CortexFlow: Topology-Aware Spectral Decoding with Confidence-Adaptive Diffusion for fMRI-to-Image Reconstruction

Anonymous authors
Paper under double-blind review

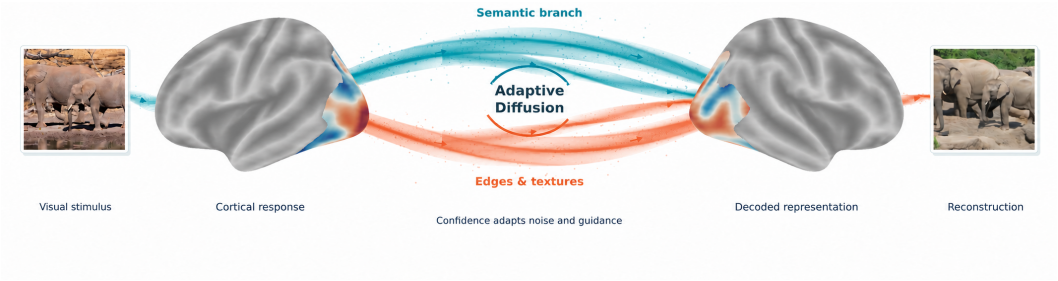


Figure 1: CortexFlow decodes a visual stimulus through complementary topology-aware structural and semantic pathways. Response-specific confidence adapts the diffusion noise level and semantic guidance to balance preservation of edges and textures with semantic reconstruction.

Abstract

Most fMRI-to-image reconstruction methods, despite advances in generative modeling, begin by representing spatially organized cortical activity as a flat vector. This overlooks the way visual cortex actually works. Cortical activity is organized across functionally specialized regions and spread over a folded cortical surface, where neighboring locations respond continuously. Beyond that, the relative importance of low-level structure and high-level semantics varies across scenes. Natural landscapes may depend more on spatial layout and texture, whereas scenes depicting events or activities may require a stronger emphasis on object identities and their interactions. A fixed diffusion schedule and guidance strength ignore this variability. We propose CortexFlow, which builds a within-hemisphere graph over the fsaverage cortical surface, extracts its Laplacian eigenmodes, and integrates branch-specific conditional flows in these spectral coordinates to incorporate cortical topology into both the VGG structural branch and the CLIP semantic branch. Two confidence gates estimate the feature-level agreement of the respective predictions. Structural confidence adapts the diffusion start, while semantic confidence controls the strength of semantic guidance. CortexFlow achieves state-of-the-art reconstruction performance on NSD across a broad range of low- and high-level metrics, with particularly strong structure-sensitive performance in SSIM. Cortical projections of topology-induced updates further provide a model-level interpretation of the decoded representations.

1 Introduction

Reconstructing visual content from brain activity provides a unique window into how the brain represents the visual world (Kay et al., 2008; Allen et al., 2022) and supports potential applications in brain–computer interfaces and clinical assessment of visual awareness. In a typical fMRI-to-image reconstruction setting, participants view images while their brain activity is recorded using functional magnetic resonance imaging (fMRI), producing paired images and neural responses. At test time, the model receives only an fMRI response and reconstructs the image that elicited it. A successful reconstruction should preserve the specific stimulus viewed by the participant, since visual plausibility alone does not establish fidelity to the original image.

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.

Modern fMRI-to-image reconstruction systems increasingly decode neural responses into deep visual representations that condition generative models, while also exploiting information shared across subjects(Shen et al., 2019; Takagi & Nishimoto, 2023; Ozcelik & VanRullen, 2023; Scotti et al., 2024; Wang et al., 2024a). Most existing decoders make limited use of cortical topology, while their generative stages typically apply fixed structural and semantic conditioning across responses.

The visual cortex contains functionally specialized regions arranged on a folded cortical surface (Kandel et al., 2013; Grill-Spector & Weiner, 2014; Wandell et al., 2007; Dumoulin & Wandell, 2008). Most reconstruction methods flatten cortical activity into one-dimensional vectors, overlooking spatial relationships on the cortical surface. Brain-IT models interactions among functionally similar voxel groups (Beliy et al., 2026), whereas *From Flat to Round* preserves cortical geometry in its tokenization (Yu et al., 2025). To combine functional grouping with local cortical surface organization, we adopt branch-specific conditional flow matching over a shared cortical graph. Laplacian heat drift provides a spatial smoothness prior throughout the flow, while control fields conditioned on functional-region tokens learn task-dependent adjustments for structural and semantic decoding.

A second challenge is that structural and semantic predictions can differ in reliability across fMRI responses: natural landscapes often rely more on low-level properties such as spatial layout, colors, and textures, whereas complex scenes with multiple objects and interactions require richer semantic representations. In our structure-initialized diffusion, starting noise determines how much decoded layout may change, whereas semantic guidance determines how strongly generation follows predicted content. We therefore learn separate feature-agreement scores that adapt these controls without assigning scenes to predefined categories.

We introduce **CortexFlow**, combining **topology-conditioned dual-branch decoding** with **confidence-adaptive diffusion**. A shared cortical graph provides spectral coordinates in which branch-specific flows refine functional-region representations for structural VGG and semantic CLIP prediction. Separate confidence gates estimate feature agreement and adapt the diffusion start and semantic guidance, respectively. Thus, cortical organization informs both decoded representations, while their estimated feature agreement controls how they enter generation.

Our contributions are:

- **Topology-conditioned visual decoding.** We incorporate cortical surface neighborhoods into both VGG and CLIP decoding through conditional spectral flows, sharing anatomical coordinates while learning branch-specific updates to functional region tokens.
- **Confidence-adaptive diffusion.** Two feature-agreement predictors control initialization noise and semantic guidance separately, allowing the same response to carry different confidence in layout and content.
- **State-of-the-art reconstruction and analysis.** CortexFlow achieves state-of-the-art performance among published NSD results across most reported metrics, while matched four-subject comparisons show consistent improvements in both pixel-level fidelity and semantic identification.

2 Related Work

2.1 Neural Decoding

Visual decoding has progressed from statistical reconstruction (Kay et al., 2008; Naselaris et al., 2009; Nishimoto et al., 2011) to deep features (Guclu & van Gerven, 2015; Shen et al., 2019) and diffusion priors (Chen et al., 2023; Takagi & Nishimoto, 2023; Ozcelik & VanRullen, 2023). Shared-subject models improve data efficiency (Scotti et al., 2024; Wang et al., 2024a; Gong et al., 2025), with extensions to unified and multimodal decoding (Xia et al., 2024b; Wang et al., 2024b; Shen et al., 2024). Brain-IT instead relates functionally similar voxel groups (Beliy et al., 2026), while *From Flat to Round* retains cortical geometry through surface tokens (Yu et al., 2025). CortexFlow combines these complementary structures by transporting both structural and semantic region tokens on a shared surface graph.

2.2 Adaptive and Topological Generation

Diffusion reconstruction usually fixes initialization noise and semantic guidance (Takagi & Nishimoto, 2023; Scotti et al., 2023), although starting noise controls preservation (Meng et al., 2021) and sample-adaptive generation can respond to input quality (Fabian et al., 2024). Flow Matching learns continuous transport (Lipman et al., 2023), and Topological Flow Matching adds Laplacian-dependent dynamics (Wyrwal et al., 2026). CortexFlow couples these ideas by using a cortical graph to refine dual visual decoders and separately adapting their influence on generation.

3 Method

3.1 Reconstruction pipeline

CortexFlow reconstructs a viewed scene through structural and semantic pathways (Figure 2). The structural branch predicts localized VGG features for Deep Image Prior (DIP) inversion, while the semantic branch predicts adapted CLIP image tokens for diffusion conditioning. Both branches refine their functional-region representations with a Topology-Conditioned Cortical Flow (TCCF; Section 3.2).

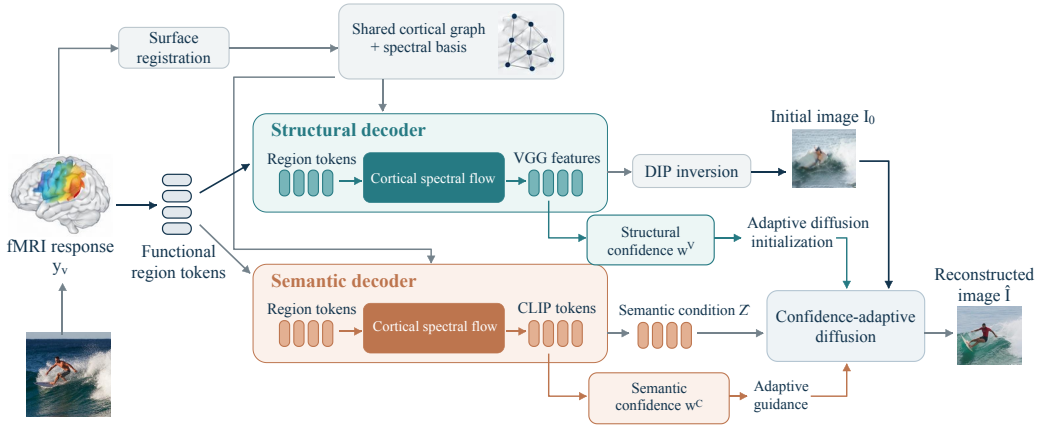


Figure 2: CortexFlow architecture. Both decoders receive their own cortical-flow updates. Structural confidence controls the perturbation of the DIP image; semantic confidence controls the guidance applied during denoising. The graph is shared, while learned topology modules and gates are branch-specific.

Because the two predictions can differ in reliability, branch-specific confidence scores separately control structural preservation and semantic guidance during diffusion (Sections 3.4–3.5). Training uses the target image for feature and confidence supervision; inference requires only the fMRI response and its subject-specific cortical registration.

3.2 Topology-conditioned dual-branch decoding

Functional grouping collects voxels with similar response patterns, even when they occupy distant cortical locations, whereas surface adjacency captures local organization on the cortical sheet. We combine these two views by transporting the registered cortical signal in a shared spectral coordinate system and using the transported signal to refine functional region tokens. The structural and semantic branches share the cortical graph and spectral basis, but learn independent transport dynamics under their respective visual objectives.

Regional Encoding. We begin with the functional-region representation used by two visual decoders. The branches share fixed functional-group assignments but maintain separate learned voxel embeddings, attention blocks, and visual queries. Each tokenizer weights learned voxel embeddings by their measured activations and aggregates them through attention into $C = 128$ region tokens of dimension $d = 512$. A region therefore represents a functional voxel group rather than a contiguous anatomical parcel. The structural decoder predicts five scales of localized VGG features $\hat{V}$, while

the semantic decoder predicts adapted CLIP tokens $\hat{Z}$. We index the branches by $b \in \{V, C\}$, with $F_V = \{V_\ell\}_{\ell=1}^{5}$ and $F_C = \{Z\}$.

Surface registration and graph. To provide a common topological coordinate system across subjects, we register voxel responses to the fsaverage cortical surface. For subject s, let $y_v \in \mathbb{R}^{N_v}$ denote the voxel response and $M_s \in \mathbb{R}^{N \times N_v}$ its fixed registration. We compute $y = M_s y_v$, averaging entries assigned to the same surface node. Because not every surface node is covered by an observed voxel, we additionally define a binary observation mask $m \in \{0,1\}^N$, where $m_i = 1$ indicates that node i receives at least one observed voxel; unobserved nodes are zero-filled. We restrict the common cortical surface to the union of the NSD `nsdgeneral` and retrosplenial cortex (RSC) masks, resulting in $N = 39{,}548$ surface nodes shared across subjects.

