning tokens:

$$c_\nu = \mathrm{LN}(\mathrm{MHA}_\nu(P_\nu, F, F)W_\nu)\,, \qquad \nu \in \{s, g\}. \tag{10}$$

ROIs without valid time points are excluded as attention keys. The semantic tokens $c_s \in \mathbb{R}^{77\times1024}$ condition cross-attention in a pretrained Stable Diffusion 2.1 U-Net. The structural tokens $c_g \in \mathbb{R}^{257\times1024}$ are mapped to a spatial hint for a trainable ControlNet branch that supplies 13 residual controls (Huo et al., 2024; Zhang et al., 2023). The pretrained U-Net and VAE remain frozen, while the encoder, readouts, and control pathway are trained.

Let z_τ denote the noisy VAE latent at diffusion step τ, obtained using Gaussian noise ϵ, and write the control residuals as $r_\tau = \mathrm{Control}(c_g, z_\tau, \tau, c_s)$. We use the latent diffusion noise-prediction objective (Rombach et al., 2022):

$$\mathcal{L}_{\text{diff}} = \mathbb{E}\Big[\|\epsilon - \epsilon_{\text{SD}}(z_\tau, \tau; c_s, r_\tau)\|_2^2\Big]\,. \tag{11}$$

BLIP generates captions for training images (Li et al., 2022), and a frozen CLIP-family text encoder provides target tokens (Radford et al., 2021). Semantic supervision $\mathcal{L}_{\text{sem}}$ minimizes token-wise MSE

between c_s and these targets. The total objective is

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_s \mathcal{L}_{\text{sem}} + \lambda_c \mathcal{L}_{\text{CHI}} + \lambda_b \mathcal{L}_{\text{BOLD}} + \lambda_r \mathcal{L}_{\text{reg}}, \tag{12}$$

where $\mathcal{L}_{\text{reg}}$ regularizes the recovered-state variance (Appendix B). At inference, the HRF bank and synthetic branches are removed. DDIM sampling (Song et al., 2021) uses BOLD-derived conditions and random noise, without stimulus images or captions (algorithm in Appendix B.7).

3.5 Operator properties

Permutation equivariance. For an ROI permutation matrix P, deterministic evaluation satisfies $E_\theta(XP^\top, MP^\top, PL_GP^\top)_t = PH_t$. Shared ROI maps, masked pooling, and graph filtering preserve this consistent relabeling, while the query readouts are invariant.

Conditional stability. The Frobenius cap $\|A_m\|_F \le c < 1$ bounds the exact linear history operator. Under deterministic evaluation with fixed inputs, masks, and mode weights, let $\widehat{\Phi}_{t,k}$ include the implemented polynomial filters and masks, and let L_σ bound the Lipschitz constant of the normalized nonlinearity. The sufficient condition

$$\rho = L_\sigma \sup_t \sum_{k=1}^{K} \left\| \widehat{\Phi}_{t,k} \right\|_{F \to F} < 1 \tag{13}$$

ensures that the maximum history difference between two trajectories with identical driving inputs is at most ρ times its previous bound after each K-step window. This condition applies to a single recurrent block and accounts for normalization and filter approximation. Proofs appear in Appendix C.

4 Experiments

4.1 Experimental setup

Datasets and evaluation. We evaluate DCG-CHI on NSD subjects S01, S02, S05, and S07 (Allen et al., 2022) and three THINGS-fMRI subjects (Hebart et al., 2023). Each input is a 15×360 ROI sequence on a preprocessed 1-second grid at stimulus-relative times $-2, -1, \ldots, 12$ seconds. Repetitions are grouped by stimulus identity, and the ROI graph is constructed from training data only. Evaluation covers PixCorr, SSIM, four feature-identification scores (AlexNet-2/5, Inception, and CLIP), and two feature distances (EfficientNet-B and SwAV). Higher values are preferred except for the two distances. Appendix D specifies the splits, metrics, and aggregation.

Implementation details. Internal controls share the generation architecture and sampling settings, so the comparisons examine changes in BOLD encoding and consistency training. The full encoder uses six blocks of width 64, four propagation modes, and Chebyshev order six. The independent learned-query readouts each use $N_h = 8$ attention heads; concatenated outputs pass through W^O, a projection to 1024 dimensions, and LayerNorm, producing 77 semantic and 257 structural tokens, respectively. Controlled runs use 12,000 optimization steps and an effective batch size of 36. Generation uses 50 denoising steps, classifier-free guidance scale 5, and 512×512 resolution. Table 8 and Appendix E provide the complete configuration and control definitions.

4.2 Reconstruction performance

Quantitative reconstruction. Table 1 evaluates reconstruction in both pixel and feature spaces. DCG-CHI achieves SSIM/CLIP identification of 0.468/0.963 on NSD and 0.366/0.755 on THINGS-fMRI, alongside PixCorr of 0.332 and 0.114. The results show a favorable combination of structural similarity and semantic discrimination across the two datasets, supporting the use of the learned BOLD representation to guide both aspects of generation. External scores are taken from the cited papers. The internal controls below assess the contributions of the proposed encoder and training objective. Additional metric profiles appear in Figure 8. The fixed quality index applied to the rounded full-model means gives $Q_{\text{tab}} = 0.8813/0.6142$ on NSD/THINGS-fMRI (Figure 5C; Appendix H.1); these are derived summaries of the same reconstruction metrics.

Qualitative reconstruction. In the examples in Figure 4, DCG-CHI more closely preserves object identity, scene layout, and background detail than the illustrated NeuroPictor and BrainFLORA outputs. The correspondence extends beyond the main object to its surrounding scene, illustrating how structural and semantic information can be retained in the same reconstruction. These selected examples complement the dataset-level metrics; Appendix G provides additional reconstructions.

4.3 Ablations and hemodynamic robustness

Spatial–temporal encoding. Table 2 compares DCG-SSM with a token-wise residual MLP (G0), graph-only processing (G1), delay-only processing (G2), parallel spatial and temporal branches

Table 1: **Reconstruction results on NSD and THINGS-fMRI.** External scores are taken from the cited papers. Arrows indicate the preferred direction.

	Method	PixCorr↑	SSIM↑	Alex2↑	Alex5↑	Incep.↑	CLIP↑	Eff-B↓	SwAV↓
NSD	DREAM (Xia et al., 2024)	0.288	0.338	0.950	0.975	0.948	0.952	0.638	0.413
	MindBridge (Wang et al., 2024)	0.151	0.263	0.877	0.955	0.924	0.947	0.712	0.418
	MindEye2 (Scotti et al., 2024)	0.322	0.431	0.961	0.986	0.954	0.930	0.619	0.344
	NeuroPictor (Huo et al., 2024)	0.229	0.375	0.965	0.984	0.945	0.933	0.639	0.350
	NeuralDiffuser (Li et al., 2025b)	0.299	0.318	0.953	0.983	0.954	0.954	0.663	0.401
	MindHier (Zhang et al., 2026)	0.326	0.461	0.931	0.980	0.959	0.954	**0.613**	0.345
	DCG-CHI (Ours)	**0.332**	**0.468**	**0.968**	**0.990**	**0.966**	**0.963**	0.620	**0.332**
THINGS-fMRI	BrainFLORA (Li et al., 2025a)	0.079	0.348	0.595	0.622	0.604	0.576	**0.812**	0.661
	MindHier (Zhang et al., 2026)	0.109	0.357	0.710	0.819	0.719	0.731	0.900	0.589
	DCG-CHI (Ours)	**0.114**	**0.366**	**0.735**	**0.845**	**0.748**	**0.755**	0.825	**0.565**

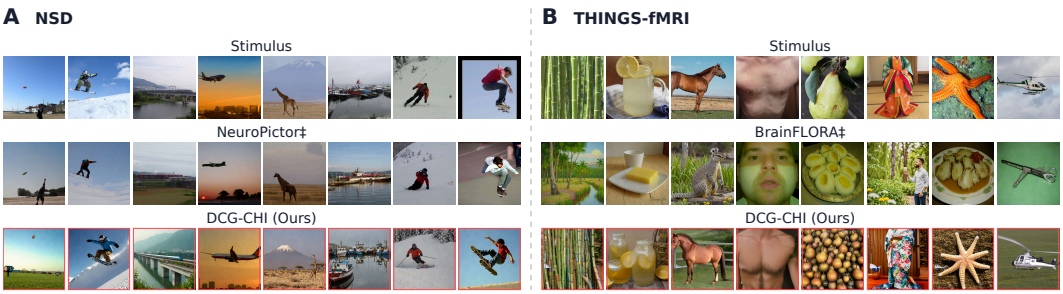

Figure 4: **Qualitative fMRI-to-image reconstruction.** Rows show stimuli, baseline outputs, and DCG-CHI reconstructions for each dataset. Baselines are NeuroPictor (NSD) and BrainFLORA (THINGS-fMRI). ‡ marks results reproduced from the papers or released code.

(G3), and a serial spatial-then-temporal encoder. All retain CHI and the downstream readouts (Appendix E.1). Replacing DCG-SSM with G0 reduces SSIM/CLIP identification by 0.009/0.010 on NSD and 0.010/0.025 on THINGS-fMRI. Parallel and serial combinations improve both scores over graph-only and delay-only processing, indicating the value of using regional interactions together with temporal history. DCG-SSM further improves SSIM/CLIP over Serial by 0.002/0.001 on NSD and 0.002/0.002 on THINGS-fMRI, with a tie on NSD AlexNet-5. Under the shared generation architecture, these results favor the complete delay-coupled recurrence in Equation 4 over the evaluated separate-branch and sequential alternatives.

Hemodynamic consistency. With DCG-SSM fixed, C0 removes cross-view agreement and anchoring; C1 applies both constraints to generic noise and masking views; C2 uses HRF views without anchoring; and C3 is the full model. All retain the same BOLD re-synthesis and variance losses (Appendix E.2). C1 improves on C0, showing that generic consistency already benefits reconstruction. Full CHI further improves SSIM/CLIP over C1 by 0.003/0.002 on NSD and 0.004/0.004 on THINGS-fMRI. Removing anchoring also lowers both scores on both datasets. Thus, HRF-based view construction adds value beyond the tested generic consistency, while anchoring improves the HRF-view formulation. These comparisons support tailoring representation regularization to the modeled BOLD observation process. The separate *w/o CHI** run in Figures 8 and 5A/B is documented in Appendix E.3.

Table 2: **Encoder and consistency comparisons.** G-variants retain CHI. C-variants retain DCG-SSM. G0 also serves as the DCG-SSM removal ablation. Each metric cell reports NSD / THINGS-fMRI. The full model, shared by G4 and C3, is listed once. Arrows indicate the preferred direction.

Configuration		PixCorr↑	SSIM↑	Alex2↑	Alex5↑	Incep.↑	CLIP↑	Eff-B↓	SwAV↓
G0	MLP	0.325/0.108	0.459/0.356	0.960/0.705	0.985/0.815	0.957/0.716	0.953/0.730	0.623/0.829	0.349/0.594
G1	Graph only	0.328/0.111	0.462/0.359	0.964/0.720	0.987/0.829	0.960/0.731	0.958/0.742	0.622/0.827	0.342/0.581
G2	Delay only	0.327/0.110	0.464/0.361	0.962/0.717	0.987/0.832	0.961/0.734	0.957/0.740	0.624/0.830	0.340/0.577
G3	Parallel	0.330/0.112	0.465/0.362	0.966/0.730	0.989/0.839	0.963/0.741	0.960/0.749	0.621/0.826	0.336/0.571
G_{serial}	Serial	0.331/0.113	0.466/0.364	0.967/0.733	**0.990**/0.843	0.965/0.746	0.962/0.753	0.621/0.826	0.334/0.568
C0	No cons.	0.328/0.110	0.463/0.360	0.966/0.728	0.988/0.837	0.963/0.740	0.960/0.747	0.624/0.831	0.339/0.576
C1	Generic	0.330/0.112	0.465/0.362	0.967/0.731	**0.990**/0.841	0.964/0.744	0.961/0.751	0.622/0.828	0.336/0.571
C2	No anchor	0.329/0.111	0.466/0.363	0.967/0.733	0.989/0.842	0.965/0.745	0.962/0.752	**0.620**/0.827	0.335/0.569
DCG-CHI (Ours)		**0.332**/**0.114**	**0.468**/**0.366**	**0.968**/**0.735**	**0.990**/**0.845**	**0.966**/**0.748**	**0.963**/**0.755**	**0.620**/**0.825**	**0.332**/**0.565**

HRF transfer perturbations. We next assess sensitivity to temporal observation changes using the same continuous-run HRF transfer perturbations for all compared models (Appendix K). At an absolute shift of two seconds, averaging positive and negative shifts, full CHI achieves SSIM of 0.432/0.338 on NSD/THINGS-fMRI, compared with 0.404/0.315 for generic consistency. With CHI fixed, delay-coupled encoding achieves 0.432/0.338, compared with 0.412/0.321 for parallel encoding (Tables 17 and 18). The SSIM margins over generic consistency (0.028/0.023) and parallel encoding (0.020/0.017) exceed their clean-input counterparts. The larger margins connect the controlled ablations to the motivation for DCG-SSM and CHI: integrating delayed regional evidence and regularizing hemodynamic observations help preserve reconstruction under the specified HRF-transfer perturbations.

4.4 Representation Analysis

Semantic and structural conditioning. Figure 5D exchanges semantic conditions, structural conditions, or both between stimuli, and evaluates two-way identification of the original stimulus using SSIM and CLIP cosine similarity (Appendix H.2). Semantic exchange primarily reduces semantic identification, whereas structural exchange has a larger effect on structural identification; exchanging both reduces both scores. These differentiated responses suggest complementary roles for the two readouts in conveying the shared BOLD representation to the generator. The analysis therefore supports retaining distinct conditioning paths for semantic content and image structure.

