SANE: STIMULUS-ADAPTIVE NEURAL EVIDENCE FOR FMRI VISUAL RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

What part of an fMRI response actually supports a decoded visual prediction? Modern fMRI-to-image systems can reconstruct increasingly rich visual content, yet reconstruction accuracy alone does not reveal which neural measurements support a particular prediction. We introduce Stimulus-Adaptive Neural Evidence (SANE), which scores anatomically defined cortical parcels, selects a fixed-budget stimulus-adaptive subset before cross-parcel contextualization, and tests the selected evidence through retention, removal, and cross-stimulus substitution for *sufficiency*, *functional importance*, and *content specificity*, respectively. On the official 1,000-image NSD shared test set across four subjects, a $K = 32$ evidence budget retains only 39.15% of available cortical parcels while preserving 97.89% of dense DINOv3 decoding performance, reducing neural-side computation from 1.073 to 0.445 GFLOPs. Yet this strong compressibility does not imply interchangeability: removing SANE-selected evidence reduces target cosine similarity by 0.1428, compared with 0.0354 for uniform-random removal and 0.0843 for voxel-count-matched removal, while cross-stimulus substitution produces a donor-directed shift of 0.4523, compared with 0.2226 and 0.3869 for the corresponding controls. The selected evidence also shows a reproducible stimulus-dependent component: repeated measurements of the same image yield more similar selected parcel sets than measurements of different images. Finally, the same frozen evidence mechanism remains usable across five heterogeneous pretrained reconstruction systems through lightweight backbone-specific adapters. These results reveal a striking asymmetry in fMRI visual decoding: dense cortical responses are highly redundant for sufficiency, yet the compact evidence identified by SANE remains functionally structured and stimulus-dependent. SANE thus shifts the question from how much neural data can be discarded to which measured neural signals actually support a decoded visual prediction.

1 INTRODUCTION

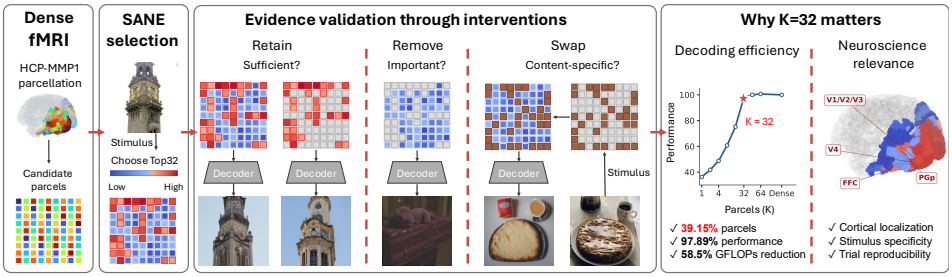


Figure 1: SANE reframes fMRI decoding as neural evidence identification. A compact, stimulus-adaptive subset of anatomically defined parcels preserves decoding performance while exhibiting functional importance and stimulus-specific content under intervention.

Modern fMRI-to-image reconstruction systems can recover increasingly rich visual content from brain activity by combining neural encoders with powerful pretrained vision and generative models (Scotti et al., 2024; Feng et al., 2026; Xia et al., 2026; Mayo et al., 2024). Yet this progress leaves

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.

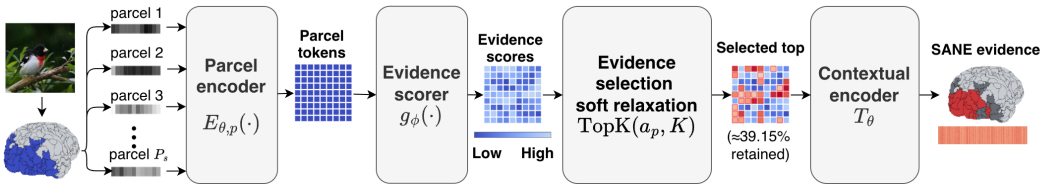


Figure 2: Overview of SANE. Each anatomical parcel is independently encoded and scored from the current fMRI response, and the top-K parcels are selected before cross-parcel contextualization, enabling stimulus-adaptive evidence selection while preserving anatomical identity.

a basic question unresolved: *which measured neural signals actually support a decoded visual prediction?* Reconstruction accuracy alone cannot answer this question. A dense cortical response may contain substantial redundancy, multiple measurements may carry overlapping information, and a powerful downstream model may compensate for missing neural input. Thus, an accurate reconstruction does not reveal which parts of the measured response actually support it.