We construct a fixed graph over this surface using sampled-mesh edges together with 12 spherical nearest neighbors within each hemisphere. For an ordinary edge (i, j),

$$W_{ij} = \exp\left(-\frac{d_{ij}^2}{2\ell_i \ell_j}\right), \qquad D_{ii} = \sum_j W_{ij}, \qquad L_{\mathrm{sym}} = I - D^{-1/2} W D^{-1/2}, \tag{1}$$

where d_{ij} is the spherical chord distance and ℓ_i is the median distance from node i to its 12 spherical nearest neighbors. The local scales ℓ_i and ℓ_j account for variation in surface sampling density, so that edge strength reflects proximity relative to the local neighborhood rather than absolute template distance alone. We use spherical neighborhoods to define a shared template-space notion of cortical proximity across subjects; for local neighbors, chord distance closely approximates spherical geodesic distance while remaining simple to compute from the template coordinates. The adjacency is symmetrized, and disconnected masked islands are connected within each hemisphere as described in Appendix B.1. Registration, graph weights, and the resulting spectral basis are computed before training and held fixed. These template-space distances should not be interpreted as geodesics on each subject's folded cortical surface.

Spectral representation. Rather than operating directly on the full cortical graph, we represent the registered response on a low-frequency spectral basis that provides a compact coordinate system for topology-aware transport. We retain the $K = 128$ lowest eigenpairs of L_{sym}, written as $U_K \in \mathbb{R}^{N \times K}$ and $\Lambda_K = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$. The spectral coordinates of the registered response and their retained surface reconstruction are $x_1 = U_K^\top D^{-1/2} y$,$y_K = D^{1/2} U_K x_1$. For a spectral state x, the normalized coordinates $q = U_K x$ correspond to the surface signal $D^{1/2} q$, with graph variation $q^\top L_{\mathrm{sym}} q = x^\top \Lambda_K x$. The transported cortical state is restricted to this K-dimensional subspace, while the original functional region tokens are retained through a residual path. Thus spectral truncation introduces low-frequency cortical context without replacing information already represented by the regional decoder.

Conditional spectral flow. We model cortical transport as a continuous process rather than a sequence of discrete graph updates. Specifically, we use the heat-process specialization of Topological Flow Matching (TFM), whose reference dynamics are defined directly by the graph spectrum. For branch b, let $R_b^{(0)} \in \mathbb{R}^{128 \times 512}$ denote its tokenizer output before regional self-attention. A learned projection of its pooled representation provides the branch condition $\mathbf{c}_b = \Pi_{c,b}(\mathrm{Mean}_{\mathrm{region}}\, R_b^{(0)})$. Each branch then integrates

$$\frac{d\tilde{x}_t^b}{dt} = -\kappa \Lambda_K \tilde{x}_t^b + v_{\theta_b}(t, \tilde{x}_t^b, \mathbf{c}_b), \qquad \tilde{x}_0^b = x_0^{\mathrm{ref}}, \quad \kappa = 1. \tag{2}$$

Here the heat term contracts higher-eigenvalue modes more strongly, providing a topology-dependent bias toward signals that vary gradually over the cortical surface, while the learned control v_{θ_b} can counteract this drift when required by the data. Both branches use the measured cortical coordinates x_1 as their flow target. They share the same anatomical coordinates but have independent conditions and control networks, allowing structural and semantic supervision to shape different transported cortical states. We solve the dynamics with 16 RK4 steps; architectural and numerical details are given in Appendix B.2.

Region-token update. The transported cortical state must finally be returned to the functional representation used by the visual decoders. We reconstruct the terminal surface signal as $\tilde{y}_b = D^{1/2} U_K \tilde{x}_1^b$ and pool it back to the functional groups. Let $A \in \mathbb{R}_+^{C \times N}$ denote the fixed assignment from surface nodes to functional groups, accumulated over the training subjects. For each group c, we compute transported and observed summaries

$$g_{b,c} = \frac{\sum_i A_{ci} \tilde{y}_{b,i}}{\sum_i A_{ci} + \epsilon}, \qquad y_c = \frac{\sum_i A_{ci} m_i y_i}{\sum_i A_{ci} m_i + \epsilon}, \qquad \epsilon = 10^{-8}. \tag{3}$$

A group with no observed mass receives $y_c = 0$. Defining $\chi_c = \mathbf{1}[\sum_i A_{ci} > 0]$ to indicate support in the training cohort, we update each original region token through a branch-specific residual adapter:

$$\widetilde{R}_{b,c} = R_{b,c}^{(0)} + \chi_c \tanh(\rho_b)\ \mathrm{MLP}_{\mathrm{topo},b}\Big([R_{b,c}^{(0)}, g_{b,c}, y_c]\Big). \tag{4}$$

Equation 4 injects the topology-derived summaries directly into the functional region tokens: the adapter output is added to the tokenizer representation $R_b^{(0)}$, producing the topology-enhanced tokens $\widetilde{R}_b$. These updated tokens replace $R_b^{(0)}$ as the input to the subsequent regional self-attention and visual-query readout. Thus, TCCF acts as a residual refinement of the original functional representation rather than as a separate prediction pathway. Groups without surface support receive no topology residual. The resulting tokens produce hidden representations h_b and visual predictions $\hat{F}_b$.

3.3 Branch-specific training

The structural and semantic pathways use the same objective form but have disjoint trainable parameters. Each branch receives cortical flow supervision from the registered response and visual supervision from the corresponding structural or semantic image features.

Voxel sampling. During training, observed voxels are randomly subsampled before the fixed registration M_s, with the observation mask updated accordingly. Inference uses all available voxels.

Flow supervision. The control field transports a Gaussian reference state toward the measured cortical endpoint. For target coordinates x_1 from Equation 3.2, we sample $x_0^{\mathrm{tr}} \sim \mathcal{N}(0, I_K)$ and $t \sim \mathcal{U}(0,1)$, evaluate the analytic heat bridge x_t, and use $u_t^\star = \dot{x}_t + \kappa \Lambda_K x_t$ as the target:

$$\mathcal{L}_{\mathrm{TCCF}}^b = \mathbb{E}\left[\frac{1}{K}\, \| v_{\theta_b}(t, x_t, \mathbf{c}_b) - u_t^\star \|_2^2\right]. \tag{5}$$

The bridge is given in Appendix B.2; feature prediction integrates Equation 2 from the fixed inference reference state.

Visual supervision. The two branches are supervised in their respective visual feature spaces. Following the structural backbone, the structural branch targets five selected levels of VGG-16-BN features, providing multi-scale supervision for visual structure. The semantic branch instead targets the 256×1664 image tokens of the adapted CLIP conditioner used by the diffusion model. This keeps semantic prediction and diffusion conditioning in one feature space.

For either branch $b \in \{V, C\}$, feature regression is averaged over all scalar entries:

$$\mathcal{L}_{\mathrm{MSE}}^b = \frac{\sum_{\ell=1}^{L_b} \| \hat{F}_{b,\ell} - F_{b,\ell} \|_F^2}{\sum_{\ell=1}^{L_b} |F_{b,\ell}|}, \qquad L_V = 5, \quad L_C = 1. \tag{6}$$

Token-level InfoNCE preserves discriminative correspondence using matched positions as positives; sampling details are in Appendix B.3. Each branch is optimized with

$$\mathcal{L}_b = 0.75\, \mathcal{L}_{\mathrm{NCE}}^b + 0.25\, \mathcal{L}_{\mathrm{MSE}}^b + 0.25\, \mathcal{L}_{\mathrm{TCCF}}^b. \tag{7}$$

Visual losses update the readout, topology-enhanced tokens, and integrated state, while $\mathcal{L}_{\mathrm{TCCF}}^b$ supervises the control field. The branches have disjoint parameters, and neither backpropagates through DIP or diffusion.

3.4 CONFIDENCE-ADAPTIVE DIFFUSION

The structural and semantic feature streams need not be equally reliable for a given brain response. We therefore estimate their reliability separately and use the resulting scores to adapt diffusion. Specifically, the structural score controls how strongly the DIP reconstruction is preserved, while the semantic score controls the strength of CLIP conditioning. Rather than predicting reconstruction quality directly, each gate is trained to estimate the agreement between its decoded visual features and the corresponding target features. The gate inputs depend only on decoded quantities available at inference and require no scene-category labels.

Branch-specific descriptors. For each branch b, we construct a descriptor from its predicted visual features and visual-query hidden tokens:

$$\bar{F}_b = \frac{1}{L_b} \sum_{\ell=1}^{L_b} \Pi_{b,\ell} \left(\mathrm{Mean}_{\mathrm{tok}}\, \hat{F}_{b,\ell} \right), \qquad H_b = \Pi_{H,b}(h_b), \tag{8}$$

$$a_b = \left[\bar{F}_b; \mathrm{Mean}_{\mathrm{tok}}\, H_b; \mathrm{Std}_{\mathrm{tok}}\, H_b \right]. \tag{9}$$

A branch-specific MLP maps this descriptor to a scalar agreement estimate

$$w_b = \mathrm{sigmoid}(g_{\omega_b}(a_b)) \in (0, 1). \tag{10}$$

Architectural details of the descriptors and gate networks are provided in Appendix B.4.

Agreement supervision. During gate training, the target agreement is computed from the decoded and ground-truth visual features. For each feature scale, we flatten the predicted and target token arrays in the same positional order and compute their cosine similarity. Averaging across scales and mapping cosine similarity from $[-1, 1]$ to $[0, 1]$ gives

$$r_b = \frac{1}{2} \left[1 + \frac{1}{L_b} \sum_{\ell=1}^{L_b} \cos \left(\mathrm{vec}\, \hat{F}_{b,\ell}, \mathrm{vec}\, F_{b,\ell} \right) \right]. \tag{11}$$

The gate is trained to regress this target using a Huber loss,

$$\mathcal{L}^b_{\mathrm{gate}} = H_{0.1} \left(w_b - \mathrm{stopgrad}(r_b) \right). \tag{12}$$

The structural target averages agreement equally across its five VGG feature levels, while the semantic target measures agreement over the adapted CLIP tokens.

Gate supervision is constructed from a held-out set of stimuli that is disjoint from both topology training and image evaluation. Ground-truth visual features are used only to construct r_b during gate training and are not required at inference. Additional architectural, numerical, and training details are given in Appendix B.4.

The resulting scores w_V and w_C should be interpreted as learned estimates of structural and semantic feature agreement, respectively, rather than probabilities that the final reconstructed image is correct. We next use these scores through a fixed policy to control structural preservation and semantic guidance during diffusion.

3.5 IMAGE RECONSTRUCTION

Image reconstruction combines VGG-based structural inversion with confidence-adaptive semantic diffusion.

Structural inversion. For each response, a DIP generator g_ψ maps a fixed noise input ξ to an image and is optimized to match the decoded VGG features across the five structural feature levels:

$$\mathcal{L}_{\mathrm{DIP}}(\psi) = \frac{1}{5} \sum_{\ell=1}^{5} \frac{\left\| \phi_\ell(\mathcal{R}_{112}(g_\psi(\xi))) - \mathcal{U}_\ell(\hat{V}_\ell) \right\|_F^2}{m_\ell}. \tag{13}$$

Here ϕ_ℓ extracts the corresponding VGG feature map, $\mathcal{U}_\ell$ converts the predicted tokens back to their spatial feature layout, and m_ℓ denotes the number of feature entries. The resulting DIP reconstruction, denoted I_0, provides the structural initialization for diffusion. Additional inversion details are provided in Appendix B.5.