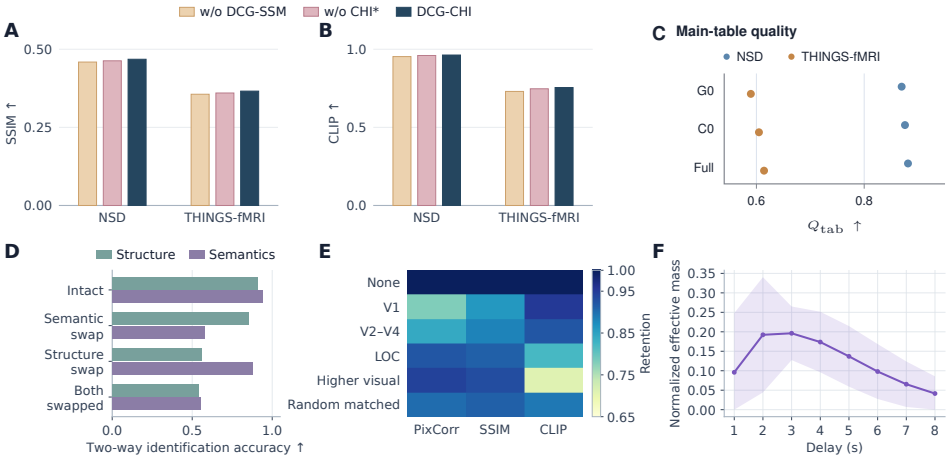

Figure 5: **Component comparisons and representation diagnostics.** (A/B) SSIM and CLIP identification under component removal. (C) Quality references derived from Table 2. (D) Structural/semantic two-way identification of the original stimulus after condition exchange, using SSIM/CLIP cosine. (E) ROI masking and retained performance. (F) Learned effective-delay mass; shading denotes standard deviation across layer–mode profiles.

Regional sensitivity. Masking V1 reduces PixCorr more than masking higher visual regions, whereas the reverse pattern holds for CLIP identification (Figure 5E). The different effects on pixel agreement and semantic discrimination indicate that reconstruction draws on complementary regional inputs. This pattern is consistent with the goal of preserving distributed visual information in the BOLD representation and characterizes the trained decoder's use of those inputs.

Learned delays and cortical graph influence. The effective-delay profile peaks at three seconds, with most mass between two and four seconds (Figure 5F), revealing nonuniform weighting of the retained history. Figure 6 complements this temporal profile with checkpoint-derived graph propagation maps over total, early, and late model lags. Together, these diagnostics make the encoder's temporal weighting and regional graph-filter structure interpretable. The cortical scores summarize graph-kernel parameters rather than sample-specific contributions, omitting input-dependent mode gates and channel mixing. Appendix H provides the definitions and aggregation procedures.

4.5 Additional Evaluations

Subject transfer and adaptation. Protocol comparisons use the common quality scale $L_{\text{protocol}} = \overline{Q}$ (Appendix D.4). Multi-source training improves the protocol score over single-source training on both datasets (Figure 7A/C). With ten target calibration examples, the score increases from 0.7315 to 0.8284 on NSD and from 0.4913 to 0.5712 on THINGS-fMRI (panel D). These results in-

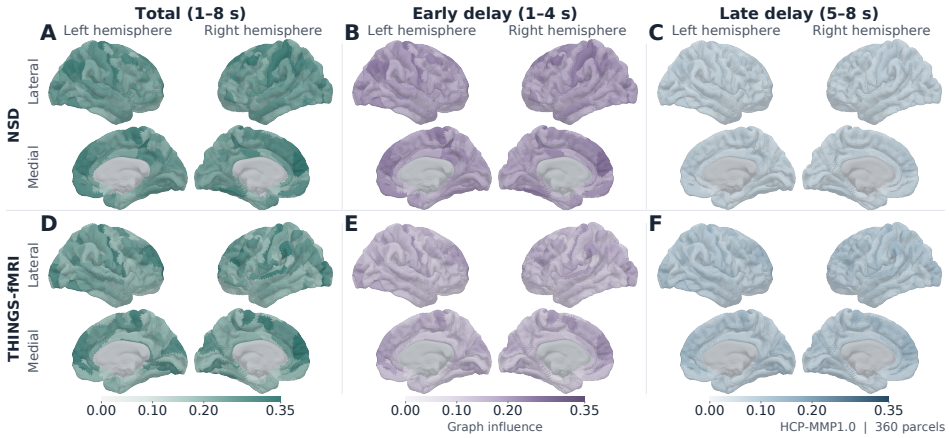

Figure 6: **Cortical graph influence across model-lag ranges.** NSD (A–C) and THINGS-fMRI (D–F): total (1–8 s), early (1–4 s), and late (5–8 s). Raw cumulative values share a 0–0.35 color scale without panel-wise renormalization (Appendix H.3).

dicate that the representation supports within-dataset transfer and benefits from limited target-subject calibration. Appendix I distinguishes joint training, held-out-target transfer, and adaptation.

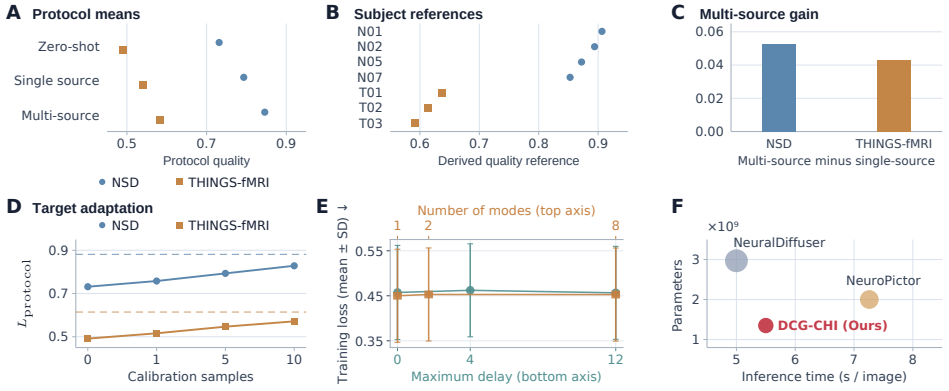

Figure 7: **Subject transfer, adaptation, and computation.** (A) Protocol means; zero-shot matches zero calibration in D. (B) Full-model subjectwise Q_{tab} from Table 5. (C) Multi-source minus single-source quality gain. (D) Target adaptation on $L_{\text{protocol}} = \overline{Q}$; dashed lines mark full-model main-table Q_{tab}. (E) Separate delay and mode-count scans with training-loss SD. (F) Parameters versus inference time; circle area is proportional to parameter count.

Computational cost. DCG-CHI has 1.35B total parameters, of which 0.20B are trainable. The reported 50-step configuration takes 5.50 seconds per image. The auxiliary HRF-view passes are confined to training, so CHI's regularization does not require synthetic-view encoding at inference. Appendix N details the resource accounting, and Figure 7F summarizes the reported configurations. Optimization scans appear in Appendix F; retrieval, additional robustness, and sampling analyses are provided in Appendices J–M.

5 Discussion and Conclusion

We presented DCG-CHI, a framework that combines delay-coupled graph encoding with hemodynamic consistency training for fMRI-to-image reconstruction. DCG-SSM explicitly couples graph diffusion scales to temporal lags, while CHI regularizes a shared encoder across synthetic hemodynamic observations without adding an inference-time branch. Experiments on NSD and THINGS-fMRI show competitive structural fidelity and semantic consistency. Ablations and controlled comparisons support the contributions of both components, while tests under the evaluated HRF perturbations show improved reconstruction over the corresponding controls. Together, these findings support integrating structured temporal encoding with observation-level regularization for visual brain decoding. Our evaluation is limited to two datasets, and CHI uses an HRF family shared across ROIs; broader validation across acquisition settings remains future work.

REPRODUCIBILITY STATEMENT

Appendices B, E, C, and D provide implementation details, control settings, theoretical proofs, and evaluation protocols, respectively.

ETHICS STATEMENT

This study uses only publicly available data from NSD and THINGS-fMRI, with no new human-subject data collection. Potential applications of visual brain decoding should protect participants' privacy and respect informed consent.

AI USE STATEMENT

An OpenAI language-model assistant supported manuscript language editing and the drafting of portions of the code. The authors remain fully responsible for the manuscript and code.

REFERENCES

Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25:116–126, 2022.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision*, pp. 382–398, 2016.

Yunpeng Bai, Xintao Wang, Yan-Pei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. DreamDiffusion: High-quality EEG-to-image generation with temporal masked signal modeling and CLIP alignment. In *Proceedings of the European Conference on Computer Vision*, 2024.

Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: Toward real-time reconstruction of visual perception. In *International Conference on Learning Representations*, 2024.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, pp. 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*, pp. 22710–22720, 2023.

Simon Dahan, Gabriel Bénédict, Logan Williams, Yourong Guo, Daniel Rueckert, Robert Leech, and Emma Robinson. SIM: Surface-based fMRI analysis for inter-subject multimodal decoding from movie-watching experiments. In *International Conference on Learning Representations*, 2025.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10041–10071, 2024.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Adrien Doerig, Tim C. Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. High-level visual representations in the human brain are aligned with large language models. *Nature Machine Intelligence*, 7:1220–1234, 2025.

Karl J. Friston, Oliver Josephs, Geraint Rees, and Robert Turner. Nonlinear event-related responses in fMRI. *Magnetic Resonance in Medicine*, 39(1):41–52, 1998.

Matthew F. Glasser, Timothy S. Coalson, Emma C. Robinson, Carl D. Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F. Beckmann, Mark Jenkinson, Stephen M. Smith, and David C. Van Essen. A multi-modal parcellation of human cerebral cortex. *Nature*, 536:171–178, 2016.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling*, 2024.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

Martin N. Hebart, Oliver Contier, Lina Teichmann, Adam H. Rockter, Charles Y. Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I. Baker. THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12:e82580, 2023.

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 *Proceedings of the European Conference on Computer Vision*, 2024.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Xiangtao Kong, Kexin Huang, Ping Li, and Lei Zhang. Toward generalizing visual brain decoding to unseen subjects. In *International Conference on Learning Representations*, 2025.

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.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via EEG embeddings with guided diffusion. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Dongyang Li, Haoyang Qin, Mingyang Wu, Chen Wei, and Quanying Liu. BrainFLORA: Uncovering brain concept representation via multimodal neural embeddings. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025a.

Haoyu Li, Hao Wu, and Badong Chen. NeuralDiffuser: Neuroscience-inspired diffusion guidance for fMRI visual reconstruction. *IEEE Transactions on Image Processing*, 2025b.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning*, volume 162, pp. 12888–12900, 2022.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018.

Martin A. Lindquist, Ji Meng Loh, Lauren Y. Atlas, and Tor D. Wager. Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling. *NeuroImage*, 45(1 Suppl):S187–S198, 2009.

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

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

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

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 *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763, 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*, 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. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Paul Steven Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, 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 International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44038–44059, 2024.

Jimmy T. H. Smith, Andrew Warrington, and Scott W. Linderman. Simplified state space layers for sequence modeling. In *International Conference on Learning Representations*, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

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*, pp. 14453–14463, 2023.

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

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4566–4575, 2015.

Haibao Wang, Jun Kai Ho, Fan L. Cheng, Shuntaro C. Aoki, Yusuke Muraki, Misato Tanaka, Jong-Yun Park, and Yukiyasu Kamitani. Inter-individual and inter-site neural code conversion without shared stimuli. *Nature Computational Science*, 5:534–546, 2025.

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

Haitao Wu, Qing Li, Changqing Zhang, Zhen He, and Xiaomin Ying. Bridging the vision-brain gap with an uncertainty-aware blur prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2246–2257, 2025.

Weihao Xia and Cengiz Oztireli. Exploring the visual feature space for multimodal neural decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4370–4379, 2025.

Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. DREAM: Visual decoding from reversing human visual system. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 8226–8235, 2024.

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

Jacob Yeung, Andrew F. Luo, Gabriel Sarch, Margaret M. Henderson, Deva Ramanan, and Michael J. Tarr. Reanimating images using neural representations of dynamic stimuli. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5331–5343, 2025.

Chenglong Yu, Shuai Shen, Xiangsheng Li, and Yang Li. D^2-FOSA: Dual-diffusion guided EEG-to-image reconstruction with frequency-oriented semantic alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Qi Zhang, Zixuan Gong, Zhuojia Wu, and Duoqian Miao. MindSimulator: Exploring brain concept localization via synthetic fMRI. In *International Conference on Learning Representations*, 2025.

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

A SUPPLEMENTARY ORGANIZATION AND NOTATION

This supplement provides implementation details, theoretical analysis, evaluation protocols, and additional results for DCG-CHI. Appendix B describes the encoder, hemodynamic views, and training procedure. Appendix C provides the operator proofs. Appendices D and E document evaluation and controlled comparisons. Subsequent sections expand the optimization, qualitative, intervention, subject-transfer, retrieval, robustness, and resource analyses presented in the main text. Brain time t, physical lag $k\Delta t$, and diffusion timestep τ are kept distinct throughout.

Table 3: Notation. Brain time t and diffusion time τ index different processes.

Symbol	Meaning
T, R, d	Sequence length, number of ROIs, and state width
X, M	BOLD sequence and binary validity mask
L_G, D_t	Normalized graph Laplacian and current ROI validity diagonal
M_q, K, J	Propagation modes, retained causal lags, and polynomial order
u_t, h_t, v_t	Block input, recurrent state, and residual output
$\pi_{t,m}, q_m(k)$	Input-dependent mode weight and normalized delay mass
α_m, β_m, A_m	Temporal decay, graph diffusion rate, and channel transition
H, H^a, H^b	Encoded BOLD trajectory and re-encoded synthetic-view trajectories
$\widetilde{n}, \psi_v, \mathcal{O}_v$	Projected latent trajectory, HRF kernel, and synthetic observation operator
c_s, c_g	Semantic and structural conditioning tokens
τ, z_τ	Diffusion step and noisy image latent
Q	Fixed, direction-aligned eight-metric quality index

B IMPLEMENTATION DETAILS

B.1 INPUT REPRESENTATION, GRAPH, AND MASKING

The encoder accepts $X \in \mathbb{R}^{B\times T\times R}$ and a validity mask of shape $B \times T$ or $B \times T \times R$. A temporal mask is broadcast across ROIs. A shared affine map lifts each measurement and its validity indicator to the state width, followed by LayerNorm and SiLU. Invalid outputs remain zero. We use HCP-MMP1.0 (Glasser et al., 2016), with 180 ROIs per hemisphere ($R = 360$), ordered left then right consistently for inputs and graph nodes. Regional signals are equal-weight voxel means for NSD and surface-vertex means for THINGS-fMRI. Each trial is represented by a 15×360 sequence. Out-of-range samples are zero-filled and masked. NSD normalization uses fixed training-set subject–ROI statistics. THINGS-fMRI uses run-level preprocessing followed by fixed ROI normalization.