Compression alone does not resolve this ambiguity. If a small subset of fMRI measurements preserves decoding performance, then that subset is sufficient for the prediction, but it need not be especially important or specific to the current stimulus. Comparable performance could arise because many measurements are interchangeable or because the decoder compensates for missing information. We therefore ask a stronger question: *can the neural evidence supporting a prediction be identified and tested?* We define three complementary properties. *Sufficiency* asks whether selected measurements preserve relevant information when retained. *Functional importance* asks whether removing them from an independently trained dense decoder causes greater degradation than matched controls. *Content specificity* asks whether replacing them with measurements from another stimulus redirects the decoded representation toward that stimulus. These tests distinguish measurements that are merely predictive from measurements whose role can be probed through intervention (Fig. 1).

We introduce Stimulus-Adaptive Neural Evidence (SANE, Fig. 2) to address this problem. SANE uses the HCP-MMP1 atlas (Glasser et al., 2016) to divide the measured visual cortex into fixed anatomical parcels, providing identifiable candidate regions whose selected measurements can be traced to their cortical locations. Each parcel is encoded independently into a local neural token, and a shared scorer assigns an evidence score from that parcel's response. For each fMRI sample, SANE retains a fixed number K of highest-scoring parcels. The atlas defines the candidate units, while SANE determines which units provide evidence for the current stimulus. The central design principle is *identify before integrate*. Parcel evidence is scored and selected before information is propagated across cortical regions. This prevents the evidence score of one parcel from depending on information already mixed in from other parcels, preserving an explicit correspondence between selected evidence and its anatomical source. Only after selection are the retained parcels integrated by a contextual encoder. SANE is trained against a frozen DINOv3 visual representation rather than an image generator, separating evidence identification from the downstream process used to visualize the decoded content. This formulation reveals a striking asymmetry in fMRI visual decoding.

On the official 1,000-image NSD (Allen et al., 2022) shared test set for Subjects 1, 2, 5, and 7, retaining $K = 32$ parcels selects only 39.15% of the available cortical parcels on average while preserving 97.89% of dense repeated-trial DINOv3 decoding performance. At the same time, neural-side computation decreases from 1.073 to 0.445 GFLOPs. Thus, dense fMRI decoding is highly redundant with respect to *sufficiency*: much of the measured response can be discarded without substantially changing the decoded representation. Yet high compressibility does not imply arbitrary interchangeability. Post-hoc intervention on an independently trained dense decoder shows that removing SANE-selected evidence reduces target cosine similarity by 0.1428, compared with 0.0354 for uniform-random removal and 0.0843 for voxel-count-matched removal. Replacing the selected measurements with anatomically corresponding measurements from another stimulus produces a donor-directed shift of 0.4523, compared with 0.2226 for random substitution and 0.3869 for the voxel-count-matched control. Repeated measurements of the same image also produce more similar selected parcel sets than measurements from different images, revealing a reproducible stimulus-dependent component. Thus, the neural response is highly redundant for what is sufficient, yet the compact evidence selected by SANE is not arbitrarily interchangeable.

We further test whether this evidence remains useful beyond the representation and decoder used to learn it. After freezing SANE, we align its $K = 32$ evidence representation to five heterogeneous pretrained reconstruction systems, MindEye2 (Scotti et al., 2024), NeuroPictor (Huo et al., 2024), NeuroAdapter (Feng et al., 2026), NeuroFlow (Mai et al., 2026), and Brain-IT (Beliy et al., 2026), using lightweight, backbone-specific adapters while keeping their downstream generative components frozen. The evidence remains usable across all five systems, although the degree of fidelity preservation depends on the backbone. This controlled experiment tests whether the identified neural evidence remains informative when connected to different reconstruction pipelines, rather than whether SANE universally outperforms their native encoders. Our contributions are threefold:

i. We distinguish predictive sufficiency from functional importance and content specificity, and evaluate these properties through retention, removal, and cross-stimulus substitution.

ii. SANE selects a fixed-budget subset of anatomically defined cortical parcels before cross-parcel contextualization, preserving the anatomical identity of selected measurements while allowing their interactions to be modeled after selection.

iii. Across four NSD subjects, a small subset of cortical parcels preserves nearly all dense decoding performance, yet the selected evidence has disproportionate effects under removal, redirects decoded content under cross-stimulus substitution, and exhibits a reproducible stimulus-dependent component. The same evidence also remains usable after lightweight alignment to heterogeneous reconstruction systems.