Confidence-adaptive diffusion start. We encode the structural initialization as $z_0 = \mathcal{E}(I_0)$ using the image autoencoder. The structural agreement estimate w_V determines the starting point of the diffusion trajectory:

$$j(w_V) = \mathrm{round}(10 + 8w_V), \qquad j(w_V) \in \{10, \ldots, 18\}, \qquad z_j = z_0 + \sigma_j \epsilon'. \tag{14}$$

The inherited perturbation is detailed in Appendix B.5. A larger w_V selects a later, lower-noise start and therefore preserves more DIP-derived structure.

Confidence-weighted semantic guidance. The semantic agreement estimate w_C independently determines the strength of classifier-free guidance (CFG) (Ho & Salimans, 2022):

$$\gamma(w_C) = 4 + 2w_C, \qquad D_\gamma = D_{\mathrm{ref}} + \gamma(w_C)\,(D_{\mathrm{cond}} - D_{\mathrm{ref}})\,, \tag{15}$$

where D_{cond} and D_{ref} are the decoded-token and reference-conditioned denoiser outputs. Larger w_C strengthens guidance toward the decoded semantic representation.

The independently predicted scores need not agree or sum to one. Decoding the terminal latent yields $\hat{I}$; complete procedures are in Appendices B.5 and B.6.

4 Experiments

4.1 Data, implementation, and evaluation

Dataset and splits. We use the Natural Scenes Dataset (NSD) (Allen et al., 2022), comprising 7T fMRI responses to COCO images (Lin et al., 2014). Training draws on all eight NSD subjects. Topology training, gate fitting, gate development, and testing are disjoint by image ID across subjects; the gates use held-out responses from all eight subjects. We evaluate Subjects 1, 2, 5, and 7 on 1,000 shared test stimuli, compute each metric per subject, and report the unweighted four-subject mean. Appendix B.1 details registration.

Matched reconstruction settings. To isolate topology and confidence adaptation, all component comparisons use the same regional decoders, image generator, DIP inversion, and diffusion sampler. The baseline and topology-only conditions use start 14 and CFG 5, whereas CortexFlow uses Equations 14 and 15. Random streams are paired across conditions; full settings are in Appendix B.3.

Metrics. We report pixel correlation (PixCorr) and grayscale SSIM, plus two-way identification in AlexNet-2, AlexNet-5, Inception, and CLIP feature spaces. Identification compares each reconstruction's matched target against every other target in the evaluation set. EfficientNet-B1 and SwAV-ResNet50 are correlation distances; lower is better. We report all eight metrics separately rather than combining them into one score.

4.2 Comparison with existing reconstruction methods

Table 1 compares CortexFlow with prior NSD reconstruction methods. Because their subject sets, training data, and evaluation protocols differ, we treat published values as external references and use the matched component comparison below for controlled attribution. The lower block follows the one-hour subject-transfer protocol: each method is trained on seven source subjects, then adapted to a held-out target subject using one hour of the target subject's fMRI. Duala reports this one-hour setting; From Flat to Round uses joint four-subject training with approximately 9K surface voxels.

Pixel metrics test spatial fidelity, while identification tests whether a reconstruction remains closer to its target than to distractors. Under our four-subject protocol, CortexFlow obtains 0.392 PixCorr, 0.502 SSIM, and 98.6%, 99.3%, 97.8%, and 96.6% identification. Because published protocols differ, Table 1 provides external context rather than a controlled ranking; the matched comparison below supports component attribution.

4.3 Analysis

Component Ablation. Adding topology increases PixCorr by 0.021 (0.366 to 0.387) and SSIM by 0.039 (0.452 to 0.491), while improving all six feature-space metrics. Enabling the confidence gates

Table 1: Reconstruction quality on NSD. Identification scores are percentages.

Method	Low-Level PixCorr ↑	SSIM ↑	Alex2 ↑	High-Level Alex5 ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
MindEye (Scotti et al., 2023)	0.319	0.360	92.8%	96.9%	94.6%	93.3%	0.648	0.377
Brain-Diffuser (Ozcelik & VanRullen, 2023)	0.273	0.365	94.4%	96.6%	91.3%	90.9%	0.728	0.421
Takagi–Nishimoto (Takagi & Nishimoto, 2023)	0.246	0.410	78.9%	85.6%	83.8%	82.1%	0.811	0.504
DREAM (Xia et al., 2024a)	0.274	0.328	93.9%	96.7%	93.4%	94.1%	0.645	0.418
NeuroVLA (Shen et al., 2024)	0.265	0.357	93.1%	97.1%	96.8%	**97.5%**	0.633	0.321
MindBridge (Wang et al., 2024a)	0.151	0.263	87.7%	95.5%	92.4%	94.7%	0.712	0.418
NeuroPictor (Huo et al., 2024)	0.229	0.375	96.5%	98.4%	94.5%	93.3%	0.639	0.350
MindEye2 (Scotti et al., 2024)	0.322	0.431	96.1%	98.6%	95.4%	93.0%	0.619	0.344
From Flat to Round (Yu et al., 2025)	0.165	0.305	78.2%	89.0%	85.1%	88.3%	0.733	0.398
Brain-IT (Beliy et al., 2026)	0.366	0.452	96.4%	98.8%	96.4%	95.4%	0.599	0.326
CortexFlow(Ours)	**0.392**	**0.502**	**98.6%**	**99.3%**	**97.8%**	96.6%	**0.574**	**0.311**
Results on 1 hour (out of 40)								
MindEye2	0.195	0.419	84.2%	90.6%	81.2%	79.2%	0.810	0.468
MindTuner (Gong et al., 2025)	0.224	0.420	87.8%	93.6%	84.8%	83.5%	0.780	0.440
Duala (Li et al., 2026)	0.230	0.416	87.9%	93.5%	85.4%	83.5%	0.781	0.445
Brain-IT	0.332	0.469	97.2%	97.8%	94.4%	92.5%	0.651	0.372
CortexFlow(Ours)	**0.338**	**0.506**	**97.7%**	**98.3%**	**95.7%**	**93.9%**	**0.641**	**0.353**

Table 2: Ablation study of topology and confidence gates under the same four-subject evaluation protocol.

Configuration	Low-Level PixCorr ↑	SSIM ↑	Alex2 ↑	High-Level Alex5 ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
Baseline	0.366	0.452	96.4%	98.8%	96.4%	95.4%	0.599	0.326
Baseline+topology	0.387	0.491	98.2%	99.1%	97.7%	95.9%	0.581	0.316
Baseline+confidence gates	0.373	0.472	97.1%	98.9%	96.9%	95.7%	0.588	0.322
CortexFlow(Ours)	**0.392**	**0.502**	**98.6%**	**99.3%**	**97.8%**	**96.6%**	**0.574**	**0.311**

then increases PixCorr by a further 0.005 and SSIM by 0.011, reaching 0.392 and 0.502, respectively, with further improvements in all six feature-space metrics.

4.4 Qualitative reconstruction examples

Figure 3 compares seven Subject 1 test stimuli with their CortexFlow reconstructions.

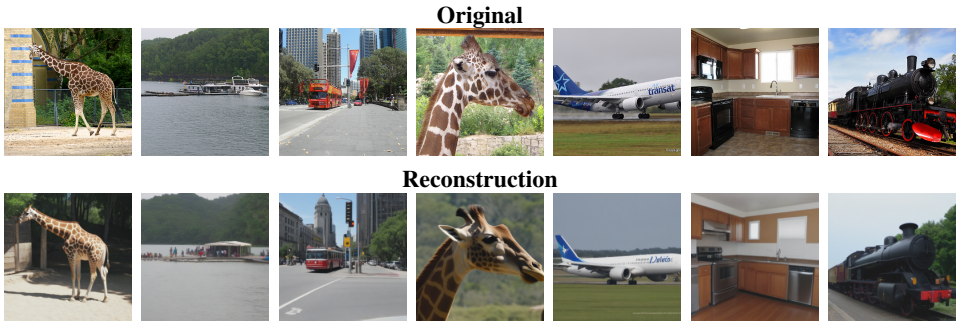


Figure 3: Subject 1 qualitative comparisons. Columns align each viewed stimulus in the top row with its CortexFlow reconstruction in the bottom row.

4.5 Cortical analysis of topology-induced updates

For response n, branch $b \in \{V, C\}$, and group c, we measure the normalized topology-induced token update

$$a_{b,c}^{(n)} = \frac{\|\widetilde{R}_{b,c}^{(n)} - R_{b,c}^{(0,n)}\|_2}{\|R_{b,c}^{(0,n)}\|_2 + \epsilon}. \tag{16}$$

We project the group-level contributions to the common surface through the fixed assignment A from Equation 3, using $H_b^{(n)}(i) = \sum_c A_{ci} a_{b,c}^{(n)} / (\sum_c A_{ci} + \epsilon)$. Figure 4 shows the Subject 1 median over 1,000 test stimuli on a shared scale. Appendix E gives normalized contrasts, parcel and cross-subject summaries, diagnostics, and example flows.

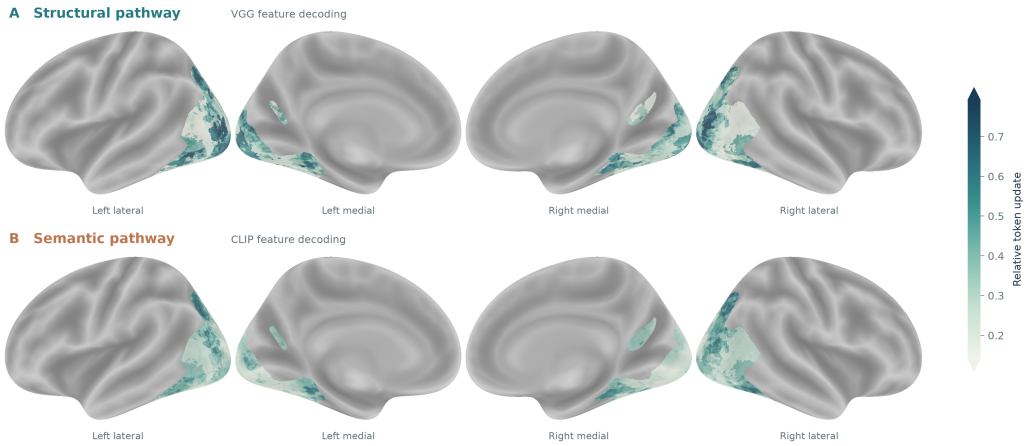


Figure 4: **Cortical projection of branch-specific token updates.** Subject 1 median relative token updates over 1,000 test stimuli after a fixed 128-mode basis (VGG top; CLIP bottom). Both branches share a scale; gray is outside support. Values are model updates, not neural activation or causality.

5 Discussion and Conclusion

CortexFlow addresses two coupled sources of structure in fMRI reconstruction: the organization of cortical responses and the unequal reliability of decoded visual evidence. A shared cortical basis supplies comparable coordinates, while independent VGG and CLIP flows allow surface context to serve distinct structural and semantic objectives. The confidence gates then determine how those decoded signals enter generation rather than treating every response as equally reliable. The matched progression in Table 2 supports both stages: topology raises PixCorr/SSIM from 0.366/0.452 to 0.387/0.491, and confidence adaptation further raises them to 0.392/0.502; all six feature-space metrics follow the same ordering. This controlled comparison establishes the benefit of the proposed topology-aware refinement as a whole, while isolating the contribution of continuous spectral transport from simpler graph operators remains an important direction for future work.