For symmetric nonnegative adjacency W and degree matrix D,

$$L_G = I_R - D^{-1/2} W D^{-1/2}. \tag{14}$$

The controlled configuration constructs a top-8 positive-correlation graph from training data, takes the symmetric union of the neighbor sets, removes negative edges, and uses the normalized Laplacian.

Before every recurrent update, retained states are masked by the current ROI validity pattern. With $D_t = \mathrm{diag}(M_{t,:})$, the history buffer is overwritten as $h_{t-k} \leftarrow D_t h_{t-k}$ before graph propagation. New states are normalized, passed through dropout, and masked. Residual outputs add the block input before a final shared normalization and mask. The buffer stores h_t, whereas v_t feeds the next block.

The encoder returns the final block's masked, normalized residual output, $H_t = v_t^{(N_{\mathrm{blk}})}$. Recurrent history is initialized to zero for each encoder call. The controlled encoder uses six blocks of width 64, four propagation modes, and Chebyshev order six. On the preprocessed $\Delta t = 1$ s model grid, its $K = 8$ positive lags correspond to 1–8 seconds (Table 8).

B.2 DELAY AND TRANSITION PARAMETERIZATION

The delay distribution in Equation 3 uses

$$\mu_m = 1 + (K-1)\,\mathrm{sigmoid}(a_m), \qquad \sigma_m = 0.35 + \mathrm{softplus}(b_m). \tag{15}$$

The positive temporal and graph rates are

$$\alpha_m = \mathrm{softplus}(a'_m), \qquad \beta_m = \mathrm{softplus}(b'_m) + 10^{-4}. \tag{16}$$

Delay mass is indexed by steps. Decay and graph propagation depend on physical lag $k\Delta t$. The learned delay distribution governs recurrent history and is distinct from the HRF pair used to synthesize CHI observations. The channel transition is parameterized as

$$A_m = \widetilde{A}_m \min\left(1, \frac{c_A}{\max(\|\widetilde{A}_m\|_F, 10^{-8})}\right), \qquad c_A = 0.95. \tag{17}$$

Thus $\|A_m\|_2 \le \|A_m\|_F \le c_A$. This bounds the linear channel map. Appendix C derives the corresponding sufficient condition for contraction of the recurrent state update.

B.3 Sparse polynomial evaluation

Let $\widetilde{L} = L_G - I_R$, with spectrum in $[-1, 1]$. For a retained state h,

$$\begin{aligned} T_0(\widetilde{L})h &= h, & T_1(\widetilde{L})h &= \widetilde{L}h, \\ T_{j+1}(\widetilde{L})h &= 2\widetilde{L}T_j(\widetilde{L})h - T_{j-1}(\widetilde{L})h. \end{aligned} \tag{18}$$

At interpolation nodes $\theta_i = \pi(i + 1/2)/N_{\text{int}}$, the coefficients are

$$c_j(s) = \frac{2 - \mathbf{1}_{\{j=0\}}}{N_{\text{int}}} \sum_{i=0}^{N_{\text{int}}-1} e^{-s(1+\cos\theta_i)} \cos(j\theta_i), \qquad N_{\text{int}} = 32. \tag{19}$$

The $j = 0$ normalization is already included in $c_0(s)$. No additional factor of one half is applied in the polynomial sum. The coefficients are differentiable with respect to the diffusion scale and hence β_m. For each retained state, the polynomial basis is shared across propagation modes. Different retained states require their own bases. No eigendecomposition is performed in the forward pass.

The DCG-SSM block is illustrated in Figure 3A.

B.4 HRF parameterization and observation views

The implemented HRF is a difference of two Gaussian-shaped components on a causal grid. The two kernels define a synthetic observation family shared across ROIs. Trainable base scalars a, b, c parameterize the peak location, width, and undershoot amplitude:

$$p = 3 + 5\,\text{sigmoid}(a), \qquad w = 0.8 + \text{softplus}(b), \qquad \eta = 0.05 + 0.25\,\text{sigmoid}(c). \tag{20}$$

The two views share these parameters and use fixed peak offsets and width scales:

$$\begin{aligned} p_v &= \max(p + \delta_v, \Delta t), & w_v &= \max(w\kappa_v, 0.35), \\ (\delta_a, \kappa_a) &= (-0.5\,\text{s}, 0.85), & (\delta_b, \kappa_b) &= (0.75\,\text{s}, 1.20). \end{aligned} \tag{21}$$

For nonnegative grid time s, define

$$\widetilde{\psi}_v(s) = \exp\left[-\frac{(s - p_v)^2}{2w_v^2}\right] - \eta \exp\left[-\frac{(s - p_v - 6)^2}{2(1.6w_v)^2}\right]. \tag{22}$$

Each kernel is divided by its discrete sum, with the denominator clamped below at 10^{-6}. The default bank uses 13 samples at $s = 0, 1, \ldots, 12$ seconds and causal zero padding. This is distinct from the 15-point input window at stimulus-relative times $-2, \ldots, 12$ seconds. The same HRF pair is shared across ROIs. The fixed double-Gamma test filter is specified separately in Appendix K.

A shared scalar projection gives $\widetilde{n} = p_\omega(H)$. We use $\widetilde{n}$ to distinguish this task-trained latent projection from the unobserved neural drive n in Equation 1. The view operator is

$$\mathcal{O}_v(U) = \psi_v *_{\text{causal}} p_\omega(U), \qquad X^v = \mathcal{O}_v(H), \qquad H^v = E_\theta(X^v, M, L_G). \tag{23}$$

Counterfactual refers to changing the synthetic observation operator for a shared latent trajectory. The original and synthetic observations use the same encoder parameters, graph, and mask.

The CHI training branches are illustrated in Figure 3B.

B.5 Loss reduction, gradient paths, and regularization

For states with d_U channels, the masked error averages over valid scalar entries, including feature channels. Suppressing the batch axis,

$$\text{MSE}_M(U, V) = \frac{\sum_{t,r} M_{t,r} \|U_{t,r,:} - V_{t,r,:}\|_2^2}{d_U \max(1, \sum_{t,r} M_{t,r})}. \tag{24}$$

The scalar-observation case has $d_U = 1$. CHI compares complete state trajectories. The anchor target $\text{sg}(H)$ is detached, while synthesis and shared re-encoding remain differentiable. The BOLD re-synthesis target is the synthetic observation X^v from the common HRF cycle, retained in all C0–C3 variants. Write its recovered states as $H^{h,v}$. In the full model, $H^{h,v} = H^v$ in Equation 7. Diffusion and semantic supervision connect these representation constraints to image reconstruction.

Default coefficients in Equation 12 are $\lambda_s = 0.1$, $\lambda_c = 0.1$, $\lambda_b = 0.05$, and $\lambda_r = 0.001$. The mode-wise recurrent-gain diagnostic and its layer summary are

$$\mathfrak{b}_{\ell,m} = 0.9\ \text{sigmoid}(s_\ell) \sum_{k=1}^{K} q_{\ell,m}(k) e^{-\alpha_{\ell,m} k \Delta t}, \qquad \mathfrak{b}_\ell = \max_m \mathfrak{b}_{\ell,m}, \tag{25}$$

where s_ℓ is the layer gain logit, $\alpha_{\ell,m} = \mathrm{softplus}(a'_{\ell,m})$, and $q_{\ell,m}$ is normalized over lags. This auxiliary diagnostic is distinct from the contraction condition in Appendix C and from the training regularizer, which controls the mean recovered-state variance:

$$\mathcal{L}_{\mathrm{reg}} = \left[\frac{1}{Rd} \sum_{r=1}^{R} \sum_{j=1}^{d} \mathrm{Var}_{(n,t):M_{ntr}=1} \left(H^{h,a}_{ntrj} \right) - 1 \right]^2 . \tag{26}$$

Here n indexes batch examples. Population variance is computed across valid sample–time positions in the first common HRF branch. These variances are averaged over ROI–channel pairs before taking the squared deviation from one. The reported configuration uses $p_{\mathrm{CHI}} = 1$. Active consistency views and the common HRF cycle are evaluated at every training step.

B.6 Training, conditioning, and inference

Independent learned-query readouts form semantic and structural tokens from masked temporal ROI means. Semantic tokens condition the cross-attention layers of the pretrained Stable Diffusion 2.1 U-Net. Structural tokens are mapped to a spatial hint for a trainable ControlNet branch supplying 13 residual controls (Huo et al., 2024; Zhang et al., 2023). The semantic and structural token arrays have shapes 77×1024 and 257×1024, respectively. The U-Net and VAE remain frozen, while the brain encoder, readouts, and control pathway are trained. The full training objective is given in Equation 12. BLIP captions and a frozen CLIP-family text encoder provide semantic targets (Li et al., 2022; Radford et al., 2021). Image training uses the latent-diffusion noise objective (Rombach et al., 2022), and inference uses DDIM sampling (Song et al., 2021).

Table 4: Training and inference procedures. Frozen U-Net weights are not updated, but gradients through its computation reach the trainable conditioning modules.

Stage	Procedure
Training 1	Read (X, M, I) and the training-image caption. Load the fixed training-derived graph.
Training 2	Compute $H = E_\theta(X, M, L_G)$ and masked temporal ROI means.
Training 3	Obtain c_s and c_g using the two independent attention readouts.
Training 4	Encode I with the frozen VAE, sample diffusion noise and time, and predict noise with the U-Net and trainable controls.
Training 5	Calculate token-wise semantic MSE against frozen text-encoder targets when captions are supplied.
Training 6	Evaluate the common HRF cycle, BOLD re-synthesis and variance losses. Apply the variant's active consistency losses (Table 7).
Training 7	Back-propagate the weighted total loss and update only trainable components.
Inference 1	Encode the BOLD sequence and obtain semantic and structural tokens from the input, mask, and graph.
Inference 2	Initialize the image latent with random noise and run the declared sampler with inferred conditions.
Inference 3	Decode the final latent using the frozen VAE, without stimulus images, captions, or image embeddings.

When enabled, semantic dropout replaces semantic tokens with the frozen empty-text embedding while retaining structural conditions. The controlled comparisons use AdamW (Loshchilov & Hutter, 2019). Optimization and sampling settings are listed in Table 8.

B.7 Training and inference algorithm

Algorithm 1 summarizes the default full-model procedure with $p_{\mathrm{CHI}} = 1$. Let Θ contain encoder parameters θ, independent readout parameters φ_s, φ_g, control-pathway parameters ξ, observation-projection parameters ω, and HRF-bank parameters ζ. The control pathway includes the spatial-hint mapping and ControlNet. $\mathrm{Readouts}(H, M)$ performs masked temporal pooling and independent semantic and structural attention readout, excluding ROIs without valid observations from the attention keys. Training captions y are generated from training images using BLIP. $\mathcal{T}$ denotes the frozen text encoder. $\mathrm{AddNoise}$ follows the pretrained diffusion schedule, and $\mathrm{GuidedPredict}$ uses the frozen U-Net and trained control pathway with the configured classifier-free guidance rule. Losses are reduced over the minibatch. MSE_M additionally averages over valid entries and feature channels. Each encoder call starts from zero recurrent history. Gradients through the frozen U-Net update the trainable conditioning pathways.

C Operator analysis

C.1 Equivariance to consistent ROI relabeling

We retain the main text's row-vector convention for $X \in \mathbb{R}^{T \times R}$. At a fixed time, $H_t \in \mathbb{R}^{R \times d}$ stores ROI features by row.

Algorithm 1 Training and inference of DCG-CHI

Require: Training data $\mathcal{D}_{\mathrm{tr}} = \{(X, M, I, y)\}$; validation data $\mathcal{D}_{\mathrm{val}}$; fixed training-derived graph L_G.
Require: Optimization steps N_{opt}; validation interval V; loss weights $\lambda_s, \lambda_c, \lambda_b, \lambda_r$.
Require: Test input $(X_{\mathrm{test}}, M_{\mathrm{test}})$; DDIM steps N, guidance scale γ, and schedule $\tau_N > \cdots > \tau_0 = 0$.
Ensure: Selected parameters Θ^* and reconstructed image $\widehat{I}$.