A review of closely related work is provided in Appendix A; below, we introduce SANE.

2 METHOD

SANE addresses a simple question: *which parts of an fMRI response provide evidence for the visual content being decoded?* Rather than first mixing information across the entire measured cortex and then attempting to determine which signals matter, SANE treats anatomically defined cortical parcels as candidate evidence units. Each parcel is encoded and scored from its own fMRI response, and a fixed budget K retains the highest-scoring parcels for each response. Only the selected parcels are then combined to model their interactions and predict a visual representation of the stimulus. This *identify-before-integrate* design makes evidence selection stimulus adaptive while keeping each selected unit explicitly tied to its anatomical source. After training, the learned evidence representation is frozen and can be connected to heterogeneous pretrained reconstruction architectures through lightweight, backbone-specific adapters, replacing their native neural interfaces while leaving their downstream reconstruction components unchanged. Below, we introduce SANE and describe how it identifies, integrates, and transfers stimulus-adaptive neural evidence.

2.1 SANE: STIMULUS-ADAPTIVE NEURAL EVIDENCE

For subject s, let $\mathcal{D}_s = \{(\boldsymbol{x}_{s,i}^{(r)}, I_i)\}_{i,r}$ denote paired fMRI responses and visual stimuli, where I_i is stimulus i and $\boldsymbol{x}_{s,i}^{(r)} \in \mathbb{R}^{V_s}$ is the visual-cortex response to I_i at repetition r. Rather than training through an image generator, SANE uses a frozen visual encoder Φ to define the target representation

$$\boldsymbol{y}_i = \Phi(I_i), \tag{1}$$

where Φ denotes the frozen DINOv3 encoder (Siméoni et al., 2025) with ℓ_2 normalization applied to its output, and $\boldsymbol{y}_i \in \mathbb{R}^{d_v}$ is the resulting visual representation of stimulus I_i. This generator-independent target allows SANE to learn which neural measurements support the visual representation without optimizing the downstream image-generation process.

Parcel encoding and evidence scoring. We use the externally defined HCP-MMP1 atlas to partition the measured visual cortex into anatomically identifiable candidate evidence units. For subject s, let P_s denote the number of valid measured parcels, and let $\boldsymbol{x}_{s,i,p}^{(r)} \in \mathbb{R}^{V_{s,p}}$ denote the fMRI response within parcel p for stimulus i and repetition r. Each parcel is encoded independently into a d-dimensional token,

$$\boldsymbol{h}_{s,i,p}^{(r)} = E_{\theta,p}\left(\boldsymbol{x}_{s,i,p}^{(r)}\right) \in \mathbb{R}^d, \tag{2}$$

where $E_{\theta,p}$ denotes the parcel-specific encoder that maps the voxel response of parcel p into the common token space. No information is exchanged across parcels before evidence scoring, so each token and its resulting evidence score depend only on the response within the corresponding anatomical parcel. A shared evidence scorer maps each parcel token to a scalar score,

$$a_{s,i,p}^{(r)} = g_\phi\left(\boldsymbol{h}_{s,i,p}^{(r)}\right) \in \mathbb{R}, \tag{3}$$

where g_ϕ is a shared learnable scoring function applied to every parcel. The score $a_{s,i,p}^{(r)}$ determines the priority of parcel p for selection under the evidence budget. Because the score is computed from each parcel's own token, it can vary across stimuli and repetitions, while the parcel index p continues to identify the same anatomical location. Thus, the HCP-MMP1 atlas specifies the anatomical locations that can provide evidence, while the shared scorer determines which locations receive the highest scores for each fMRI response.