The cortical projection offers a model-level interpretation of this improvement. Related but non-identical update maps are expected because the branches share anatomical coordinates but not visual supervision or trainable parameters. This comparison is more informative than assigning either branch to a predefined set of areas, but it remains an analysis of the decoder: spatial alignment does not establish neural selectivity, biological transport, or causal mechanism.

Limitations. First, fsaverage registration, a fixed neighborhood graph, and the fixed $K = 128$ spectral truncation approximate subject-specific cortical geometry and may suppress fine-scale variation; sensitivity to spectral resolution and comparison with simpler graph operators remain to be studied. Second, feature agreement is a tractable supervision signal for confidence, not a calibrated probability of reconstruction correctness; future work should test calibration under acquisition noise and distribution shift. Third, the branches are trained independently and interact only through the generator, so the model does not learn an explicit cross-branch consistency objective. Finally, the continuous flow is a numerical transport mechanism, and its trajectory should not be interpreted as neural dynamics in biological time.

Ethics Statement

Reconstructions are uncertain estimates, not private-mind readings; neuroimaging use requires consent, privacy protection, and dataset compliance.

REPRODUCIBILITY STATEMENT

Appendices B.1–B.6 specify implementation, training, inference, and evaluation details; Section 4 specifies the data splits and matched comparison protocol.

AI USE STATEMENT

Yes, I use it to assist with writing tasks, such as improving clarity, refining expressions, correcting grammar, and polishing the overall style of written content.

REFERENCES

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, 2022.

Roman Beliy, Navve Wasserman, Amit Zalcher, and Michal Irani. The wisdom of a crowd of brains: A universal brain encoder, 2024. URL https://arxiv.org/abs/2406.12179.

Roman Beliy, Amit Zalcher, Jonathan Kogman, Navve Wasserman, and Michal Irani. Brain-IT: Image reconstruction from fMRI via brain-interaction transformer. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c9fd326fd03eaf52f672c31cde9658af-Abstract-Conference.html.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments, 2021. URL https://arxiv.org/abs/2006.09882.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations, 2019. URL https://arxiv.org/abs/1806.07366.

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, 2023. URL https://arxiv.org/abs/2211.06956.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering, 2017. URL https://arxiv.org/abs/1606.09375.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

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

Zalan Fabian, Berk Tinaz, and Mahdi Soltanolkotabi. Adapt and diffuse: Sample-adaptive reconstruction via latent diffusion models, 2024. URL https://arxiv.org/abs/2309.06642.

Bruce Fischl, Martin I. Sereno, and Anders M. Dale. Cortical surface-based analysis.II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9(2):195–207, 1999. doi: 10.1006/nimg.1998.0396.

Zixuan Gong, Qi Zhang, Guangyin Bao, Lei Zhu, Rongtao Xu, Ke Liu, Liang Hu, and Duoqian Miao. MindTuner: Cross-subject visual decoding with visual fingerprint and semantic correction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(13):14247–14255, 2025. doi: 10.1609/aaai.v39i13.33560.

Kalanit Grill-Spector and Kevin S Weiner. The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15(8):536–548, 2014. doi: 10.1038/nrn3747.

U. Guclu and M. A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015. ISSN 1529-2401. doi: 10.1523/jneurosci.5023-14.2015. URL http://dx.doi.org/10.1523/JNEUROSCI.5023-14.2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Jingyang Huo, Yikai Wang, Xuelin Qian, Yun Wang, Chong Li, Jianfeng Feng, and Yanwei Fu. NeuroPictor: Refining fMRI-to-image reconstruction via multi-individual pretraining and multi-level modulation. In *Computer Vision – ECCV 2024*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/6746_ECCV_2024_paper.php.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015. URL https://arxiv.org/abs/1502.03167.

Eric R Kandel, James H Schwartz, Thomas M Jessell, Steven A Siegelbaum, and A J Hudspeth. *Principles of Neural Science*. McGraw-Hill Education, New York, 5th edition, 2013.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022. URL https://arxiv.org/abs/2206.00364.

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

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2016. URL https://arxiv.org/abs/1609.02907.

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. URL https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

Shumeng Li, Jintao Guo, Jian Zhang, Yulin Zhou, Luyang Cao, and Yinghuan Shi. Duala: Dual-level alignment of subjects and stimuli for cross-subject fMRI decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42722–42731, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014. URL https://arxiv.org/abs/1405.0312.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017. URL https://arxiv.org/abs/1711.05101.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations, 2021. URL https://arxiv.org/abs/2108.01073.

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

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

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

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis, 2023. URL https://arxiv.org/abs/2307.01952.

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, 2021. URL https://arxiv.org/abs/2103.00020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

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. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4ddab70bf41ffe5d423840644d3357f4-Abstract-Conference.html.

Paul S. 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. 2024. URL https://arxiv.org/abs/2403.11207.

Guobin Shen, Dongcheng Zhao, Xiang He, Linghao Feng, Yiting Dong, Jihang Wang, Qian Zhang, and Yi Zeng. Neuro-vision to language: Enhancing brain recording-based visual reconstruction and language interaction, 2024. URL https://arxiv.org/abs/2404.19438.

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

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014. URL https://arxiv.org/abs/1409.1556.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2020a. URL https://arxiv.org/abs/2010.02502.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2020b. URL https://arxiv.org/abs/2011.13456.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015. URL https://arxiv.org/abs/1512.00567.

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

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2019. URL https://arxiv.org/abs/1905.11946.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

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

Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. MindBridge: A cross-subject brain decoding framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11333–11342, 2024a.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/tip.2003.819861. URL https://doi.org/10.1109/tip.2003.819861.

Zicheng Wang, Zhen Zhao, Luping Zhou, and Parashkev Nachev. UniBrain: A unified model for cross-subject brain decoding, 2024b. URL https://arxiv.org/abs/2412.19487.

Navve Wasserman, Roman Beliy, Roy Urbach, and Michal Irani. Functional Brain-to-Brain Transformation with No Shared Data, 2024. URL https://arxiv.org/abs/2404.11143.

Kacper Wyrwal, İsmail İlkan Ceylan, and Alexander Tong. Topological flow matching. In *International Conference on Learning Representations (ICLR)*, 2026.

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 (WACV)*, pp. 8226–8235, 2024a.

Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. UMBRAE: Unified multimodal brain decoding. In *Computer Vision – ECCV 2024*, 2024b. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/1133_ECCV_2024_paper.php.

Sijin Yu, Zijiao Chen, Wenxuan Wu, Shengxian Chen, Zhongliang Liu, Jingxin Nie, Xiaofen Xing, Xiangmin Xu, and Xin Zhang. From flat to round: Redefining brain decoding with surface-based fmri and cortex structure. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 6248–6257, October 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. URL https://arxiv.org/abs/1801.03924.

A More Ablations and Analysis

Table 3: Effect of the number of nearest neighbors k. Our default is $k = 12$.

	Low-Level				High-Level			
k	PixCorr ↑	SSIM ↑	Alex2 ↑	Alex5 ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
8	0.391	0.501	98.5%	99.4%	97.4%	96.2%	0.572	0.317
12 (default)	0.392	0.502	98.6%	99.3%	97.8%	96.6%	0.574	0.311
16	0.394	0.497	98.2%	99.2%	97.5%	96.8%	0.569	0.314

Neighborhood size. Table 3 shows modest variation across k. The default $k = 12$ gives the highest SSIM and lowest SwAV distance, while $k = 16$ improves PixCorr, CLIP, and EfficientNet distance. Smaller neighborhoods may limit the spatial context supplied by the spectral representation; larger ones may mix less-local responses and reduce spatial specificity. The results support $k = 12$ as a compromise rather than a uniformly optimal setting.

Table 4: Effect of the diffusion start-index mapping range. Our default is $[10, 18]$.

	Low-Level				High-Level			
Start-index range	PixCorr ↑	SSIM ↑	Alex2 ↑	Alex5 ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
$[12, 16]$	0.390	0.499	98.7%	99.0%	97.4%	96.1%	0.577	0.318
$[10, 18]$ (default)	0.392	0.502	98.6%	99.3%	97.8%	96.6%	0.574	0.311
$[8, 20]$	0.394	0.497	98.3%	99.5%	97.6%	96.8%	0.574	0.315

Diffusion start range. In Table 4, $[10, 18]$ achieves the highest SSIM and lowest SwAV distance, whereas $[8, 20]$ improves PixCorr, Alex5, and CLIP. Narrowing the range around index 14 limits adaptation to structural confidence. Widening it permits more extreme starts: earlier, noisier starts may overwrite useful DIP structure, while later starts may retain initialization errors. The default balances structural preservation and diffusion refinement.

Table 5: Effect of the CFG strength mapping range. Our default is $[4, 6]$.

	Low-Level				High-Level			
CFG range	PixCorr ↑	SSIM ↑	Alex2 ↑	Alex5 ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
$[2, 8]$	0.379	0.476	98.7%	98.9%	97.6%	96.8%	0.584	0.328
$[4, 6]$ (default)	0.392	0.502	98.6%	99.3%	97.8%	96.6%	0.574	0.311
$[4.5, 5.5]$	0.390	0.499	98.3%	98.7%	97.4%	96.1%	0.577	0.315

CFG range. Table 5 favors $[4, 6]$ over $[4.5, 5.5]$ on all reported metrics. The wider $[2, 8]$ range slightly improves Alex2 and CLIP but reduces PixCorr and SSIM to 0.379 and 0.476. Weak guidance may underuse decoded semantics, while strong guidance may amplify prediction errors. Conversely, a very narrow range limits response-specific adjustment. These results support moderate guidance variation around 5.

B Implementation Details

B.1 Registration and cortical graph

Node ordering and coverage. We use a cortical surface representation (Fischl et al., 1999). The graph uses the union of the NSD `nsdgeneral` and RSC fsaverage masks: 19,004 left-hemisphere and 20,544 right-hemisphere vertices. Nodes are ordered by ascending template vertex index within each hemisphere, left before right. Subject-specific response indices map into this common ordering. If multiple observed entries map to one node, their values are averaged. Missing nodes are filled with zero and tracked by a per-sample observation mask m. Subject-specific mappings preserve each participant's observed vertices within the common template. Training uses all eight NSD subjects, including the construction of the shared surface-to-group assignment. Evaluation uses Subjects 1, 2, 5, and 7, with 1,000 shared test stimuli per subject, and averages the four subject-level scores. The evaluation subset does not restrict the subjects contributing to training.

Edges and weights. We retain fsaverage mesh triangles whose three vertices lie in the mask and add their edges. Within each hemisphere, we also add $k = 12$ nearest-neighbor edges in the spherical template. Let d_{ij} be spherical chord distance and ℓ_i the median distance from node i to its k neighbors. Ordinary edges have weights

$$W_{ij} = \exp\left(-\frac{d_{ij}^2}{2\ell_i \ell_j}\right). \tag{17}$$

Chord distance provides a practical surface-distance proxy across masked gaps. For each remaining disconnected island, we add its shortest spherical connection to the largest component in the same hemisphere, with the median ordinary-edge weight. No edge crosses hemispheres. Symmetrization gives two connected components. We compute the lowest 128 eigenpairs of the sparse normalized Laplacian and fix each eigenvector's sign by making its largest-magnitude entry positive. The resulting basis is stored with the model and reused at inference, including the chosen coordinates within repeated-eigenvalue subspaces. For this normalized Laplacian, the nullspace is spanned by the degree-weighted component indicators, $D^{1/2}\mathbf{1}_{\mathrm{LH}}$ and $D^{1/2}\mathbf{1}_{\mathrm{RH}}$.