Training

Freeze the VAE encoder/decoder, text encoder $\mathcal{T}$, and SD2.1 U-Net ϵ_{SD}.
for $j = 1, \ldots, N_{\mathrm{opt}}$ **do**
  Sample a minibatch (X, M, I, y) from $\mathcal{D}_{\mathrm{tr}}$.
  $H \leftarrow E_\theta(X, M, L_G)$; $(c_s, c_g) \leftarrow \mathrm{Readouts}_{\varphi_s, \varphi_g}(H, M)$.
  $z_0 \leftarrow \mathcal{E}_{\mathrm{VAE}}(I)$; $c_I \leftarrow \mathcal{T}(y)$; sample diffusion step τ and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$.
  $z_\tau \leftarrow \mathrm{AddNoise}(z_0, \tau, \epsilon)$.
  $r_\tau \leftarrow \mathrm{Control}_\xi(c_g, z_\tau, \tau, c_s)$; $\widehat{\epsilon} \leftarrow \epsilon_{\mathrm{SD}}(z_\tau, \tau; c_s, r_\tau)$.
  $\mathcal{L}_{\mathrm{diff}} \leftarrow \|\epsilon - \widehat{\epsilon}\|_2^2$; $\mathcal{L}_{\mathrm{sem}} \leftarrow \mathrm{MSE}(c_s, c_I)$.
  $(\psi_a, \psi_b) \leftarrow \mathrm{HRFBank}_\zeta$; $\widetilde{n} \leftarrow p_\omega(H)$; $H_{\mathrm{anc}} \leftarrow \mathrm{sg}(H)$.
  for $v \in \{a, b\}$ **do**
    $X^v \leftarrow \psi_v *_{\mathrm{causal}} \widetilde{n}$.
    $H^v \leftarrow E_\theta(X^v, M, L_G)$.
    $\widehat{X}^v \leftarrow \psi_v *_{\mathrm{causal}} p_\omega(H^v)$.
  end for
  $\mathcal{L}_{\mathrm{cross}} \leftarrow \mathrm{MSE}_M(H^a, H^b)$.
  $\mathcal{L}_{\mathrm{anchor}} \leftarrow \sum_{v \in \{a,b\}} \mathrm{MSE}_M(H^v, H_{\mathrm{anc}})$.
  $\mathcal{L}_{\mathrm{CHI}} \leftarrow \mathcal{L}_{\mathrm{cross}} + \mathcal{L}_{\mathrm{anchor}}$; $\mathcal{L}_{\mathrm{BOLD}} \leftarrow \sum_{v \in \{a,b\}} \mathrm{MSE}_M(\widehat{X}^v, X^v)$.
  Compute $\mathcal{L}_{\mathrm{reg}}$ from the population variance of H^a (Equation 26).
  $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{diff}} + \lambda_s \mathcal{L}_{\mathrm{sem}} + \lambda_c \mathcal{L}_{\mathrm{CHI}} + \lambda_b \mathcal{L}_{\mathrm{BOLD}} + \lambda_r \mathcal{L}_{\mathrm{reg}}$.
  Update Θ with AdamW using $\nabla_\Theta \mathcal{L}$; keep pretrained weights fixed.
  if $j \bmod V = 0$ **then**
    Evaluate Q on $\mathcal{D}_{\mathrm{val}}$; retain the best checkpoint Θ^*, preferring earlier steps in a tie.
  end if
end for

Inference

Load Θ^*, switch to evaluation mode, and omit the HRF bank and synthetic-view branches.
$H_{\mathrm{test}} \leftarrow E_{\theta^*}(X_{\mathrm{test}}, M_{\mathrm{test}}, L_G)$.
$(c_s, c_g) \leftarrow \mathrm{Readouts}_{\varphi_s^*, \varphi_g^*}(H_{\mathrm{test}}, M_{\mathrm{test}})$; sample $z_{\tau_N} \sim \mathcal{N}(0, \mathbf{I})$.
for $i = N, N-1, \ldots, 1$ **do**
  $\widehat{\epsilon}_i \leftarrow \mathrm{GuidedPredict}_{\xi^*}(z_{\tau_i}, \tau_i; c_s, c_g, \gamma)$.
  $z_{\tau_{i-1}} \leftarrow \mathrm{DDIMStep}(z_{\tau_i}, \widehat{\epsilon}_i, \tau_i, \tau_{i-1})$.
end for
return Θ^* and $\widehat{I} = \mathcal{D}_{\mathrm{VAE}}(z_0)$.

Proposition 1 (Consistent ROI relabeling)**.** *Let $P \in \mathbb{R}^{R \times R}$ be an ROI permutation matrix and let $L'_G = P L_G P^\top$. With shared ROI-wise maps and deterministic evaluation,*

$$E_\theta(XP^\top, MP^\top, L'_G)_t = P E_\theta(X, M, L_G)_t \tag{27}$$

for every t. Both learned-query readouts are invariant under the same joint relabeling.

Proof. The input lift commutes with P because its parameters are shared across ROIs. Consistently permuting features and the validity mask leaves the masked ROI mean unchanged, so the selected mode weights are unchanged. For any polynomial g,

$$g(P L_G P^\top) = P g(L_G) P^\top, \tag{28}$$

and the same identity holds for the matrix exponential. Channel transforms act on a different tensor axis and therefore commute with ROI permutation. The zero history is equivariant. Induction over time, including current-mask updates of the retained buffer, establishes equivariance of the propagated states. Pointwise nonlinearities, channel LayerNorm, masking, and residual updates preserve this relation. Induction over blocks gives the encoder result.

For a query readout, joint permutation of keys and values permutes the attention-weight columns and value rows correspondingly, leaving the weighted sum unchanged. The same holds for a consistently permuted attention-key mask. Thus both readouts are invariant. Independent dropout masks preserve equivariance in distribution. Exact samplewise equality requires disabled dropout or correspondingly permuted dropout masks. □

This property concerns consistent relabeling within a fixed ROI system.

C.2 A SUFFICIENT CONTRACTION CONDITION

Consider one recurrent block with fixed inputs, masks, and mode weights under deterministic evaluation. Write its update as

$$h_t = \sigma_t\left(b_t + \sum_{k=1}^{K} \widehat{\Phi}_{t,k}[h_{t-k}]\right), \tag{29}$$

where $\widehat{\Phi}_{t,k}$ includes the implemented polynomial filters and masking operations. The map σ_t includes the normalized nonlinearity. Assume that these maps are uniformly L_σ-Lipschitz, using the Frobenius norm on states.

Proposition 2 (Conditional contraction and boundedness)**.** *If*

$$\rho = L_\sigma \sup_t \sum_{k=1}^{K} \|\widehat{\Phi}_{t,k}\|_{F\to F} < 1, \tag{30}$$

then the maximum difference over the retained histories of two trajectories with identical inputs, masks, and mode weights is reduced to at most ρ times its preceding bound after every K updates. If also $\|\sigma_t(b_t)\|_F \le C$, their state magnitudes are bounded by $\max(H_0, C/(1-\rho))$, where H_0 bounds the initial history.

Proof. For the difference e_t between two trajectories, Lipschitz continuity gives

$$\|e_t\|_F \le L_\sigma \sum_{k=1}^{K} \|\widehat{\Phi}_{t,k}\|_{F\to F}\|e_{t-k}\|_F \le \rho \max_{1\le k\le K} \|e_{t-k}\|_F. \tag{31}$$

The running maximum over the retained K-step history cannot increase. After K updates, every entry in the previous history has been replaced by an entry bounded by ρ times the previous window maximum. Iterating this argument establishes geometric forgetting over windows. Similarly,

$$\|h_t\|_F \le C + \rho \max_{1\le k\le K} \|h_{t-k}\|_F. \tag{32}$$

Induction shows that $\max(H_0, C/(1-\rho))$ is an invariant upper bound. □

For the exact heat kernel, symmetry and positive semidefiniteness of L_G imply $\|e^{-sL_G}\|_2 \le 1$ for $s \ge 0$. Since masking is nonexpansive,

$$\sum_k \|\Phi_{t,k}\|_{F\to F} \le \sum_m \pi_{t,m}\|A_m\|_2 \sum_k q_m(k) e^{-\alpha_m k\Delta t} \le c_A. \tag{33}$$

Thus exact-filter contraction follows when $L_\sigma c_A < 1$. LayerNorm must be included in L_σ. For a polynomial filter $\widehat{P}_{m,k}$ satisfying $\|\widehat{P}_{m,k} - e^{-\beta_m k\Delta t L_G}\|_2 \le \varepsilon_{m,k}$, the corresponding sufficient history bound is

$$\sum_m \pi_{t,m}\|A_m\|_2 \sum_k q_m(k) e^{-\alpha_m k\Delta t}(1+\varepsilon_{m,k}). \tag{34}$$

The implemented block's sufficient condition combines this approximation-aware history bound with the Lipschitz constant of the normalized nonlinearity.

C.3 SPECTRAL INTERPRETATION AND COMPLEXITY

For an exact Laplacian eigenpair (λ_i, v_i), the graph filter in a mode–delay term multiplies the corresponding spatial component by $e^{-\beta_m k\Delta t\lambda_i}$. Within each mode, longer lags therefore attenuate high graph frequencies more strongly before channel mixing and normalization.

For each retained state, J sparse polynomial basis evaluations cost $O(J|E|d)$, mode-wise basis combination costs $O(M_q JRd)$, and dense channel transforms cost $O(M_q Rd^2)$. Summing over T steps and K lags gives the per-block cost

$$O\big(TK[J|E|d + M_q(JRd + Rd^2)] + TRd^2\big), \tag{35}$$

up to selector and coefficient-generation overhead. For a dense graph, each sparse graph multiplication is replaced by $O(R^2 d)$ work. For fixed architectural budgets, runtime is linear in sequence length. The inference history requires $O(KRd)$ memory per block. Full-sequence back-propagation stores additional activations. For the full model, the HRF cycle adds two shared-encoder passes.

D EVALUATION SETTINGS

D.1 DATASETS, SPLITS, AND COMPARISON SCOPE

We evaluate NSD subjects S01, S02, S05, and S07 (Allen et al., 2022) and three THINGS-fMRI subjects (Hebart et al., 2023). Inputs are trial-centered ROI sequences. The controlled protocol retains the official test stimuli and assigns remaining eligible stimulus identities to training and validation at 90%/10% using a fixed hash. Repetitions are grouped by stimulus identity. Each input contains 15 samples at stimulus-relative times $-2, -1, \ldots, 12$ seconds on a preprocessed 1-second grid. Invalid or split-conflicting windows are masked or excluded. Functional correlations use training data only. Dataset-specific normalization is described in Appendix B. Caption-derived targets concern training images. Inference uses fMRI-derived conditions.

External-method results are taken from the cited papers or obtained from reproductions based on the papers or released implementations for evaluations not reported there (‡). Paper-reported results retain their original evaluation protocols. MindHier uses an updated 1,000-image NSD test set. The 982-entry retrieval gallery here specifies the candidate pool for retrieval. Models are evaluated separately on the two datasets.

Acquisition and model time grids. The acquisition TR is 1.6 s for NSD and 1.5 s for THINGS-fMRI. NSD sequences are slice-time corrected and cubically interpolated to 1 s. THINGS-fMRI sequences are resampled to an intermediate 0.5-second grid using polyphase resampling with linear boundary extension. Stimulus-aligned windows then select every other sample. Both datasets therefore use $\Delta t = 1$ s at the model input, independently of acquisition TR.

D.2 PRIMARY RECONSTRUCTION METRICS

PixCorr is pixelwise Pearson correlation after preprocessing. SSIM measures local luminance, contrast, and structure (Wang et al., 2004). AlexNet-2/5, Inception, and CLIP identification measure feature-space discriminability (Krizhevsky et al., 2012; Szegedy et al., 2016; Radford et al., 2021). EfficientNet-B and SwAV use feature correlation distances (Tan & Le, 2019; Caron et al., 2020). All eight primary metrics are reported in their native units.

For feature extractor f and similarity s_f, pairwise identification of reconstruction $\widehat{I}_i$ has the form

$$\mathrm{ID}_i = \frac{1}{N_{\mathrm{eval}} - 1} \sum_{j \neq i} \mathbf{1}\Big\{ s_f(f(\widehat{I}_i), f(I_i)) > s_f(f(\widehat{I}_i), f(I_j)) \Big\}. \tag{36}$$

For feature correlation distance,

$$d_f = 1 - \mathrm{corr}\Big(f(\widehat{I}), f(I) \Big), \tag{37}$$

smaller values are better. Paper-reported external feature-space scores are retained as reported. Table 5 gives all eight scores for each subject. Subjectwise and dataset means are rounded separately.

Table 5: Subjectwise reconstruction results for the full model. Each row reports all eight primary metrics.

	Subject	PixCorr↑	SSIM↑	Alex2↑	Alex5↑	Incep.↑	CLIP↑	Eff-B↓	SwAV↓
NSD	S01	0.347	0.484	0.985	0.999	0.986	0.985	0.597	0.307
	S02	0.339	0.476	0.976	0.999	0.976	0.974	0.609	0.320
	S05	0.327	0.462	0.961	0.988	0.958	0.955	0.629	0.341
	S07	0.316	0.450	0.949	0.974	0.944	0.939	0.646	0.359
THINGS-fMRI	S01	0.127	0.380	0.750	0.862	0.766	0.774	0.804	0.543
	S02	0.114	0.366	0.735	0.845	0.748	0.755	0.825	0.565
	S03	0.101	0.352	0.720	0.829	0.730	0.736	0.845	0.587

D.3 QUALITY INDEX AND STATISTICAL REPORTING

The fixed direction-aligned quality index used for validation checkpoint selection is

$$Q = \frac{1}{8} \sum_{i=1}^{8} \mathrm{clip}(a_i, 0, 1), \quad \boldsymbol{a} = \left(\frac{p}{0.40}, \frac{s}{0.50}, A_2, A_5, I_{\mathrm{m}}, C, \frac{1 - E}{0.75}, \frac{1 - W_{\mathrm{m}}}{0.75} \right). \tag{38}$$

Here p, s denote PixCorr and SSIM; $A_2, A_5, I_{\mathrm{m}}, C$ are the four identification scores; and E, W_{m} are EfficientNet-B and SwAV distances. The subscript distinguishes these metric symbols from an image I or adjacency W. Applied to the rounded full-model dataset means, it gives 0.8813 on NSD and 0.6142 on THINGS-fMRI. These derived references are denoted by Q_{tab} below and are distinguished from subject-first aggregation.

Across-subject SD describes participant variation. Generation-seed variation describes stochastic sampling with the trained model fixed. Tail-loss SD describes optimization fluctuations. Protocol and temporal-intervention evaluation use this quality scale. Representation consistency and reliability concern different measurement objects, as detailed below.

D.4 Definitions of auxiliary evaluation metrics

Temporal interventions and subject-transfer protocols share a reconstruction-quality scale. We also define internal representation consistency for condition exchange. Results are aggregated independently for NSD and THINGS-fMRI.

Measurement and normalization. Let s index a target subject, r an experimental repeat, and e an evaluation condition. For the fixed test stimuli $I_{s,i}$ and their reconstructions $\widehat{I}^{(e)}_{s,r,i}$, first compute the eight primary metrics over the complete test set for that subject and repeat:

$$\mathbf{m}_{s,r}(e) = \big(p,\, v,\, A_2,\, A_5,\, I_{\mathrm{m}},\, C,\, E,\, W_{\mathrm{m}}\big)_{s,r,e}. \tag{39}$$

Here p is PixCorr, v is SSIM, $A_2, A_5, I_{\mathrm{m}}, C$ are the AlexNet-2, AlexNet-5, Inception, and CLIP two-way identification rates, and E, W_{m} are the EfficientNet-B and SwAV feature correlation distances. The identification rates use frozen features and are expressed as fractions in $[0, 1]$. The symbol v distinguishes SSIM from the subject index s.

The direction-aligned component transformations and the quality index are

$$g_j(x) = \begin{cases} \mathrm{clip}(x/0.40, 0, 1), & j = 1, \\ \mathrm{clip}(x/0.50, 0, 1), & j = 2, \\ x, & j \in \{3, 4, 5, 6\}, \\ \mathrm{clip}((1-x)/0.75, 0, 1), & j \in \{7, 8\}, \end{cases} \tag{40}$$

$$Q_{s,r}(e) = \frac{1}{8} \sum_{j=1}^{8} g_j(m_{s,r,j}(e)) \,. \tag{41}$$

Every component has weight $1/8$. The constants 0.40, 0.50, and 0.75 are fixed scales shared by all methods and conditions.