Evidence selection and soft relaxation. The evidence budget K is a prescribed capacity rather than a learned quantity. Given the evidence scores, SANE selects the K highest-scoring parcels,

$$\mathbb{S}_{s,i}^{(r)} = \mathrm{TopK}_{p\in\{1,\dots,P_s\}}\left(a_{s,i,p}^{(r)}, K\right), \qquad \left|\mathbb{S}_{s,i}^{(r)}\right| = K. \tag{4}$$

Thus, K controls *how much* neural evidence is retained, while the learned scores determine *where* that evidence comes from. Different stimuli can therefore select different cortical subsets under the same fixed evidence budget. Hard Top-K selection is non-differentiable, so during training SANE uses a budget-constrained soft relaxation,

$$m_{s,i,p}^{(r)} = \sigma\left(\frac{a_{s,i,p}^{(r)} - \tau_{s,i}^{(r)}}{T}\right), \qquad \sum_{p=1}^{P_s} m_{s,i,p}^{(r)} \simeq K, \tag{5}$$

where $\sigma(z) = 1/(1+\exp(-z))$ is the sigmoid function, $\tau_{s,i}^{(r)}$ is a response-specific threshold chosen by bisection so that the total soft gate mass approximately matches the evidence budget K, and T controls the temperature of the relaxation. The resulting gated token is

$$\widetilde{\boldsymbol{h}}_{s,i,p}^{(r)} = m_{s,i,p}^{(r)} \boldsymbol{h}_{s,i,p}^{(r)}. \tag{6}$$

This relaxation provides a differentiable path from the visual prediction loss through the gates to the evidence scores and their underlying parameters. SANE therefore learns stimulus-dependent evidence scores whose ranking determines which parcels are retained under the fixed budget. The temperature is progressively reduced during training to sharpen the relaxation. At validation and inference, the soft relaxation is removed and the exact K highest-scoring parcels are retained, yielding an explicitly identifiable set of K anatomical evidence units for each response.

Evidence integration and visual prediction. After evidence scores are computed, SANE integrates the parcel-level evidence to predict a visual representation of the stimulus. During training, the evidence scores are converted into soft gates, and the resulting gated tokens $\widetilde{\boldsymbol{h}}_{s,i,p}^{(r)}$ are processed by a Transformer-based contextual encoder T_θ. Specifically, the gated parcel tokens are stacked as

$$\widetilde{\boldsymbol{H}}_{s,i}^{(r)} = \mathrm{Stack}_{p=1}^{P_s}\left(\widetilde{\boldsymbol{h}}_{s,i,p}^{(r)}\right) \in \mathbb{R}^{P_s\times d}, \tag{7}$$

and

$$\boldsymbol{E}_{s,i}^{(r)} = T_\theta\left(\widetilde{\boldsymbol{H}}_{s,i}^{(r)}\right) \in \mathbb{R}^{P_s\times d}. \tag{8}$$

Here, T_θ is a contextual encoder that models interactions among parcel tokens after evidence scoring. Because the tokens are softly gated, gradients from the visual prediction loss can propagate through the contextual encoder to the evidence scores, while preserving the *identify-before-integrate* ordering: each parcel is scored before information is exchanged across parcels.

At validation and inference, the soft relaxation is removed and the exact K highest-scoring parcels are retained. Their original local tokens are gathered as $\boldsymbol{H}_{s,i}^{(r)} = \mathrm{Stack}_{p\in\mathbb{S}_{s,i}^{(r)}}\left(\boldsymbol{h}_{s,i,p}^{(r)}\right) \in \mathbb{R}^{K\times d}$, and processed by the same contextual encoder, $\boldsymbol{E}_{s,i}^{(r)} = T_\theta\left(\boldsymbol{H}_{s,i}^{(r)}\right) \in \mathbb{R}^{K\times d}$. Thus, training uses soft gating for differentiable evidence learning, whereas validation and inference use exactly K selected

parcel tokens. In both cases, the contextual encoder operates only after parcel-level evidence scores have been computed, so it integrates the identified evidence rather than determining which parcels are selected. The contextualized tokens are then mean-pooled and mapped to the frozen visual representation space:

$$\widehat{\boldsymbol{y}}_{s,i}^{(r)} = q_\psi \left(\mathrm{MeanPool} \left(\boldsymbol{E}_{s,i}^{(r)} \right) \right), \tag{9}$$

where q_ψ denotes the prediction head with ℓ_2 normalization applied to its output, and $\widehat{\boldsymbol{y}}_{s,i}^{(r)} \in \mathbb{R}^{d_v}$ is the predicted visual representation. The prediction $\widehat{\boldsymbol{y}}_{s,i}^{(r)}$ and target $\boldsymbol{y}_i$ from Eq. 1 therefore lie in the same normalized DINOv3 representation space. This generator-independent target allows SANE to learn neural evidence that preserves the visual information needed for prediction.