Pooling to functional regions. Universal voxel-centric encoding and functional brain-to-brain transformation provide complementary approaches to combining information across individuals (Beliy et al., 2024; Wasserman et al., 2024).

Let $A \in \mathbb{R}_+^{128\times N}$ be the assignment matrix obtained by mapping the backbone's voxel-to-cluster memberships onto the surface and accumulating counts over all eight training subjects. Nonempty rows are normalized to sum to one; empty rows remain zero. For either branch, write $\tilde{y} = D^{1/2}U_K\tilde{x}_1^b$; pooling is

$$g_c = \frac{\sum_i A_{ci}\tilde{y}_i}{\sum_i A_{ci} + \epsilon}, \qquad y_c = \frac{\sum_i A_{ci} m_i y_i}{\sum_i A_{ci} m_i + \epsilon}, \quad \epsilon = 10^{-8}. \tag{18}$$

Generated signals cover the full graph, while observed signals require masked normalization. A group without observed mass receives $y_c = 0$. The cohort support mask $\chi_c = \mathbf{1}[\sum_i A_{ci} > 0]$ additionally suppresses the adapter residual for groups absent from the selected subjects. This differs from m, which records the observations available in a particular input. Any unsupported groups receive no topology-adapter residual, preserving the backbone's 128-group indexing for downstream attention.

B.2 Analytic heat bridge and control

Spectral graph filtering and graph convolutions encode neighborhood structure (Defferrard et al., 2017; Kipf & Welling, 2016), while neural differential equations model continuous representation dynamics (Chen et al., 2019).

The following formulas are the heat-process specialization of TFM (Wyrwal et al., 2026). Write $a_k = \kappa\lambda_k$ for the contraction rate of mode k. The zero-noise conditional bridge between spectral endpoints x_0 and x_1 is

$$x_{t,k} = \begin{cases} \dfrac{\sinh(a_k(1-t))}{\sinh(a_k)} x_{0,k} + \dfrac{\sinh(a_k t)}{\sinh(a_k)} x_{1,k}, & a_k > 0, \\ (1-t)x_{0,k} + t x_{1,k}, & a_k = 0. \end{cases} \tag{19}$$

The control is the path derivative minus the reference drift, i.e., $u_t^\star = \dot{x}_t + \kappa\Lambda_K x_t$, giving

$$u_{t,k}^\star = \begin{cases} \dfrac{2a_k e^{-a_k(1-t)}}{1 - e^{-2a_k}}(x_{1,k} - e^{-a_k}x_{0,k}), & a_k > 0, \\ x_{1,k} - x_{0,k}, & a_k = 0. \end{cases} \tag{20}$$

These expressions satisfy $\dot{x}_t = -\kappa\Lambda_K x_t + u_t^\star$. The zero-mode branch uses linear interpolation. The implementation evaluates exponential ratios with `expm1` and uses this branch for $|a_k| < 10^{-6}$. Training draws a Gaussian x_0^{tr} independently for every example and clamps uniformly drawn times to $[10^{-3}, 1 - 10^{-3}]$ for numerical stability. There is no optimal-transport reassignment of training endpoints.

The control network projects the 128-dimensional state, a 128-dimensional sinusoidal time embedding, and a 512-dimensional regional condition to a 512-dimensional hidden space, followed by four

residual MLP blocks and a spectral output projection. Each branch's feature-prediction pass starts from the same serialized seed-0 Gaussian state x_0^{ref} and uses RK4 with step size $1/16$. Fixing this state removes sampling variation from the added topology modules; DIP and image diffusion retain their own random streams.

B.3 Visual targets and branch-specific optimization

Structural targets. We follow the backbone's VGG-16-BN representation (Simonyan & Zisserman, 2014; Ioffe & Szegedy, 2015). Structural targets are extracted from the antialiased 112×112 image cache at feature indices 5, 12, 22, 32, and 42. The first scale groups 2×2 locations with stride 2, the second uses stride 1, and the remaining scales use one token per spatial location. Channel replication brings each token to 512 dimensions. The resulting five scales contain $(3136, 3025, 784, 196, 49)$ tokens, respectively, for 7,190 packed tokens in total.

Semantic targets. The semantic target is the 256×1664 token output of the adapted ViT-bigG/14 image conditioner (Dosovitskiy et al., 2021; Radford et al., 2021) stored with the stage-2 diffusion weights. We use these exact weights for target extraction rather than the CLIP model used for evaluation, ensuring that semantic decoding and diffusion conditioning share the same feature space. The 256×256 RGB image cache is converted from uint8 to $[-1, 1]$, resized to 224×224 using antialiased bicubic interpolation with aligned corners, mapped to $[0, 1]$, and normalized with the conditioner's CLIP mean and standard deviation. Targets are cached in float16 and converted to float32 when computing losses. The cache manifest associates each token array with its global image ID and a hash of the corresponding stage-2 weights.

Token contrastive and regression losses. For each feature scale ℓ, we sample matched spatial positions from the prediction and target. The structural branch samples $(512, 512, 128, 64, 16)$ positions across its five scales, while the semantic branch samples 128 positions from its single token scale. Flattening the batch and sampled-position dimensions gives n_ℓ paired token vectors $\{(p_i, q_i)\}_{i=1}^{n_\ell}$. We define $S_{ij} = \cos(p_i, q_j)/\tau$ with $\tau = 0.125$. For the five-scale structural branch, the symmetric token-level contrastive objective (van den Oord et al., 2019) is

$$\mathcal{L}_{\mathrm{NCE}} = -\frac{1}{5} \sum_{\ell=1}^{5} \frac{1}{2n_\ell} \sum_{i=1}^{n_\ell} \left[\log \frac{e^{S_{ii}}}{\sum_j e^{S_{ij}}} + \log \frac{e^{S_{ii}}}{\sum_j e^{S_{ji}}} \right]. \tag{21}$$

The semantic branch uses the same objective with a single scale. Negatives include other sampled positions from both the same image and other images; distributed training additionally gathers sampled features across ranks when enabled. The regression term is the mean squared error over all scalar entries of the prediction and target. For the structural branch, this naturally weights the five scales according to their token counts.

Trainable topology modules. The structural base is the regional VGG decoder, while the semantic base is restored from the same stage-2 checkpoint used by the generator. Each branch's topology-specific modules comprise the control field v_{θ_b}, condition projection $\Pi_{c,b}$, topology residual adapter $\mathrm{MLP}_{\mathrm{topo},b}$, and residual scale ρ_b. Gradients from the visual objectives pass through the regional attention and visual-query readout.

The topology residual adapter takes $[R_{b,c}^{(0)}, g_{b,c}, y_c]$ as input. Since $R_{b,c}^{(0)} \in \mathbb{R}^{512}$ and $g_{b,c}$ and y_c are scalars, its input dimension is 514. We use an MLP with dimensions $514 \to 256 \to 512$ and GELU activation. Its output layer is initialized to zero, so the topology residual is initially zero and the model starts from the original regional representation. The learnable residual scale ρ_b is initialized to 1. After the residual update, the 512-dimensional region tokens are processed by the branch-specific regional attention and visual-query readout to produce h_b and $\hat{F}_b$.

Voxel sampling. Each training example samples 15,000 voxels with replacement. The sampled response is registered using the fixed subject-specific mapping M_s, and the observation mask m is constructed from the same sample. Repeated entries mapped to the same surface node are averaged, while nodes receiving no sampled measurement are zero-filled. The spectral target x_1 used for flow supervision is computed from this same registered input. At evaluation time, voxel sampling is disabled and all observed voxels are used for feature extraction and reconstruction.

Optimization protocol. We use the loss weights $\lambda_{\mathrm{NCE}} = 0.75$, $\lambda_{\mathrm{MSE}} = 0.25$, and $\lambda_{\mathrm{TCCF}} = 0.25$ for both branches. These coefficients are treated as relative weights and are not constrained to sum to one. Both branches are trained on responses from all eight subjects using the same training-stimulus partition, but are optimized as separate jobs with disjoint trainable parameters and separate checkpoints. Each branch is trained for 30 epochs with AdamW (Loshchilov & Hutter, 2017), a learning rate of 2×10^{-4}, weight decay 10^{-4}, gradient-norm clipping at 1, batch size 16, and one GPU per job. The final epoch supplies the topology weights used for evaluation, while epoch-boundary checkpoints support resuming interrupted runs.

B.4 Confidence Gate Details

Gate architecture. The structural and semantic gates use independent projection and MLP parameters. For the structural branch, each of the five VGG feature means is projected to 512 dimensions and the projected representations are averaged across feature levels. For the semantic branch, the mean of the 1,664-channel adapted CLIP tokens is projected directly to 512 dimensions. The hidden tokens are independently projected to the same width and summarized by their token-wise mean and population standard deviation, as in Equation 9. Consequently, the feature summary, hidden mean, and hidden standard deviation are each 512-dimensional, giving $a_b \in \mathbb{R}^{1536}$.

Each branch uses an independent gate MLP with dimensions $1536 \to 256 \to 128 \to 1$ and GELU activations in the hidden layers. A final sigmoid maps the scalar output to $w_b \in (0, 1)$. Although the gate input pools over token positions, spatial correspondence is retained when constructing the agreement target r_b in Equation 11.

Agreement-target computation. For each feature level, the predicted and target arrays are flattened in the same positional order before cosine similarity is computed. Cosine denominators are clamped away from zero for numerical stability, and the resulting agreement target is clipped to $[0, 1]$ as a numerical safeguard. The structural target computes agreement independently over the five VGG feature levels and averages them with equal weight, whereas the semantic target compares the corresponding adapted CLIP token arrays.

The Huber loss used in Equation 12 is

$$H_\delta(e) = \begin{cases} e^2/2, & |e| \le \delta, \\ \delta(|e| - \delta/2), & \text{otherwise}, \end{cases} \tag{22}$$

with $\delta = 0.1$.

Gate training protocol. Gate training is performed after the structural and semantic feature decoders have been trained. Supervision uses held-out responses from all eight subjects. We partition these responses into fitting and development sets by global stimulus ID using seed 0, keeping all responses to the same image in the same partition across subjects. Both partitions are disjoint from the stimuli used for topology training and final image evaluation.

Each gate is trained for 30 epochs with AdamW, batch size 32, learning rate 10^{-4}, weight decay 10^{-4}, and gradient-norm clipping at 1. We select the checkpoint with the lowest development Huber loss independently for the structural and semantic gates. The resulting gates are used for reconstruction evaluation on Subjects 1, 2, 5, and 7.

Gate checkpoints are associated with the topology checkpoint, base decoder weights, and spectral basis used to produce their inputs. The semantic gate is additionally associated with the stage-2 diffusion weights.

B.5 DIP Inversion and Image Diffusion

DIP inversion. For each response, the DIP generator g_ψ is optimized independently to invert the decoded structural features. As defined in Equation 13, the generated image is resized to 112×112 and matched to the decoded VGG representation at five feature levels. The feature extractors ϕ_ℓ include ImageNet input normalization and VGG-16-BN feature extraction. The unpacking operator $\mathcal{U}_\ell$ converts the predicted structural tokens back to their corresponding feature-map layouts; it averages

duplicated channel halves at scales 1 and 3 and overlapping patches at scale 2. The normalization factor m_ℓ is the number of entries in the corresponding feature map.