For a condition e, the dataset-level quality score is

$$\overline{Q}(e) = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{R_s} \sum_{r=1}^{R_s} Q_{s,r}(e). \tag{42}$$

Thus the order is test-set metrics within a subject and repeat, component transformations, equal-weight quality index, repeat average, and equal-weight subject average. Clipping is applied before averaging across repeats and subjects. The index summarizes reconstruction quality on fixed component scales.

References derived from reported means. For the rounded dataset-level metric vector $\mathbf{m}_{\mathrm{tab}}$, the corresponding quality reference is

$$Q_{\mathrm{tab}} = \frac{1}{8} \sum_{j=1}^{8} g_j(m_{\mathrm{tab},j}). \tag{43}$$

Here Q_{tab} summarizes the reported table means. Clipping and rounding can make it differ from the subject-first mean $\overline{Q}$.

Temporal interventions. For intervention c, define

$$T_{s,r}(c) = Q_{s,r}(c), \qquad T(c) = \overline{Q}(c). \tag{44}$$

Normal and intervened conditions are evaluated against the same test stimuli with the same metric configuration and fixed quality scales.

Subject protocols and target adaptation. For training or adaptation protocol p, define

$$L_{s,r}(p) = Q_{s,r}(p), \qquad L_{\mathrm{protocol}}(p) = \overline{Q}(p). \tag{45}$$

Protocols include within-subject reconstruction, source-to-target transfer, joint training, zero-shot evaluation, and target adaptation. Comparisons within each target subject use the same test stimuli and evaluation configuration. The higher-is-better score L_{protocol} uses the same transformations and weights as T, indexed by protocol p.

Changes in structural similarity, semantic identification, and aggregate quality are expressed in the units of SSIM, CLIP identification, and Q. Let $\overline{m}_j(p) = S^{-1} \sum_s R_s^{-1} \sum_r m_{s,r,j}(p)$. The ten-shot

versus zero-shot differences are

$$\Delta\mathrm{SSIM} = \overline{m}_2(\text{10-shot}) - \overline{m}_2(\text{0-shot}),$$
$$\Delta\mathrm{CLIP} = \overline{m}_6(\text{10-shot}) - \overline{m}_6(\text{0-shot}),$$
$$\Delta Q = \overline{Q}(\text{10-shot}) - \overline{Q}(\text{0-shot}). \tag{46}$$

The first two changes concern structural similarity and semantic identification, respectively. ΔQ is a change in quality-index units.

Structural and semantic representation consistency. For test stimulus i, subject s, and repeat r, let $\mathbf{h}^{(c)}_{s,r,i}$ and $\mathbf{z}^{(c)}_{s,r,i}$ denote the structural control tensor and semantic conditioning tensor under exchange condition c. Superscript 0 denotes the corresponding unperturbed tensors for the same sample. Corresponding internal conditioning tensors are taken at the same extraction point with matching shape and index order.

Define $\mathbf{v}^{(c)}_{\mathrm{str},s,r,i} = \mathrm{vec}(\mathbf{h}^{(c)}_{s,r,i})$ and $\mathbf{v}^{(c)}_{\mathrm{sem},s,r,i} = \mathrm{vec}(\mathbf{z}^{(c)}_{s,r,i})$. For $k \in \{\mathrm{str}, \mathrm{sem}\}$, the sample-level consistency is

$$C^{(c)}_{k,s,r,i} = \frac{\left\langle \mathbf{v}^{(c)}_{k,s,r,i}, \mathbf{v}^{(0)}_{k,s,r,i} \right\rangle}{\left\| \mathbf{v}^{(c)}_{k,s,r,i} \right\|_2 \left\| \mathbf{v}^{(0)}_{k,s,r,i} \right\|_2}. \tag{47}$$

The cosine uses L2 normalization and has range $[-1, 1]$. Zero-norm tensors are flagged as undefined. For N_s matched test stimuli, the aggregate is

$$C_k(c) = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{R_s}\sum_{r=1}^{R_s}\frac{1}{N_s}\sum_{i=1}^{N_s} C^{(c)}_{k,s,r,i}. \tag{48}$$

Structural and semantic consistency are reported separately. With nonzero reference tensors, the intact condition compares each tensor with itself and therefore satisfies

$$C_{\mathrm{str}}(0) = C_{\mathrm{sem}}(0) = 1. \tag{49}$$

If conditions are swapped directly after the readouts, the branch whose tensor is unchanged also has consistency one. For a recipient i and donor j, the exchanged branch has cosine $\cos(\mathbf{h}_j, \mathbf{h}_i)$ or $\cos(\mathbf{z}_j, \mathbf{z}_i)$, using the vectorization in Equation 47. The measurement point is immediately after readout exchange. Representation consistency measures proximity to the unperturbed conditioning.

Repeat indexing. The index r denotes a repeat within a subject and condition. Depending on the experiment, it indexes an independent training run, an adaptation-set draw, or an image-generation seed. The corresponding sources of variation are distinguished in Section D.3.

Table 6: Evaluation quantities and their measurement objects.

Quantity	Use and comparison scope
Q	Fixed direction-aligned aggregate of the eight reconstruction metrics; validation checkpoint selection (Equation 38).
$L_{\mathrm{protocol}}(p)$	Subject-first mean $\overline{Q}(p)$ for a training or adaptation protocol (Equation 45).
$L_{\mathrm{reliability}}$	Stimulus agreement multiplied by repeat-observation consistency (Equation 58).
$T(c)$	Subject-first mean $\overline{Q}(c)$ for temporal intervention c (Equation 44).
$C_{\mathrm{str}}, C_{\mathrm{sem}}$	Separate cosine consistencies of the structural and semantic tensors against the same sample without exchange (Equations 47–48).
Input-corruption score	Normalized reconstruction score for Gaussian-noise and missing-input experiments.
$S_{\mathrm{eff}}(h)$	Equal-weight SSIM and CLIP-identification ratios to the same method's full-data result (Equation 59).
Generation consistency	Mean pairwise similarity across eight seeds with fMRI input and checkpoint fixed (Equation 60).

E Control definitions and experimental configuration

E.1 Graph and delay controls

The controlled encoder comparison retains the input representation, readouts, CHI training, and downstream generation architecture. G0 replaces graph–delay blocks with token-wise residual MLPs; G1 retains spatial graph processing without delayed recurrence; G2 retains temporal processing without cross-ROI propagation. G3 uses independent spatial and temporal branches, whereas G_{serial} applies spatial processing before temporal processing. For lifted features U, spatial operator $\mathcal{G}_A$, temporal operator $\mathcal{D}$, and token-wise projection $\mathcal{P}$, the latter two controls are

$$H_{G3} = U + \mathcal{P}([\mathcal{G}_A(U), \mathcal{D}(U)]), \quad H_{\mathrm{serial}} = U + \mathcal{P}(\mathcal{D}(\mathcal{G}_A(U))). \tag{50}$$

G4 is the full model, which couples graph filtering to the delayed transition in Equation 4. These comparisons evaluate complete encoder variants. Results appear in Table 2.

E.2 Consistency controls

Let $\mathcal{L}_{\text{pair}} = \mathcal{L}_{\text{cross}}$ denote the paired-state MSE in Equation 8; the anchor is the *sum* of the two masked errors to $\text{sg}(H)$. The control objectives are

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + 0.1\mathcal{L}_{\text{sem}} + \lambda_p\mathcal{L}_{\text{pair}} + \lambda_a\mathcal{L}_{\text{anchor}} + 0.05\mathcal{L}_{\text{BOLD}} + 0.001\mathcal{L}_{\text{reg}}. \tag{51}$$

For C3, $\lambda_p = \lambda_a = \lambda_c = 0.1$, recovering Equation 12. All four variants retain the same common HRF-cycle BOLD loss and variance regularizer. Only the consistency views and their loss weights change.

Table 7: CHI control settings. Diffusion and semantic-loss weights are 1 and 0.1 throughout. BOLD reconstruction uses the common HRF cycle in every variant.

Variant	Consistency views	λ_p	λ_a	λ_b	λ_r
C0	None	0	0	0.05	0.001
C1	Gaussian noise + masking	0.1	0.1	0.05	0.001
C2	HRF transforms	0.1	0	0.05	0.001
C3	HRF transforms	0.1	0.1	0.05	0.001

C1 constructs two independent views from the unperturbed synthetic observation $X^0 = \mathcal{O}_0(H)$:

$$\widetilde{X}^{(j)} = M \odot A^{(j)} \odot \left(X^0 + 0.05s \odot \epsilon^{(j)}\right), \qquad \epsilon^{(j)} \sim \mathcal{N}(0, I), \quad A^{(j)} \sim \text{Bernoulli}(0.9). \tag{52}$$

Here $\mathcal{O}_0$ uses the shared base HRF without a view offset or width rescaling. The ROI standard deviations s are computed from valid positions of X^0 in the current training batch and detached. Noise has 5% of this scale, and 10% of valid elements are independently masked. The original validity mask M is unchanged for re-encoding and loss reduction. C2 and C3 use the same HRF transforms in Equation 21. Thus, C3–C1 tests the view construction, C3–C2 tests anchoring, and C3–C0 tests the full consistency constraint while preserving the auxiliary losses.

E.3 Component-removal implementation and run identity

The *w/o DCG-SSM* control replaces each recurrent graph–delay block with the token-wise residual MLP branch

$$\text{LN} \to \text{Linear}(d, 2d) \to \text{GELU} \to \text{Dropout} \to \text{Linear}(2d, d). \tag{53}$$

This control removes recurrent graph–delay mixing while retaining masking and downstream readouts, including their temporal aggregation. The starred *w/o CHI** designation identifies the hemodynamic-pathway removal run used in the additional diagnostics. C0 removes the paired and anchor constraints in the primary controlled comparison. Appendix F reports the optimization summary.

E.4 Controlled configuration and comparison sources

Table 8 records the 12,000-step configuration for controlled comparisons, denoted `NEW-12K`.

Table 8: Experimental configuration for the graph–delay and consistency comparisons.

Item	Setting
Data split	Official test stimuli retained; remaining eligible stimulus identities split 90%/10% by a fixed hash; repetitions grouped by identity.
BOLD window	15×360 ROI input at stimulus-relative $-2, -1, \ldots, 12$ seconds on the 1-second grid; out-of-range samples are zero-filled and masked.
Parcellation	HCP-MMP1.0, 180 ROIs per hemisphere; NSD voxel means and THINGS-fMRI surface-vertex means, ordered left then right.
ROI graph	Training-data positive correlations; top-8 neighbors, symmetric union, negative edges removed, normalized Laplacian, fixed ROI order.
Encoder	6 blocks, width 64, 4 modes, Chebyshev order 6; $K = 8$ positive lags at 1–8 seconds on the 1-second model grid.
Optimization	AdamW; learning rate 10^{-4}, weight decay 0.01, bf16; 12,000 steps and effective batch size 36.
Validation	Eight metrics every 500 steps; checkpoint selected by fixed Q in Equation 38; earlier checkpoint preferred in a tie.
Generation	Frozen Stable Diffusion 2.1 U-Net and VAE; trainable control pathway and brain-condition readouts.
Sampling	50 denoising steps, guidance scale 5, batch size 1, 512×512; common generation seed for the primary controls.
Reporting scope	`NEW-12K`: controlled comparisons; short-budget optimization diagnostics appear in Appendix F.

Table 9 lists the paper-reported quantitative results accompanying the reproduced qualitative examples.



Figure 8: **Reconstruction metric profiles.** (A) NSD profiles for NeuroPictor and internal variants. (B) THINGS-fMRI internal variants. (C) Radar profiles with fixed, direction-aware scales. (D) Dataset metrics and selected gallery-quality strata. Literature comparisons appear in Table 1.

Table 9: Paper-reported quantitative results accompanying the reproduced qualitative comparisons.

Dataset	Reference	Reported SSIM↑ / CLIP↑
NSD	NeuroPictor (Huo et al., 2024)	0.375 / 0.933
THINGS-fMRI	BrainFLORA (Li et al., 2025a)	0.348 / 0.576

F Optimization diagnostics

Table 10 summarizes THINGS-fMRI optimization at step 6,990. Tail loss summarizes the training objective active in each configuration. Reconstruction quality is reported through the image metrics in Tables 1 and 2.

Table 10: THINGS-fMRI training summary at step 6,990. Time is measured in seconds per training step.

Configuration	s/step↓	Tail loss	Gradient L_2
DCG-CHI (Ours)	17.52	0.4482	0.2468
w/o DCG-SSM	1.18	0.4824	0.1388
w/o CHI*	9.46	0.4125	0.0309
Graph-only diagnostic	1.11	0.4743	0.1421

Table 11 reports 1,000-step scans, with SD computed over the optimization tail. The delay and mode-count scans in Figure 7E vary each factor independently. "Maximum delay 0" selects an instantaneous graph encoder, retaining current-time graph propagation, readouts, and CHI while removing delayed history. This setting replaces the positive-lag recurrence with current-time graph propagation.

G Direct-output qualitative examples and selection

Figure 4 presents a subset of reconstruction results.

The supplemental strata galleries contain fMRI-conditioned reconstructions selected post hoc using stimulus-based reconstruction metrics. For NSD, the selection score combines direction-aligned percentile ranks of six feature-based metrics with total weight 0.60, plus PixCorr and SSIM percentile ranks with weight 0.20 each. For THINGS-fMRI, the weights are 0.50 for CLIP and 0.25 each for PixCorr and SSIM.

Each dataset shows six distinct stimuli from each of the best, median, and worst strata. Stimuli shown in the main-text comparison (Figure 4) are excluded. The NeuroPictor and BrainFLORA comparison

Table 11: Short-budget THINGS-fMRI optimization diagnostics. Each scan contains 1,000 optimization steps. SD summarizes tail-loss variation.