The stimulus image is never used during this optimization: DIP receives supervision only from the VGG features predicted from the brain response. We run 1,800 optimization iterations for each response and resize the resulting image I_0 to 256×256 before diffusion refinement.

Confidence-to-diffusion mapping. The two learned agreement scores control separate components of the diffusion procedure. Structural agreement determines the diffusion start index according to

$$j(w_V) = \mathrm{round}(10 + 8w_V), \qquad j \in \{10, \ldots, 18\}, \tag{23}$$

on the 38-transition sampler grid. All nine integer start indices are selectable, with half-way ties rounded to the nearest even integer. Semantic agreement determines the classifier-free guidance coefficient according to

$$\gamma(w_C) = 4 + 2w_C, \qquad \gamma \in (4, 6). \tag{24}$$

The start-index and guidance ranges are fixed design choices and do not assume that the learned gate scores span the full interval.

Diffusion initialization. Diffusion priors build on denoising and score-based generative modeling (Ho et al., 2020; Song et al., 2020b); latent diffusion and SDXL make such priors practical at higher image resolutions (Rombach et al., 2022; Podell et al., 2023). Sampling trajectories and noise parameterizations also affect the trade-off between reconstruction fidelity and generative refinement (Song et al., 2020a; Karras et al., 2022).

Let $z_0 = \mathcal{E}(I_0)$ denote the autoencoder latent of the DIP reconstruction, and let $\sigma_0 \geq \cdots \geq \sigma_{38} = 0$ denote the noise schedule of the 38-transition sampler. At the selected start index j, initialization follows the inherited unCLIP sampling procedure:

$$z_j = z_0 + \sigma_j(\epsilon + 0.04\eta), \qquad \epsilon \sim \mathcal{N}(0, I), \quad \eta \sim \mathcal{N}(0, 1). \tag{25}$$

A single scalar η is sampled independently for each example and broadcast across its latent dimensions. Denoising then proceeds from transition j through transition 37. Consequently, larger structural agreement selects a later and less noisy initialization, preserving more of the DIP reconstruction, whereas lower agreement allows stronger diffusion refinement.

For compatibility with the inherited sampler API, the wrapper supplies $z_j/\sqrt{1+\sigma_0^2}$ as its initial state. The sampler internally multiplies this input by $\sqrt{1+\sigma_0^2}$ before the first denoising transition, recovering the state z_j in Equation 25.

Semantic guidance and output processing. At every denoising transition, the conditional branch uses the adapted CLIP tokens predicted by the semantic decoder. The reference branch uses independently sampled standard-normal reference tokens of the same shape, with the vector suffix shared between the conditional and reference inputs as in the inherited unCLIP implementation. The two denoiser outputs are combined using the response-specific coefficient $\gamma(w_C)$ from Equation 15. This coefficient is held fixed throughout the denoising trajectory for a given response; the semantic tokens themselves are not rescaled.

After sampling, the terminal latent is decoded to pixel space. The common output routine clamps pixel values to $[0, 1]$ and applies the inherited contrast transformation

$$I \leftarrow 0.8(I - 0.5) + 0.5. \tag{26}$$

The same output processing is applied to every compared reconstruction condition.

B.6 Training and inference algorithms

Algorithm 1 Independent topology optimization and confidence fitting

Fix registration, graph, spectral basis, regional base weights, and visual target encoders.
Build A from all eight training subjects and compute its support mask χ.
for branch $b \in \{V, C\}$ **do**
  Initialize its condition projection, control field, adapter, and residual scale.
  for each topology-training batch (y_v, I) **do**
    Sample voxel entries; encode $R_b^{(0)}$ and $\mathbf{c}_b$ and register the sampled input to obtain x_1.
    Draw x_0^{tr}, t; form $x_t, u_t^\star$ with Equations 19–20.
    Evaluate $\mathcal{L}_{\mathrm{TCCF}}^b$; integrate the branch ODE from x_0^{ref}.
    Update supported region tokens using χ, predict $\hat{F}_b$, and obtain the target F_b.
    Update only this branch's trainable parameters using Equation 7.
  end for
  Compute feature descriptors and r_b on gate-fitting/development stimuli.
  for each gate-training epoch **do**
    Fit this branch's gate with Equation 12; evaluate development Huber loss.
  end for
  Retain this gate's best development checkpoint.
end for

Algorithm 2 Dual-branch CortexFlow reconstruction

Input: response y_v, registration M_s, both trained decoders and gates, image generator.
Register $y = M_s y_v$ and retain observation mask m.
for branch $b \in \{V, C\}$ **do**
  Obtain $R_b^{(0)}$ and $\mathbf{c}_b$; integrate its spectral ODE with 16 RK4 steps.
  Decode its surface signal, pool generated and observed values, and update supported region tokens.
  Predict $\hat{F}_b, h_b$; compute confidence w_b from Equation 9.
end for
Set $j = \mathrm{round}(10 + 8w_V)$ and $\gamma = 4 + 2w_C$.
Invert $\hat{V}$ with DIP to obtain I_0; encode $z_0 = \mathcal{E}(I_0)$.
Initialize z_j with Equation 14; denoise transitions j through 37 under $\hat{Z}$ using the fixed per-response coefficient γ.
Return: the decoded and postprocessed image $\hat{I}$.

C Evaluation Details

Metric implementation. PixCorr and SSIM (Wang et al., 2004) compare the reconstructed and target pixels. AlexNet-2 and AlexNet-5 (Krizhevsky et al., 2012), Inception (Szegedy et al., 2015), and CLIP (Radford et al., 2021) use two-way identification against the 999 other targets within each subject's evaluation set, averaged over reconstruction–distractor pairs. We compute every metric separately for Subjects 1, 2, 5, and 7 and take the unweighted mean of the four scores. Repeated presentations of the same target across subjects do not create additional distractors. CLIP evaluation uses OpenAI ViT-L/14 with a 224×224 bilinear resize, the CLIP normalization, and no crop. SwAV (Caron et al., 2021) uses the official ResNet-50 (He et al., 2015) checkpoint, a bilinear resize, ImageNet normalization, and no crop. EfficientNet (Tan & Le, 2019) and SwAV distances are $1 - \mathrm{corr}$ in their feature spaces. Color SSIM is recorded as an additional pixel-level diagnostic. Deep-feature perceptual metrics offer complementary notions of similarity (Zhang et al., 2018).

D Additional Qualitative Examples

The following panels provide aligned method comparisons, consistency across subjects, intermediate reconstructions, and representative failure cases.

D.1 Comparison with published Brain-IT examples

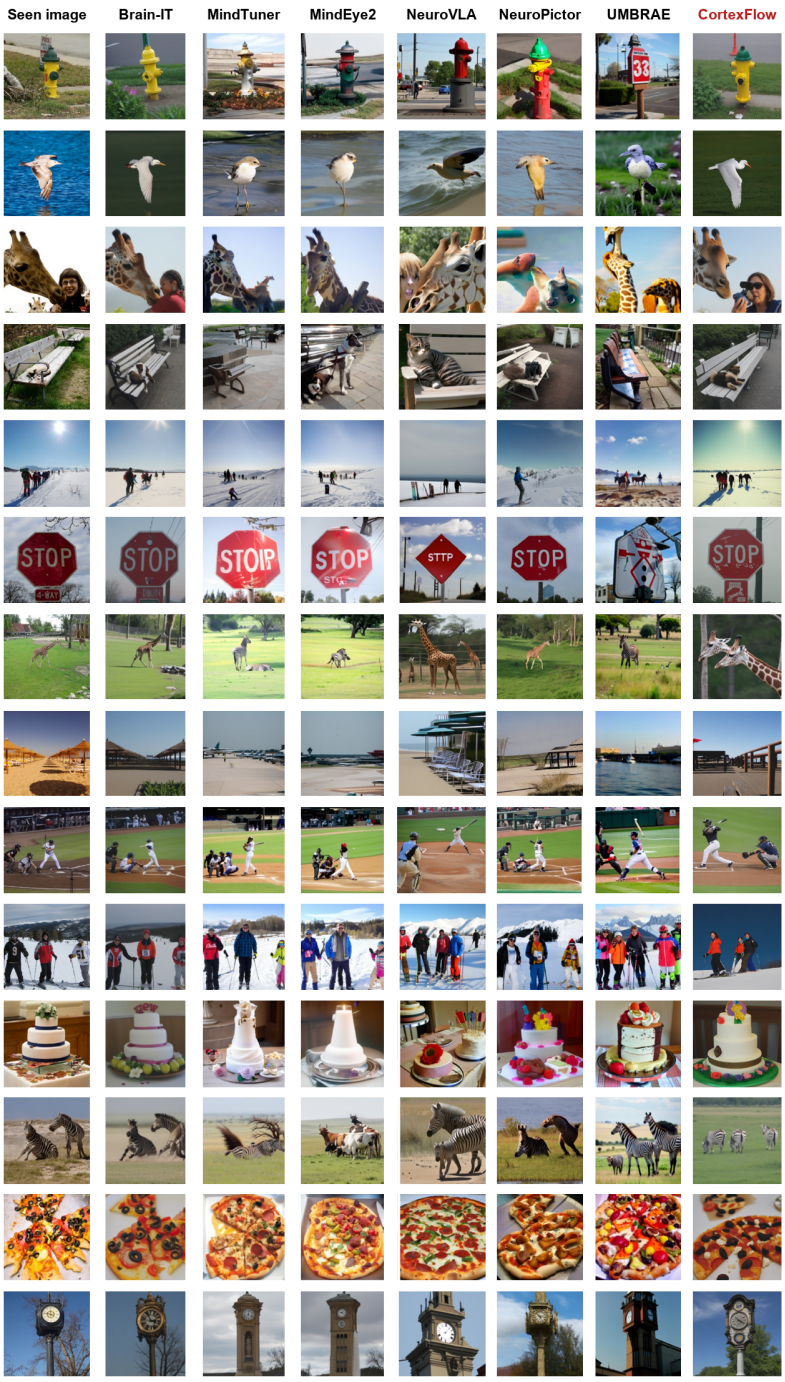


Figure 5: Qualitative comparison on matched Subject 1 stimuli. The first seven columns—Seen image, Brain-IT, MindTuner, MindEye2, NeuroVLA, NeuroPictor, and UMBRAE—are cropped from Figure S2 of the Brain-IT paper (Beliy et al., 2026); the final column contains the corresponding CortexFlow reconstructions. The baseline outputs are reproduced from that published figure rather than regenerated with our implementation, so this panel is intended as a qualitative visual comparison.