Setting	Tail loss	Tail SD
Chebyshev order 2	0.45676	0.10328
Chebyshev order 4	0.45672	0.10328
Chebyshev order 8	0.45671	0.10329
CHI consistency only	0.40798	0.10347
Maximum delay 0	0.45733	0.10462
Maximum delay 12	0.45652	0.10328
Maximum delay 4	0.46215	0.10331
$\lambda_c = 0$	0.40796	0.10346
$\lambda_c = 0.05$	0.43396	0.10334
$\lambda_c = 0.2$	0.49941	0.10326
Learning rate 2×10^{-4}	0.45016	0.10320
Learning rate 5×10^{-5}	0.45961	0.10403
Modes 1	0.45004	0.10342
Modes 2	0.45289	0.10343
Modes 8	0.45269	0.10342

galleries are reproduced using their released implementations. Their quantitative reference scores are reported in the cited papers.

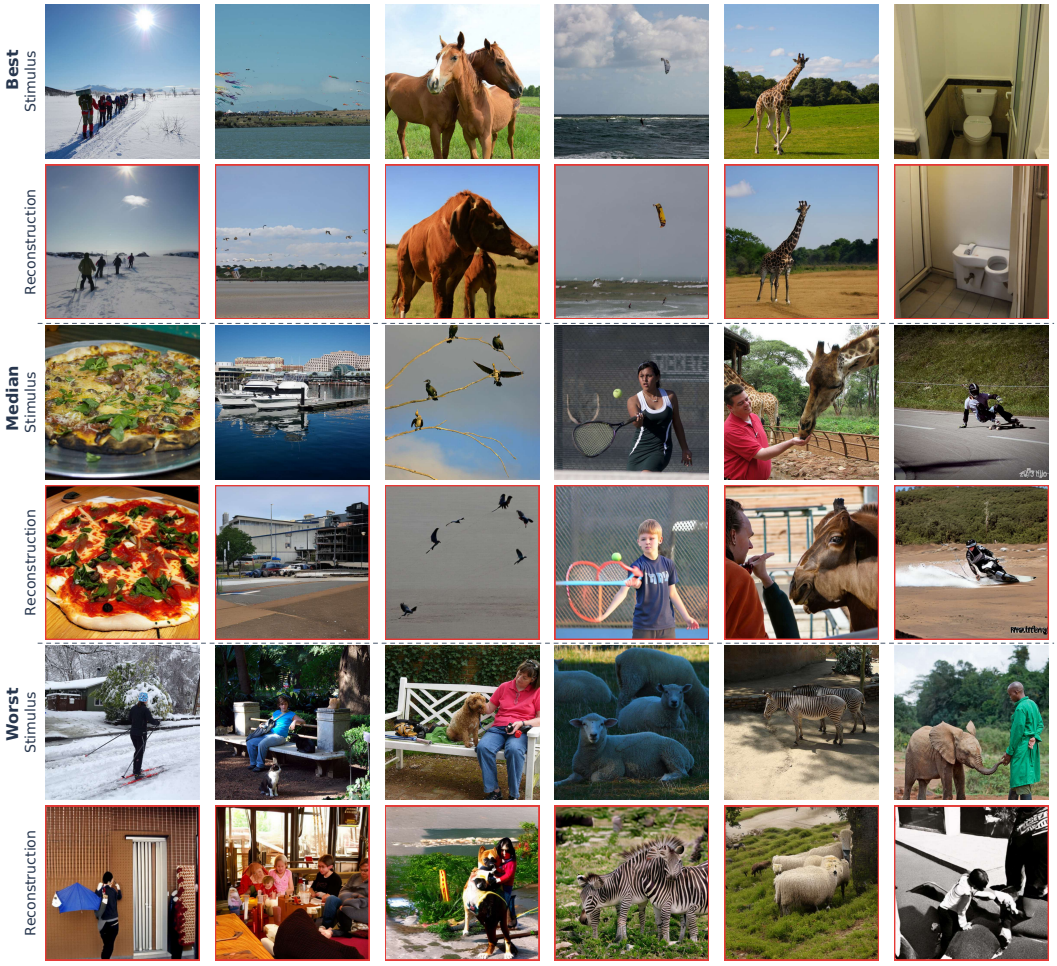

Figure 9: **NSD direct-output strata.** Six stimulus–reconstruction pairs are shown in each of the best, median, and worst tiers from top to bottom. Selection follows the stated reconstruction-metric percentile score and excludes the main-text comparison stimuli.

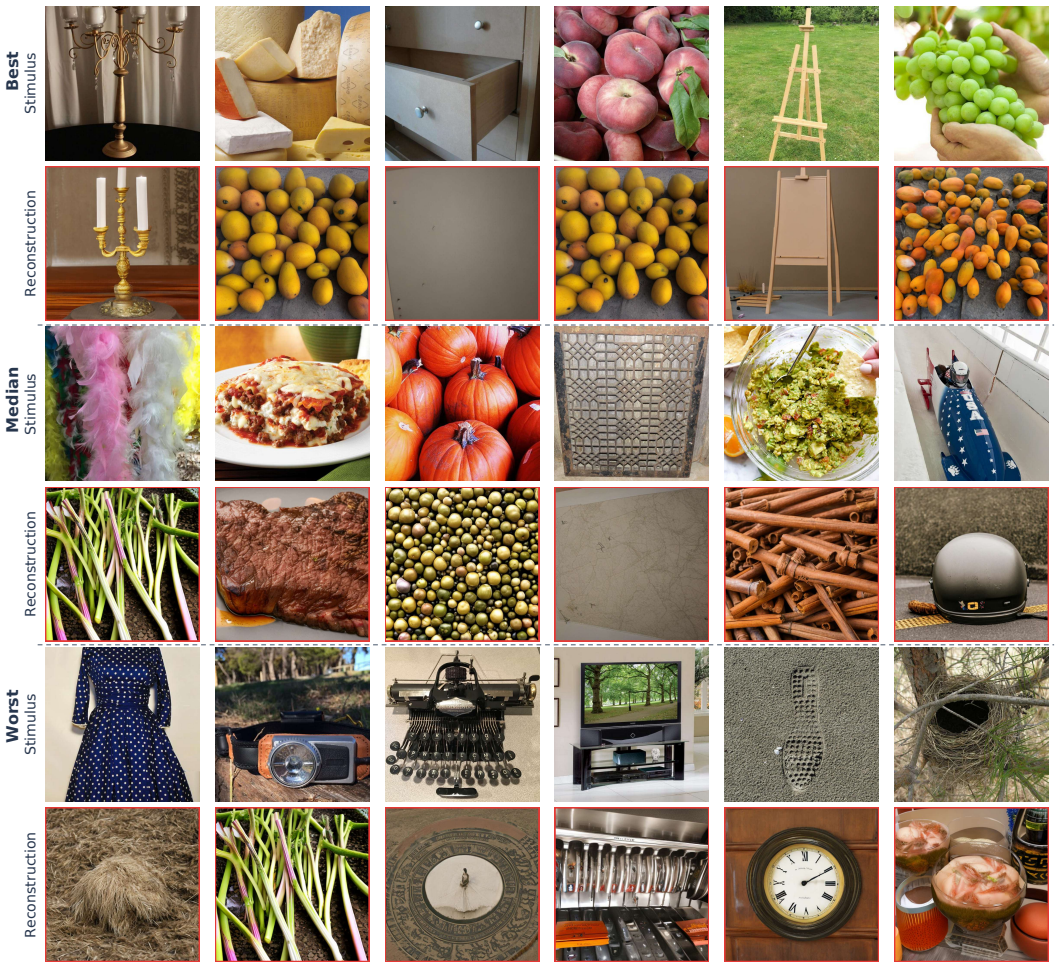

Figure 10: **THINGS-fMRI direct-output strata.** Six stimulus–reconstruction pairs are shown in each tier under the stated percentile score. The selected stimuli are disjoint from the main-text THINGS-fMRI comparison.

H TEMPORAL, CONDITIONING, AND REGIONAL ANALYSES

H.1 QUALITY REFERENCES AND TEMPORAL EVALUATION

Temporal interventions use $T(c) = \overline{Q}(c)$ from Equation 44, computed from each condition's eight reconstruction metrics on fixed test stimuli. Table 12 summarizes the unperturbed configurations using Q_{tab}, calculated from the rounded dataset means in Table 2 (Equation 43).

Table 12: Quality references Q_{tab} calculated from the eight rounded dataset means in Table 2, shown to four decimals. Rows summarize unperturbed configurations. C0 is the controlled no-consistency run.

Configuration	NSD Q_{tab} ↑	THINGS-fMRI Q_{tab} ↑
G0: w/o DCG-SSM	0.8695	0.5897
C0: No consistency	0.8757	0.6047
DCG-CHI (Ours)	**0.8813**	**0.6142**

H.2 DIRECT EXCHANGE OF SEMANTIC AND STRUCTURAL CONDITIONS

Image-level structural and semantic identification. Figure 5D evaluates retention of the original stimulus information using structural and semantic two-way identification rates. For an original stimulus I_i, a different stimulus I_j supplies the exchanged conditions. Intact uses both conditions from I_i; semantic swap replaces only its semantic condition with that of I_j; structure swap replaces only its structural condition; and both swapped replaces both. All four conditions are evaluated against the original stimulus I_i.

For each stimulus, one fixed distractor I_r is sampled randomly from the test stimuli excluding I_i and I_j. Structure compares the reconstruction's SSIM with the target and distractor. Semantics compares cosine similarities of their frozen CLIP image features. A trial scores 1 when target similarity is

higher, 0 when it is lower, and 0.5 for a tie. The SSIM implementation, image preprocessing, and CLIP encoder are the same as in the primary reconstruction evaluation.

All four conditions share the same exchange pairs, distractors, and generation seeds. Scores are first averaged over generation repeats for each stimulus, then over stimuli, and finally equally over subjects. The two datasets are evaluated separately. Both scores lie in $[0, 1]$ and are not normalized relative to Intact. Because the reference is the true stimulus rather than the reconstruction itself, Intact need not score 1. These image-level scores are computed and reported separately from the internal condition-tensor cosine consistencies below.

Internal condition-tensor consistency. For a recipient stimulus i and a donor j, direct exchange after the readouts replaces the semantic tensor, structural tensor, or both. With the same donor assignment across exchange conditions, define γ_{str} and γ_{sem} as the averages of $\cos(\mathbf{h}_j, \mathbf{h}_i)$ and $\cos(\mathbf{z}_j, \mathbf{z}_i)$, respectively, in the stimulus–repeat–subject order of Equation 48. Table 13 records the resulting algebraic identities for internal tensor consistency.

Table 13: Algebraic consistency identities for direct condition exchange after the readouts. Entries equal to one follow from unchanged, nonzero tensors. γ_{str} and γ_{sem} are mean donor–recipient cosines.

Condition	Structural consistency↑	Semantic consistency↑
Intact conditions	1	1
Semantic exchange	1	γ_{sem}
Structural exchange	γ_{str}	1
Both exchanged	γ_{str}	γ_{sem}

Exchanged-branch cosines lie in $[-1, 1]$ and depend on the donor–recipient pair. These identities apply immediately after exchange. Downstream generation depends jointly on both conditions. Image-level effects are assessed through reconstruction metrics, while C0–C3 compare CHI training objectives.

H.3 Regional sensitivity and effective delays

Under ROI masking, V1 retains 0.78 of PixCorr and 0.95 of CLIP, whereas masking higher visual regions retains 0.94 and 0.71, respectively. These results are consistent with differentiated model sensitivity to the specified regional inputs (Figure 5E).

The mean effective-delay profile across the 24 layer–mode profiles peaks at three seconds, with mass 0.1962; 0.5622 of its mass lies between two and four seconds. The ROI-resolved summary in Figure 11B also peaks at three seconds, at 0.193. The two summaries aggregate layer–mode weighting and regional influence, respectively. DCG-SSM shares mode weights across ROIs at each time step.

Cortical graph influence. For Figure 6, the influence associated with brain region i among the 360 HCP-MMP1.0 parcels and delay k is computed from the trained checkpoint as

$$I_i(k) = \frac{1}{24} \sum_{\ell=1}^{6} \sum_{m=1}^{4} q_{\ell,m}(k) e^{-\alpha_{\ell,m} k \Delta t} \sum_{j \neq i} \left| \left[e^{-\beta_{\ell,m} k \Delta t L_G} \right]_{ij} \right|, \qquad \Delta t = 1\,\mathrm{s}. \tag{54}$$

Here $L_G = I_R - D^{-1/2} W D^{-1/2}$ is the normalized Laplacian of the functional graph fixed during training, with $D_{ii} = \sum_j W_{ij}$. The checkpoint supplies the learned delay weights $q_{\ell,m}(k)$, normalized by $\sum_{k=1}^{8} q_{\ell,m}(k) = 1$, temporal decay rates $\alpha_{\ell,m}$, and graph diffusion coefficients $\beta_{\ell,m}$. The inner sum is the absolute off-diagonal row sum of the graph heat kernel, excluding the region's self-term. The factor $1/24$ gives equal weight to the six layers and four modes per layer. This summary uses the checkpoint kernel parameters without sample-dependent mode gates or channel mixing matrices.

The three columns accumulate the corresponding delay ranges:

$$I_i^{\mathrm{total}} = \sum_{k=1}^{8} I_i(k), \qquad I_i^{\mathrm{early}} = \sum_{k=1}^{4} I_i(k), \qquad I_i^{\mathrm{late}} = \sum_{k=5}^{8} I_i(k). \tag{55}$$

Thus $I_i^{\mathrm{total}} = I_i^{\mathrm{early}} + I_i^{\mathrm{late}}$ for every region. Colors encode these raw cumulative values, with no region-wise or panel-wise renormalization. All panels use a common 0–0.35 color scale.

Together these analyses characterize model sensitivity to regional inputs and learned temporal weighting. Optimization diagnostics appear in Appendix F.