D.2 RECONSTRUCTIONS ACROSS FOUR SUBJECTS

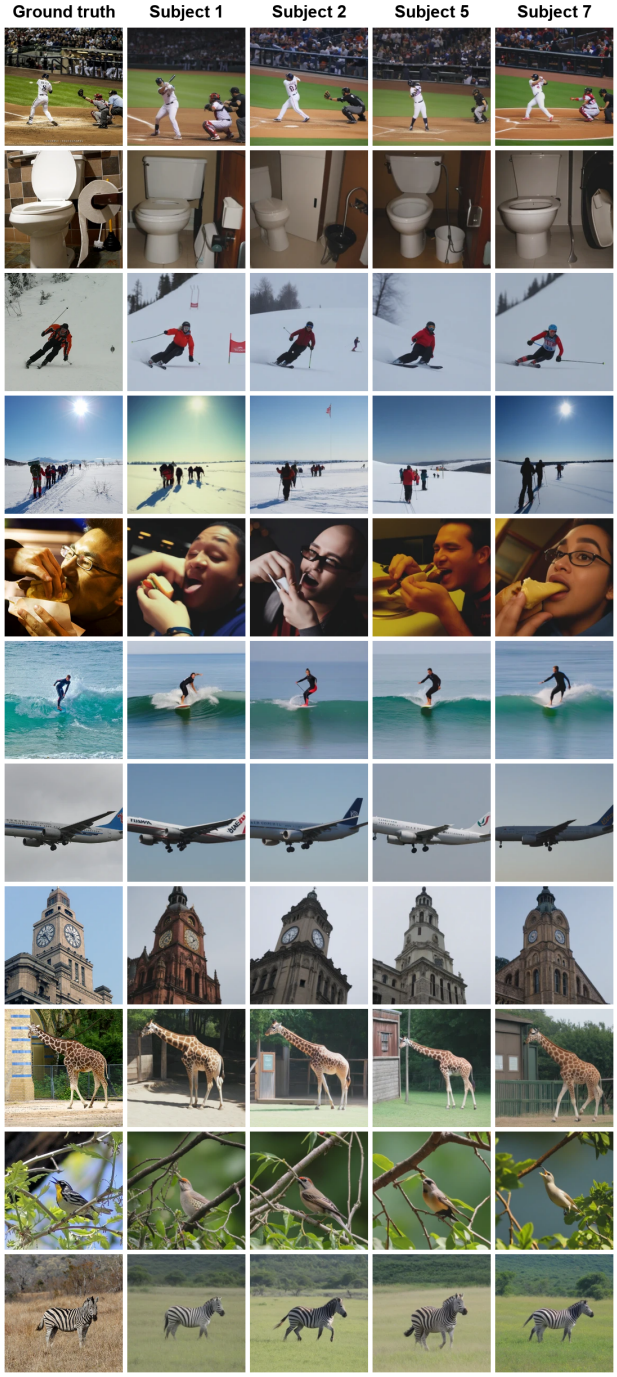


Figure 6: CortexFlow reconstructions for Subjects 1, 2, 5, and 7. Each row uses one shared viewed stimulus: the first column shows the ground-truth image and the remaining columns show the reconstruction obtained from each subject's fMRI measurements. All eleven stimuli with a complete set of four subject reconstructions in the supplied results are shown.

D.3 FAILURE CASES

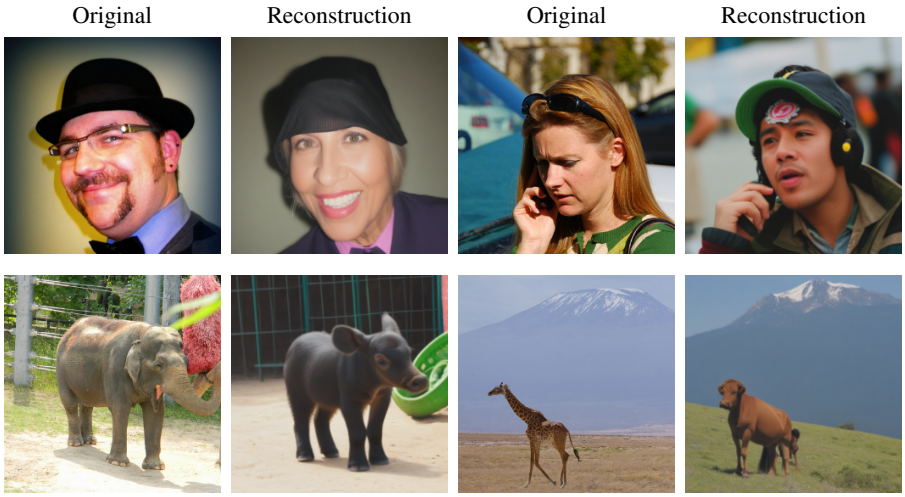


Figure 7: Subject 1 failure cases. Each pair places the viewed stimulus beside its CortexFlow reconstruction. The examples retain portions of the coarse composition while changing identity, attributes, or the central object category.

Figure 7 separates broad visual similarity from stimulus-specific fidelity. Plausible scene structure or color does not necessarily imply faithful recovery of the viewed stimulus or its central object.

D.4 RECONSTRUCTION STAGES

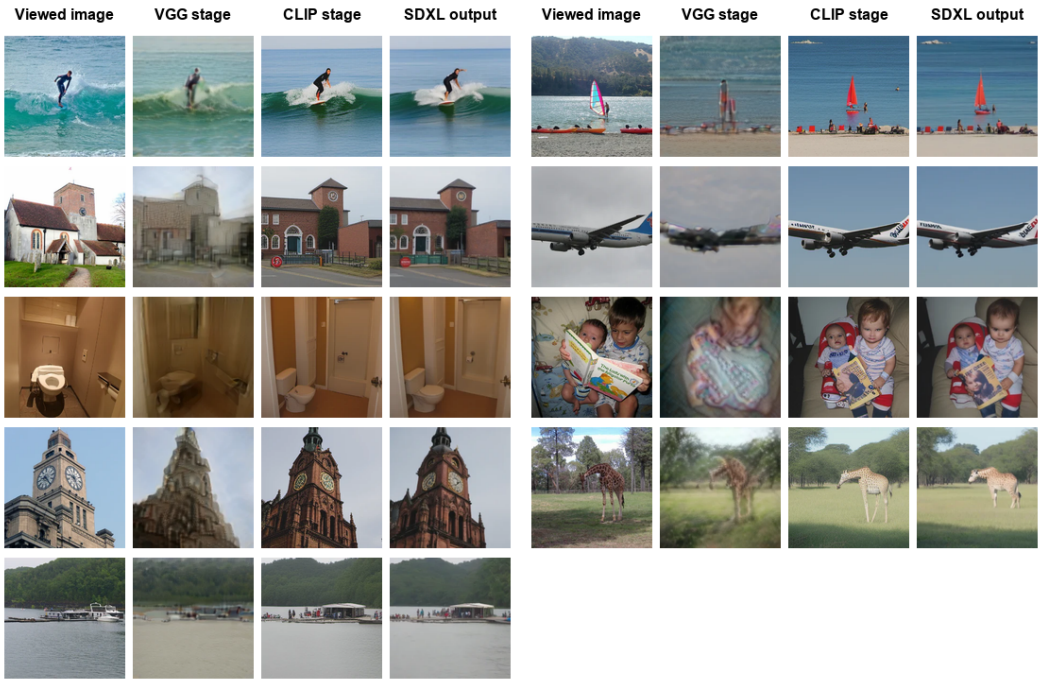


Figure 8: Stage-wise reconstruction examples. Each row corresponds to one viewed image. From left to right, the columns show the viewed image, the VGG-based structural stage, the CLIP-conditioned semantic stage, and the final SDXL output. The aligned rows illustrate how coarse structure is established before semantic and generative refinement.

E ADDITIONAL CORTICAL-REPRESENTATION ANALYSES

This section supplements Figure 4 with analyses of the intermediate cortical representation and branch-specific token updates. Unless stated otherwise, Subject 1 analyses use all 1,000 held-out stimuli. The multi-subject analyses use the same fixed subset of 32 shared stimuli for Subjects 1, 2, 5, and 7. All results describe computations inside the trained model; they are not measurements of neural activation, biological dynamics, or causal regional importance.

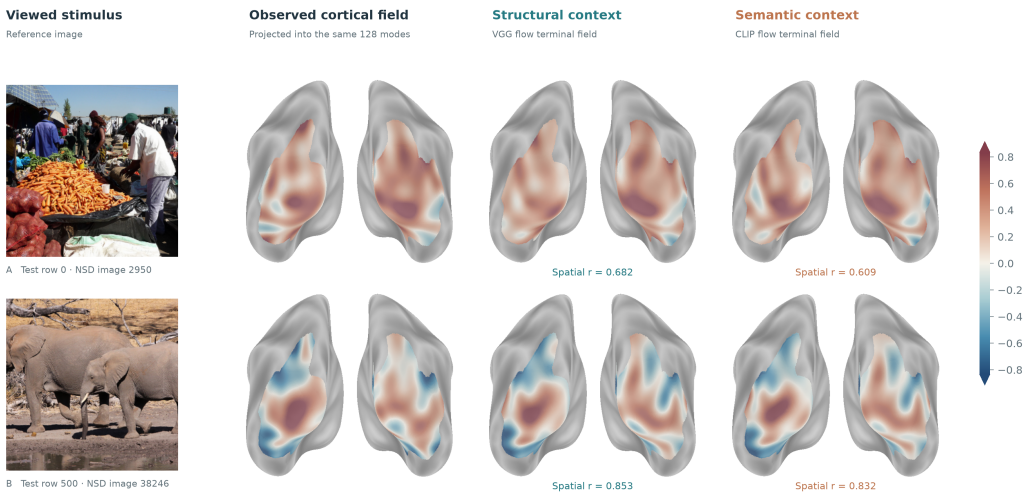


Figure 9: **Observed and generated cortical fields for two fixed test stimuli.** Each row shows a reference image, the measured response reconstructed in the retained 128-mode basis, and terminal fields from the structural and semantic flows. The examples are fixed test rows 0 and 500 for Subject 1. All fields share a symmetric color scale; reported spatial correlations use all modeled vertices and unclipped values. The observed fields are band-limited representations rather than raw activation maps.

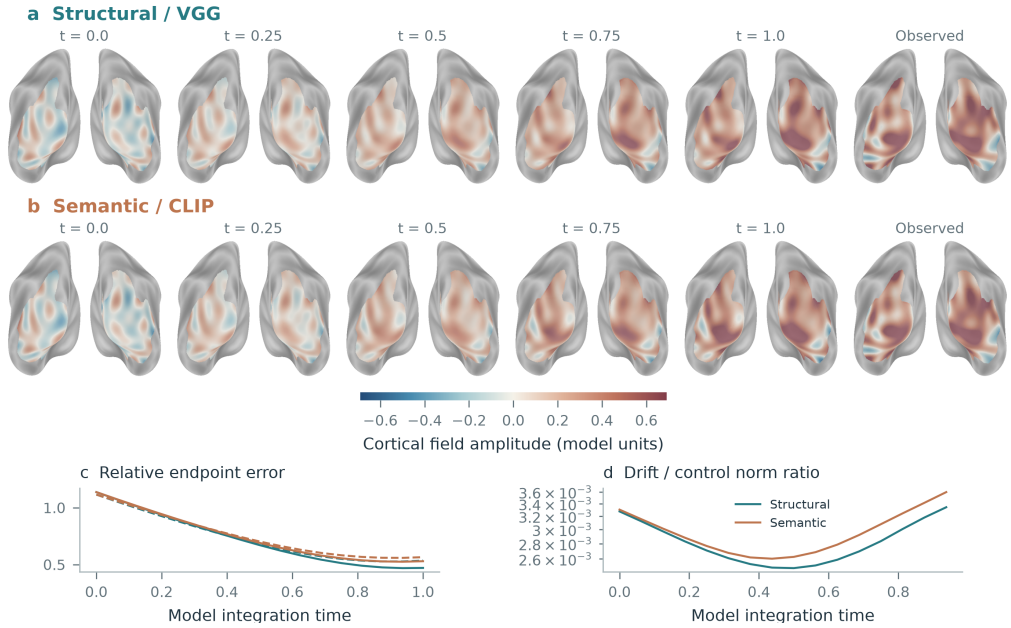


Figure 10: **Evolution of the learned cortical flow.** Structural and semantic fields are shown at selected points along the exact 16-step RK4 trajectory for fixed Subject 1 test row 0, together with the observed field in the same 128-mode basis. Curves report relative distance to the observed endpoint for rows 0 and 500 and the heat-drift-to-learned-control norm ratio for row 0. Integration time is a model coordinate, not neural time or biological dynamics.