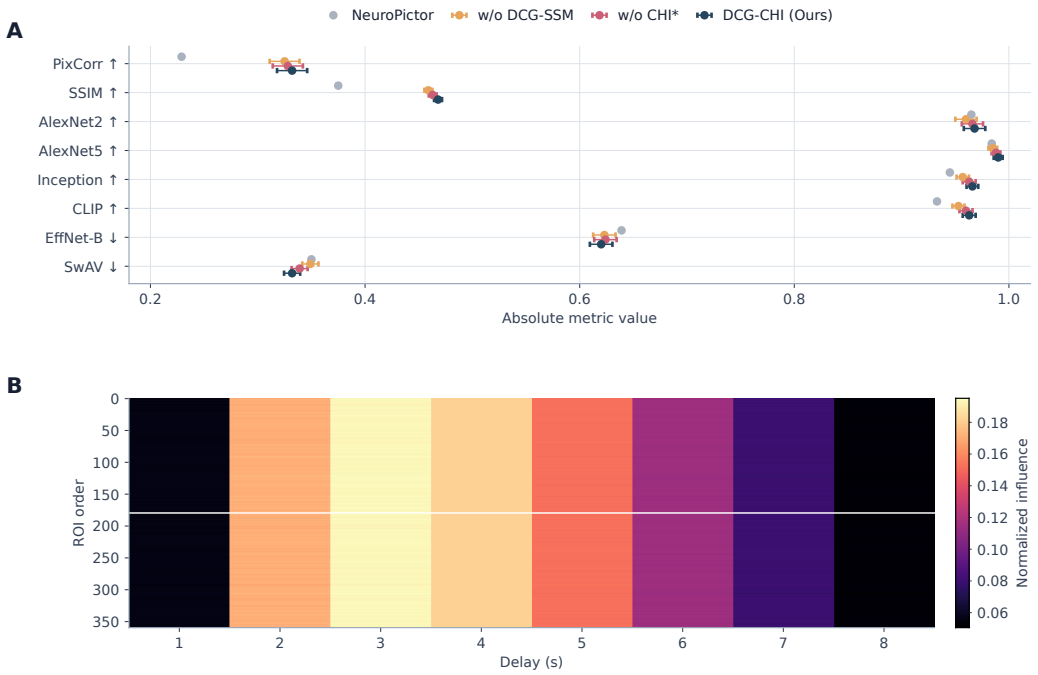

Figure 11: **Reconstruction and regional analyses.** (A) NSD absolute metrics for the reference and internal configurations. (B) ROI-resolved delay profiles with hemispheres separated. The layer–mode summary is shown in Figure 5F. Computational resources are shown in Figure 7F and detailed in Appendix N.

I Subject-transfer and adaptation evaluation

I.1 Evaluation protocols

We evaluate within-subject reconstruction, single-source transfer, joint-subject training, unseen-target zero-shot reconstruction, and target adaptation using the common quality definition $L_{\text{protocol}}(p) = \overline{Q}(p)$ (Equation 45). Comparisons for a target subject use the same test stimuli and metric configuration, and the two datasets are reported separately.

A four-subject NSD endpoint or three-subject THINGS-fMRI endpoint represents joint-training scale. Strict leave-one-subject-out evaluation permits at most three NSD or two THINGS-fMRI sources. Target adaptation adds the specified number of calibration examples to the zero-shot setting. Protocol scores use the subject-first aggregation in Equation 45.

Table 14 and Figure 7A report dataset-level protocol means. The unseen-subject zero-shot condition is the same experiment as zero calibration in the adaptation analysis. The means are 0.7315 on NSD and 0.4913 on THINGS-fMRI. Figure 7C shows the multi-source minus single-source differences. Panel B gives full-model subjectwise references obtained by applying Equation 43 to each rounded metric row of Table 5.

Table 14: Dataset-level protocol means on $L_{\text{protocol}} = \overline{Q}$ (↑). Zero-shot uses the same zero-calibration condition as Table 15. The subject and repeat aggregation is defined in Equation 45.

Dataset	Single source	Multiple sources	Zero-shot
NSD	0.7932	0.8460	0.7315
THINGS-fMRI	0.5405	0.5835	0.4913

I.2 Target-subject adaptation

Table 15 reports adaptation on the common quality scale $L_{\text{protocol}} = \overline{Q}$, also shown in Figure 7D. On NSD, the score increases from 0.7315 at zero-shot to 0.8284 at ten-shot, a gain of 0.0969. On THINGS-fMRI, it increases from 0.4913 to 0.5712, a gain of 0.0799.

The quality-index changes summarize the eight reconstruction metrics. Equation 46 defines the corresponding metric-specific differences.

Table 15: Target-subject adaptation results on $L_{\mathrm{protocol}} = \overline{Q}$ (↑). Calibration counts denote target examples. The last column gives Q_{tab} calculated from the full-model main-table metrics as a reference.

Dataset	Zero-shot	1-shot	5-shot	10-shot	Reference Q_{tab} ↑
NSD	0.7315	0.7579	0.7932	0.8284	0.8813
THINGS-fMRI	0.4913	0.5159	0.5466	0.5712	0.6142

J Extended reconstruction and representation analyses

J.1 Retrieval-pool scaling

Table 16 and Figure 12A report bidirectional top-1 retrieval for candidate pools of 50, 300, and 982 entries. At 982 candidates, fMRI-to-image accuracy is 80% on NSD and 71% on THINGS-fMRI. The reverse direction reaches 77% and 68%, respectively. Top-1 retrieval requires the matching candidate to rank first within the candidate pool. Primary feature identification averages pairwise comparisons.

Table 16: Bidirectional top-1 retrieval accuracy (↑, percent) by candidate-pool size.

Dataset	Direction	50	300	982
NSD	fMRI → image	94	87	80
NSD	image → fMRI	92	84	77
THINGS-fMRI	fMRI → image	87	79	71
THINGS-fMRI	image → fMRI	84	76	68

J.2 Training-image similarity and category variation

Figure 12B groups reconstructions by similarity to their nearest training image. Point area indicates bin count. This analysis describes how reconstruction scores vary with training-image similarity. Panel C reports normalized CLIP, SSIM, and PixCorr across seven categories: animals, food, vehicles, tools, clothing, plants, and indoor scenes. Category profiles supplement the dataset-level metrics by showing variation across content types.

J.3 Semantic evaluation

Figure 12D reports CLIP-S, CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016). The measured CLIP-S, CIDEr, and SPICE scores for DCG-CHI are 0.77, 0.66, and 0.62, respectively. The BrainFLORA reference values 44.21, 4.40, and 15.55 are divided by 100 for display, giving 0.4421, 0.0440, and 0.1555, respectively, marked by §. The caption-generation pipeline and reference set follow each method's original evaluation protocol.

K Hemodynamic perturbation evaluation

K.1 Test-input construction

Test perturbations apply a model-independent HRF transfer filter to each full continuous ROI BOLD run. This approximates regularized frequency-domain deconvolution with a fixed causal double-Gamma reference HRF followed by reconvolution with a temporally shifted HRF. The shift δ is in seconds, with positive values delaying the response. Kernels are evaluated over 0–32 s on the BOLD filtering grid and normalized by their discrete absolute sum. The filtering step is 1 s for NSD and 0.5 s for THINGS-fMRI. Fractional shifts are evaluated directly in the continuous HRF function.

The reference kernel is a double Gamma with shapes 6 and 16, scales of 1 s, and negative-component weight $1/6$. With H_0 and H_δ denoting the reference and perturbed kernel spectra, the regularized transfer spectrum is computed as

$$G_\delta(\omega) = \frac{H_\delta(\omega)\overline{H_0(\omega)}}{|H_0(\omega)|^2 + \lambda}, \qquad \lambda = 10^{-3} \max_\omega |H_0(\omega)|^2. \tag{56}$$

The implementation additionally sets the DC gain to one. Each run is reflection-padded by 32 s per side (32 NSD or 64 THINGS-fMRI samples). The FFT length is the smallest power of two at least $\max\{1024, N_{\mathrm{run}} + 2P + K_h - 1\}$, where P is the per-side padding and K_h the sampled kernel length. After filtering, the full run is cropped to its original length and stimulus-aligned windows are extracted on the 1-second model grid. Dataset-specific normalization and masks are retained. All models receive the same perturbed inputs. Clean inputs bypass filtering. Thus, the clean reference is the original input rather than the output of a zero-shift regularized transfer filter.

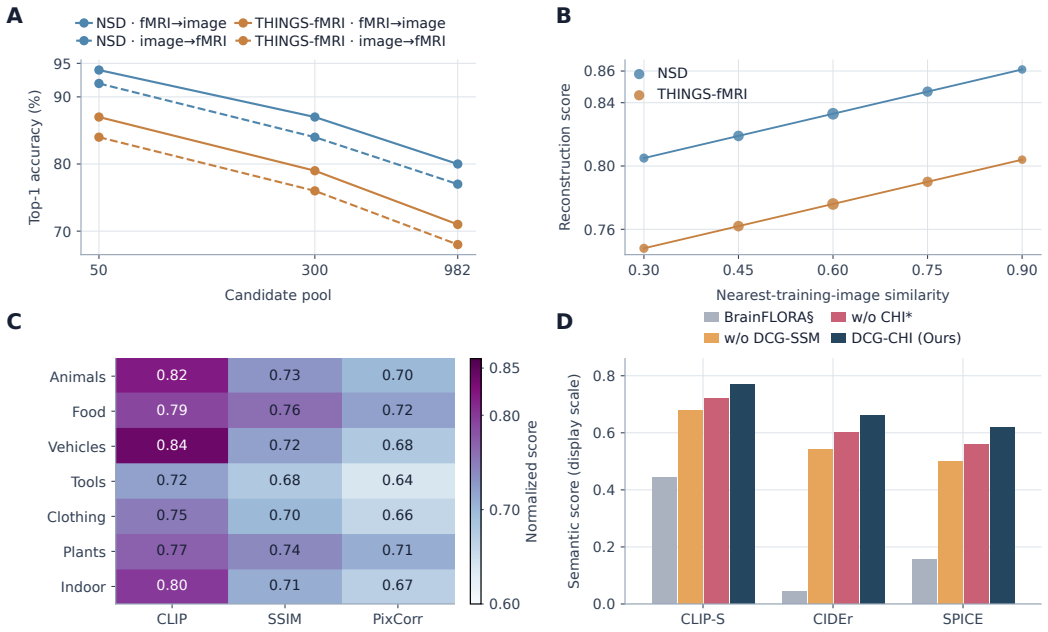

Figure 12: **Retrieval and semantic analyses.** (A) Bidirectional top-1 retrieval at three candidate-pool sizes. (B) Reconstruction score versus nearest-training-image similarity, with point area proportional to bin count. (C) Category-level normalized CLIP, SSIM, and PixCorr. (D) Semantic-evaluation scores for the internal variants and BrainFLORA reference. §: BrainFLORA reference scores divided by 100. Internal variants show measured scores.

K.2 Perturbation settings and reporting

We evaluate reconstruction under temporal HRF shifts with absolute offsets of 0, 1, and 2 seconds. Positive shifts delay the HRF. Negative shifts advance it. For $d = |\delta|$, scores average the $+d$ and $-d$ conditions equally. Clean references use the same model, test set, and generation settings.

Tables 17 and 18 report the reconstruction scores for the evaluated conditions. The zero-shift scores provide the corresponding clean references. G4 and C3 denote the same full model in the encoder and consistency comparisons, respectively, so their shared clean and perturbation results are listed consistently. These experiments evaluate reconstruction robustness to the specified continuous-run HRF transfer perturbations.

K.3 Consistency variants under temporal shifts

Table 17 compares C0–C3. At a two-second shift, full CHI reaches SSIM/CLIP of 0.432/0.889 on NSD and 0.338/0.697 on THINGS-fMRI, compared with 0.404/0.835 and 0.315/0.653 for generic consistency. Under this shift, the HRF-view objective preserves higher structural and semantic scores than the generic-view control.

Table 17: Consistency variants under temporal HRF shifts. Each nonzero absolute shift averages its positive and negative conditions. Columns report SSIM/CLIP identification pairs. Zero shift is the clean reference condition. ↑ indicates higher is better.

Model	SSIM↑ 0 s	CLIP↑ 0 s	SSIM↑ 1 s	CLIP↑ 1 s	SSIM↑ 2 s	CLIP↑ 2 s
NSD						
C0	0.463	0.960	0.440	0.913	0.379	0.786
C1	0.465	0.961	0.449	0.928	0.404	0.835
C2	0.466	0.962	0.453	0.935	0.417	0.860
C3	0.468	0.963	0.459	0.944	0.432	0.889
THINGS-fMRI						
C0	0.360	0.747	0.342	0.711	0.295	0.612
C1	0.362	0.751	0.350	0.725	0.315	0.653
C2	0.363	0.752	0.353	0.731	0.325	0.672
C3	0.366	0.755	0.359	0.740	0.338	0.697

K.4 Delay coupling under temporal shifts

Table 18 compares the parallel encoder G3 with the coupled encoder G4 while retaining CHI. At two seconds, G4 reaches SSIM of 0.432 and 0.338, compared with 0.412 and 0.321 for G3 on NSD and THINGS-fMRI, respectively. This comparison complements the clean-input encoder controls by examining the two complete architectures under temporal perturbation.

Table 18: SSIM under temporal shifts with CHI held fixed. G3 uses parallel spatial–temporal branches and G4 uses the delay-coupled transition. ↑ indicates higher is better.

Dataset	Encoder	SSIM↑ 0 s	SSIM↑ 1 s	SSIM↑ 2 s
NSD	G3	0.465	0.451	0.412
NSD	G4	0.468	0.459	0.432
THINGS-fMRI	G3	0.362	0.351	0.321
THINGS-fMRI	G4	0.366	0.359	0.338

L Input robustness, sampling, and failure analysis

L.1 Input noise and missing-input perturbations

The input-corruption experiments add Gaussian noise to standardized inputs and apply random voxel masking. At noise standard deviation 0.4, DCG-CHI achieves a normalized score of 0.8188, compared with 0.760 for *w/o CHI** and 0.722 for *w/o DCG-SSM*. At a 60% masking ratio, the corresponding scores are 0.7572, 0.696, and 0.674 (Figure 13A/B).

Table 19: Input-corruption results at the largest evaluated perturbation settings. Scores (↑) use the normalized scale of the corruption experiment.

Configuration	Noise SD 0.4	Masking ratio 60%
w/o DCG-SSM	0.722	0.674
w/o CHI*	0.760	0.696
DCG-CHI (Ours)	**0.8188**	**0.7572**

L.2 Sampling sensitivity

Figure 13C compares 10, 20, 30, and 50 denoising steps at classifier-free guidance scales of 3, 5, and 7.5. The highest quality score is 0.892 at guidance scale 5 and 50 steps. The figure plots reconstruction quality against the number of denoising steps for each guidance scale. The 50-step configuration takes 5.50 seconds per image. Resource accounting is detailed in Appendix N. The trained model is fixed throughout this sampling sweep.