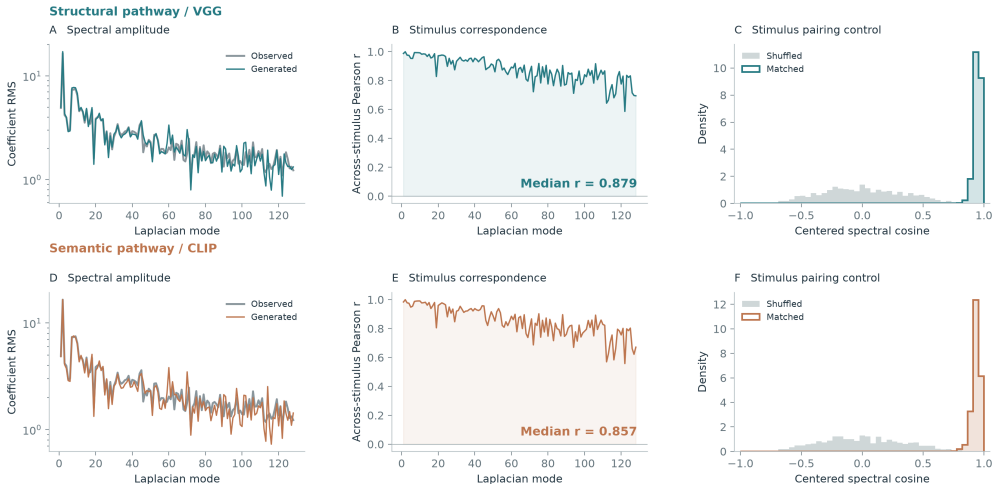


Figure 11: **Agreement between generated and observed cortical spectral representations.** For both branches, we compare coefficient RMS, per-mode Pearson correlation across 1,000 Subject 1 stimuli, and per-stimulus cosine agreement for matched versus fixed shuffled pairs after mode-wise centering. The shuffled distribution is a descriptive reference rather than a permutation significance test; agreement concerns representations derived from the input response.

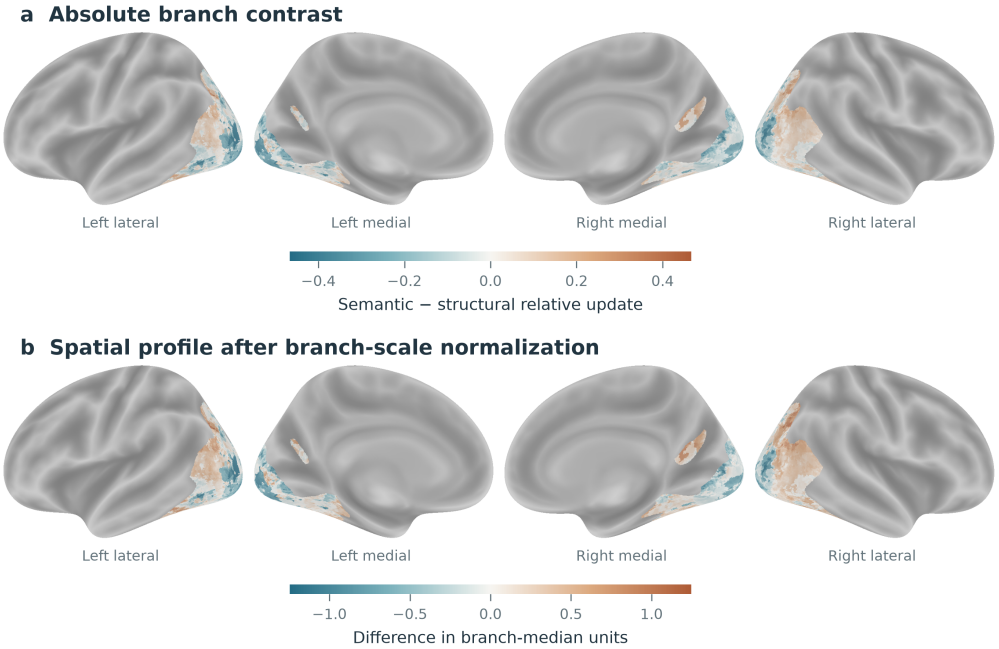


Figure 12: **Spatial contrasts between structural and semantic token updates.** Maps compare stimulus-median projected updates over 1,000 held-out Subject 1 stimuli. The top row shows semantic minus structural update; the bottom row repeats the contrast after normalizing each branch map by its own spatial median over supported vertices. Blue denotes larger structural updates and orange larger semantic updates. These contrasts describe model updates rather than biological functional specialization.

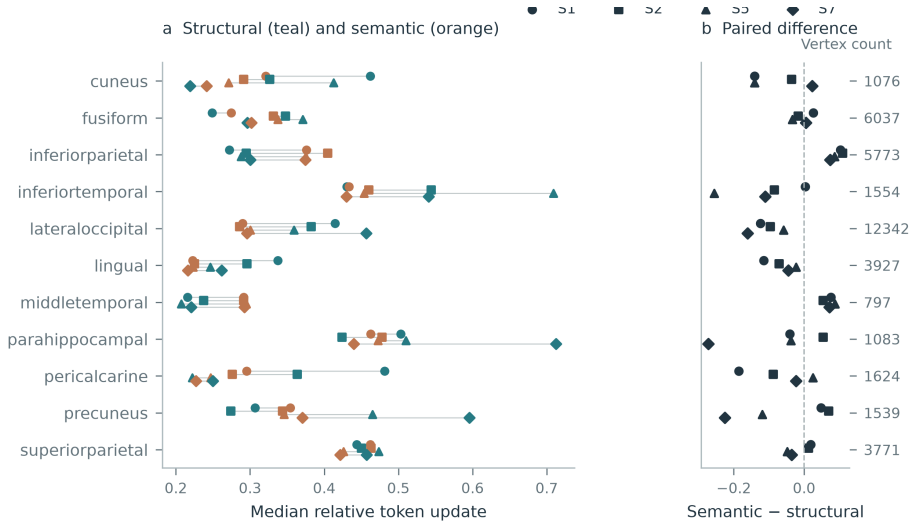


Figure 13: **Anatomical distribution of branch-specific model updates.** Projected relative updates are summarized within bilateral Desikan–Killiany parcels for Subjects 1, 2, 5, and 7 using the same 32 shared test stimuli. Markers denote subjects, lines join the two branches within a subject, and the right panel shows their paired summary difference. Only assignment-supported vertices are included. Anatomical labels provide spatial references rather than evidence of functional specialization.

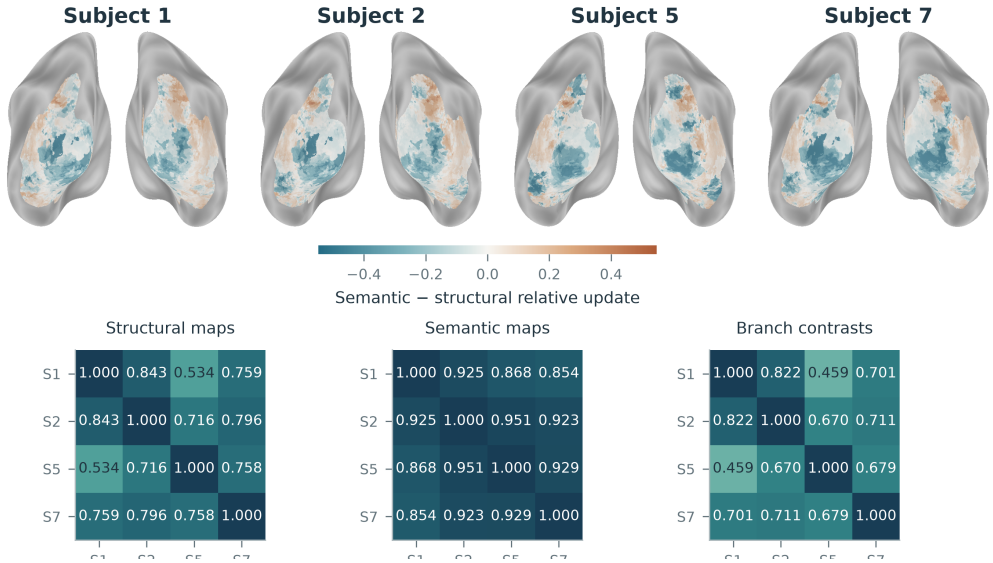


Figure 14: **Descriptive cross-subject consistency of cortical update maps.** The top row shows semantic-minus-structural contrasts for Subjects 1, 2, 5, and 7 on the identical 32-stimulus subset. The bottom row reports pairwise spatial Spearman correlations for structural maps, semantic maps, and branch contrasts over assignment-supported vertices. Shared model parameters and assignments can contribute to agreement; this analysis does not establish population-level or causal biological consistency.

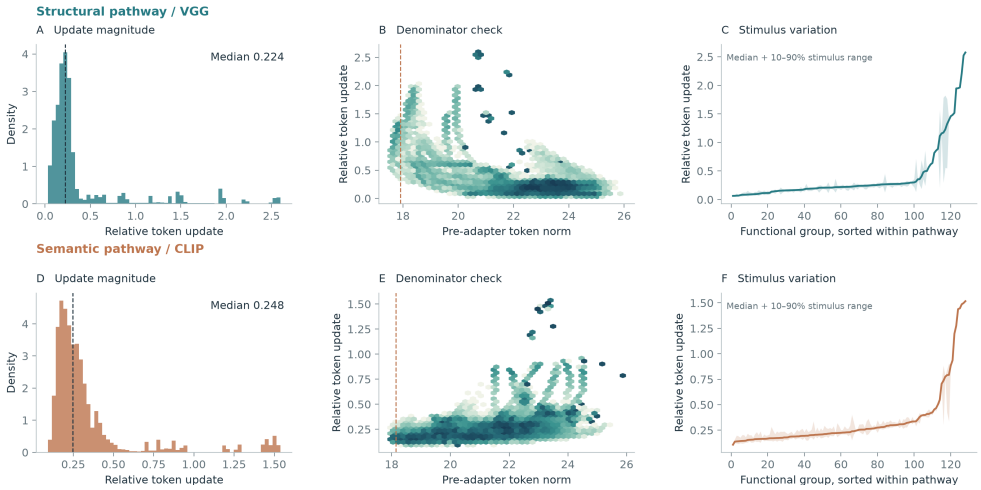


Figure 15: **Magnitude and numerical diagnostics of token updates.** Rows correspond to structural and semantic branches for Subject 1. Distributions summarize 128 functional groups over 1,000 stimuli; hexagonal bins relate relative updates to pre-adapter token norms; and group-wise curves show medians with 10th–90th percentile stimulus ranges. The ranges describe stimulus variability, not population confidence intervals.