L.3 Failure taxonomy

Figure 13D groups DCG-CHI outputs on both datasets as Acceptable, Category, Count, Attribute, Layout, Omission, or Background. The categories distinguish incorrect object identity from errors in multiplicity, appearance, spatial arrangement, missing content, and scene context. The plotted distributions summarize the evaluated reconstruction outputs. The galleries in Appendix G illustrate representative successful and unsuccessful cases.

M Subject reliability, data efficiency, and seed consistency

M.1 Subjectwise reliability

For reliability and seed consistency, let $e(I)$ be the L2-normalized image embedding from frozen CLIP ViT-L/14 and define

$$c(I, J) = \frac{1 + e(I)^\top e(J)}{2}. \tag{57}$$

For stimulus I_i, reconstruct its $J_i \geq 2$ repeated fMRI observations using the same generation seed. The stimulus-anchored reliability is

$$R_i = \left[\frac{1}{J_i} \sum_{j=1}^{J_i} c(\widehat{I}_i^{(j)}, I_i) \right] \left[\frac{2}{J_i(J_i - 1)} \sum_{j<k} c(\widehat{I}_i^{(j)}, \widehat{I}_i^{(k)}) \right]. \tag{58}$$

The factors measure stimulus agreement and stability across input observations, respectively. We average stimulus scores within each subject and then weight subjects equally, separately for each dataset. This mean is $L_{\text{reliability}}$ in Table 6.

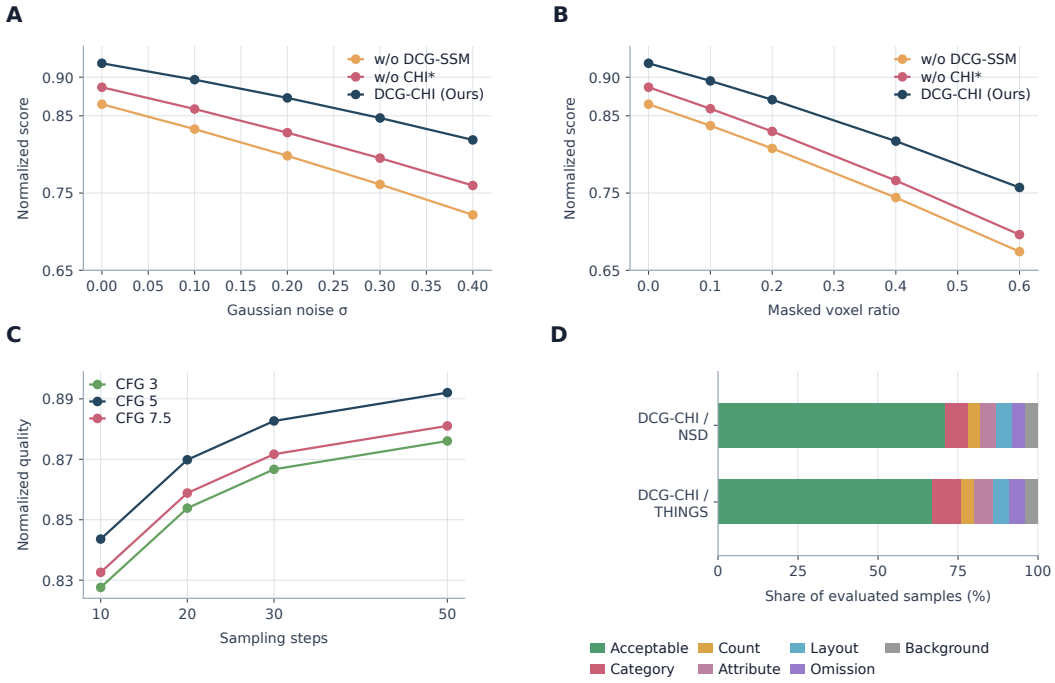

Figure 13: **Input robustness, sampling, and failure analysis.** (A) Gaussian noise on standardized inputs. (B) Random voxel masking. (C) Reconstruction quality versus the number of denoising steps at different guidance scales. (D) Output categories for DCG-CHI reconstructions on both datasets.

The stimulus-anchored reliability means are 0.965/0.950 for DCG-CHI on NSD/THINGS-fMRI, compared with 0.948/0.933 for *w/o CHI** and 0.933/0.918 for *w/o DCG-SSM* (Figure 14A). Table 20 reports these means together.

Table 20: Mean stimulus-anchored reliability (↑) on the two datasets.

Configuration	NSD	THINGS-fMRI
w/o DCG-SSM	0.933	0.918
w/o CHI*	0.948	0.933
DCG-CHI (Ours)	**0.965**	**0.950**

M.2 Training-data budgets

Figure 14B compares DCG-CHI at training budgets of 1, 2, 4, and 10 hours with its full-data setting. For each subject s, the normalized SSIM–CLIP score is

$$S_{\mathrm{eff},s}(h) = \frac{1}{2}\left[\frac{\mathrm{SSIM}_s(h)}{\mathrm{SSIM}_s(\mathrm{Full})} + \frac{\mathrm{CLIP}_{\mathrm{ID},s}(h)}{\mathrm{CLIP}_{\mathrm{ID},s}(\mathrm{Full})}\right]. \tag{59}$$

CLIP here is two-way identification. Each denominator is the full-data DCG-CHI result for the same subject and test set. The two ratios receive equal weight, followed by an equal-weight subject average. The Full endpoint is one. h measures available training data in hours. This score measures retained SSIM–CLIP performance relative to the full-data setting.

M.3 Generation-seed consistency

We fix the checkpoint, fMRI observation and sampling settings, varying only eight generation seeds (42–49). Using the image similarity in Equation 57, define

$$C_i = \frac{2}{8 \times 7} \sum_{1 \le j < k \le 8} c(\hat{I}_i^{(j)}, \hat{I}_i^{(k)}). \tag{60}$$

This is the average of 28 reconstruction pairs for a fixed stimulus. Scores are averaged over stimuli within subjects, then equally over subjects. For a subject with N test stimuli, seed j has score

$$S_j = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{7} \sum_{k \ne j} c(\hat{I}_i^{(j)}, \hat{I}_i^{(k)}), \qquad j = 1, \ldots, 8. \tag{61}$$

Panel C shows eight scores per box, each averaged equally over subjects. Their mean equals the mean of C_i: 0.902 on NSD and 0.857 on THINGS-fMRI for the full model. Boxes show quartiles, center lines medians, and whiskers the minimum–maximum range. Each plotted observation is a seed-level summary with the checkpoint and fMRI observation fixed.

M.4 Computational-resource visualization

Panel D shows parameter counts and inference times for the same configurations as Figure 7F and Table 21. The DCG-CHI point is 5.50 seconds per image and 1.35B parameters. Circle area is proportional to parameter count. Resource configurations are described in Appendix N.

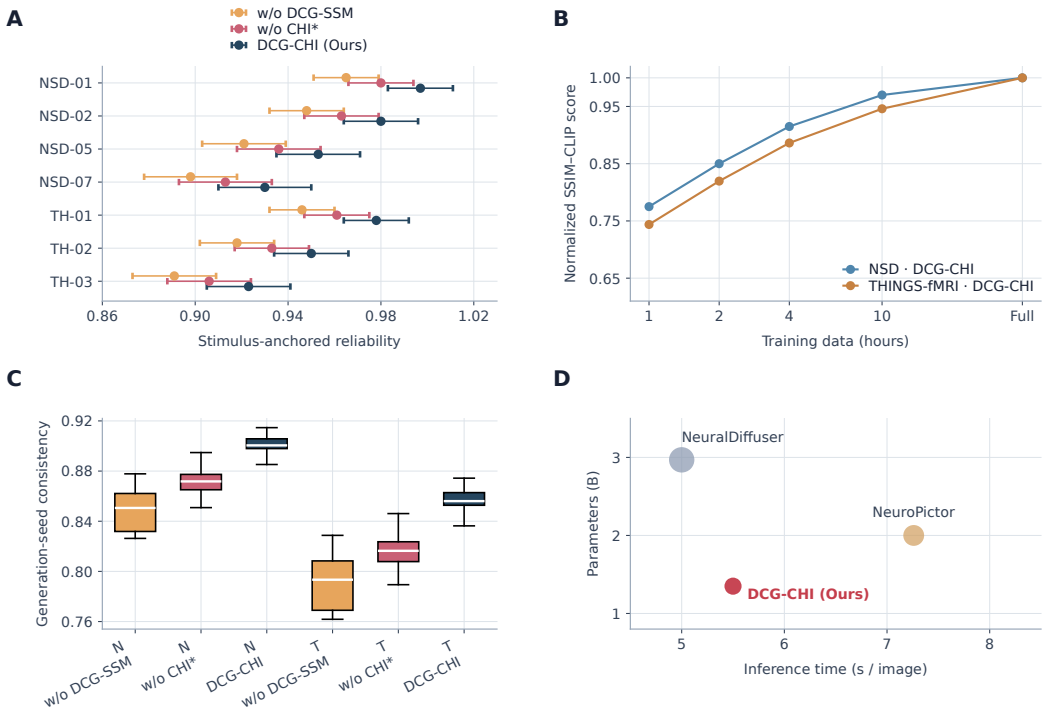

Figure 14: **Reliability, data efficiency, and generation variability.** (A) Subjectwise stimulus-anchored reliability. (B) DCG-CHI SSIM–CLIP performance relative to its full-data setting. (C) Eight seed-level scores per box (Equation 61), with minimum–maximum whiskers. (D) Parameter counts versus inference time for the configurations in Figure 7F. Circle area is proportional to parameter count.

N Computational resources

N.1 Full-model resource use

The DCG-CHI configuration contains 1.35B total parameters, of which 0.20B are trainable, or 14.8%. Its per-image runtime is 5.50 seconds with 50 generation steps. The configuration uses 188 GFLOPs, 15.60 GB VRAM, and 54 GPU-hours of training. CHI adds shared-encoder passes during training but no synthetic-view branch at inference.

N.2 Resource comparison and counting conventions

Table 21 compares resources for the listed experimental configurations. Table 22 defines the parameter, training-cost, and inference-cost quantities.

MindBridge and MindEye2 runtimes are reported in MindHier's comparison table. NeuralDiffuser reports timing on an RTX 3090. NeuroPictor reports 13.6 GB VRAM with fp32 and 50 DDIM steps on an RTX A6000. BrainFLORA reports 7.36M parameters for its compact projection module (Li et al., 2025a). This is a module count rather than a total-model count. Each resource entry uses the hardware, precision, sampling configuration, and parameter-count convention specified by its source.

Table 21: Parameter counts and inference times from the cited papers and our experiments. A dash denotes an unreported value. ↓ indicates lower is better.

Method	Total params.↓ (B)	Time↓ (s/image)	Source
MindBridge	—	15.98	Zhang et al. (2026)
MindEye2	—	12.14	Zhang et al. (2026)
NeuralDiffuser	2.969	5.00	Li et al. (2025b)
MindHier	—	2.64	Zhang et al. (2026)
NeuroPictor	2.00	7.26	Huo et al. (2024)
DCG-CHI	1.35	5.50	This work

Table 22: Resource quantities and counting scopes.

Quantity	Reporting scope
Total parameters	Parameters in the reported model configuration, including frozen components where counted by that configuration.
Trainable parameters	Parameters updated during training within the specified model components.
Inference time	Seconds per generated image under the corresponding sampling, precision, and hardware configuration.
GFLOPs and VRAM	Computational and memory figures reported for the corresponding pipeline configuration.
Training GPU-hours	Training resource use, reported separately from per-image inference latency.
Generation steps	Sampling/generation iterations of the corresponding pipeline.

O Extended related work

Generative visual brain decoding. Latent diffusion reconstruction and Brain-Diffuser connect fMRI features to pretrained image generators (Takagi & Nishimoto, 2023; Ozcelik & VanRullen, 2023). MinD-Vis and MindEye improve brain representations through masked modeling and contrastive alignment (Chen et al., 2023; Scotti et al., 2023). DREAM, NeuralDiffuser, NeuroPictor, and MindHier explore complementary visual attributes, diffusion guidance, multilevel modulation, and hierarchical generation (Xia et al., 2024; Li et al., 2025b; Huo et al., 2024; Zhang et al., 2026). DCG-CHI combines delay-coupled BOLD encoding with consistency across synthetic hemodynamic observations.

Cross-subject alignment and generalization. MindBridge, MindEye2, and subsequent multi-subject models study alignment, data efficiency, and generalization (Wang et al., 2024; Scotti et al., 2024; Kong et al., 2025). SIM uses cortical and multimodal alignment (Dahan et al., 2025). Neural code conversion and intertwined autoencoders address individual differences (Wang et al., 2025; Xu et al., 2025). CHI complements subject alignment through consistency across hemodynamic observations.

Temporal information in brain decoding. MEG decoding emphasizes temporally resolved representations (Benchetrit et al., 2024). Mind-Animator and BrainNRDS model semantic, structural, and motion information for video reconstruction (Lu et al., 2025; Yeung et al., 2025). In DCG-CHI, time indexes the BOLD response to a static image.

Semantic alignment and synthetic brain responses. BrainFLORA, caption-embedding studies, and VINDEX examine neural semantic alignment and visual feature spaces (Li et al., 2025a; Doerig et al., 2025; Xia & Oztireli, 2025). Uncertainty-aware blur addresses neural–visual mismatch (Wu et al., 2025). MindSimulator synthesizes concept-conditioned fMRI (Zhang et al., 2025), whereas CHI synthesizes HRF-transformed observations of a shared encoded trajectory.

EEG-to-image reconstruction. DreamDiffusion and the Adaptive Thinking Mapper learn EEG representations for conditioned image generation (Bai et al., 2024; Li et al., 2024). D^2-FOSA combines frequency-oriented encoding and dual-diffusion alignment (Yu et al., 2026).

Graph and state-space models. Graph convolution and diffusion recurrence model relational structure (Kipf & Welling, 2017; Li et al., 2018). S4, S5, Mamba, and Mamba-2 develop structured and selective state-space computation (Gu et al., 2022; Smith et al., 2023; Gu & Dao, 2024; Dao & Gu, 2024). DCG-SSM couples finite distributed delays with graph diffusion, using Chebyshev polynomial filtering for sparse evaluation (Defferrard et al., 2016). Appendix C analyzes equivariance and contraction of this recurrence.