2.2 Learning and Evidence Transfer

SANE is trained in two stages. We first learn the stimulus-adaptive evidence representation against a frozen visual target, using a dense decoder to initialize the shared representation modules. We then freeze the learned SANE evidence pathway and test whether the resulting representation can replace the native neural interface of heterogeneous pretrained reconstruction backbones.

Evidence learning. SANE is trained to predict the frozen visual target $\boldsymbol{y}_i$ from the restricted neural evidence selected under budget K. We minimize the visual prediction loss

$$\mathcal{L}_{\mathrm{vis}} = 1 - \widehat{\boldsymbol{y}}_{s,i}^{(r)\top} \boldsymbol{y}_i, \tag{10}$$

where $\boldsymbol{y}_i$ and $\widehat{\boldsymbol{y}}_{s,i}^{(r)}$ are the ℓ_2-normalized target and prediction defined in Eqs. 1 and 9. The amount of retained neural evidence is controlled directly by the prescribed budget K. During training, gradients from $\mathcal{L}_{\mathrm{vis}}$ propagate through the contextual encoder and soft-gated parcel tokens to the evidence scores $a_{s,i,p}^{(r)}$. This jointly updates the evidence scorer and the parcel encoders, so that the learned scores determine which parcels are retained under the fixed evidence budget.

Dense initialization and training. Before learning stimulus-adaptive evidence, we train a subject-specific dense decoder using all available parcels. The dense decoder shares the parcel encoder, contextual encoder, and prediction head with SANE, but does not perform evidence selection. Its best validation checkpoint initializes these components for SANE, while the evidence scorer is newly initialized. SANE then learns which parcel tokens can preserve the visual prediction under the fixed evidence budget K. For each subject and evidence budget, SANE is trained on a stimulus-disjoint training/validation split. Model selection uses the hard Top-K inference procedure, and the held-out test set is reserved for final evaluation. We further evaluate evidence portability across pretrained reconstruction backbones using lightweight adapters; details are provided in Appendix E.1.

2.3 Testing Neural Evidence Through Interventions

Prediction accuracy alone cannot establish that selected parcels constitute neural evidence, since a compact subset may preserve performance when multiple measurements are interchangeable. We therefore test the learned evidence on held-out stimuli along three complementary dimensions: *sufficiency*, *functional importance*, and *content specificity*. These tests are performed post hoc and are not part of the SANE training objective.

Sufficiency. We retain only the selected parcels and compare their decoding performance with the corresponding dense model. This tests whether the selected evidence preserves the information required for visual prediction under the evidence budget K.

Functional importance. Sufficiency alone does not imply that the selected parcels are functionally distinguished. We therefore remove the selected measurements from an independently trained, frozen dense decoder and measure the resulting loss in target similarity. We compare this degradation with random and voxel-count-matched removal controls (see Appendix D.4) to test whether SANE-selected evidence has a disproportionate functional effect.

Content specificity. We replace the recipient's measurements at the donor's selected parcel locations with the corresponding donor measurements, leaving all other recipient measurements unchanged. We test whether the representation moves away from the recipient target and toward the

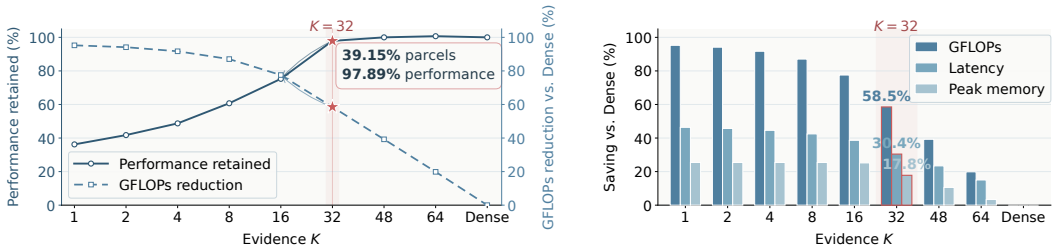
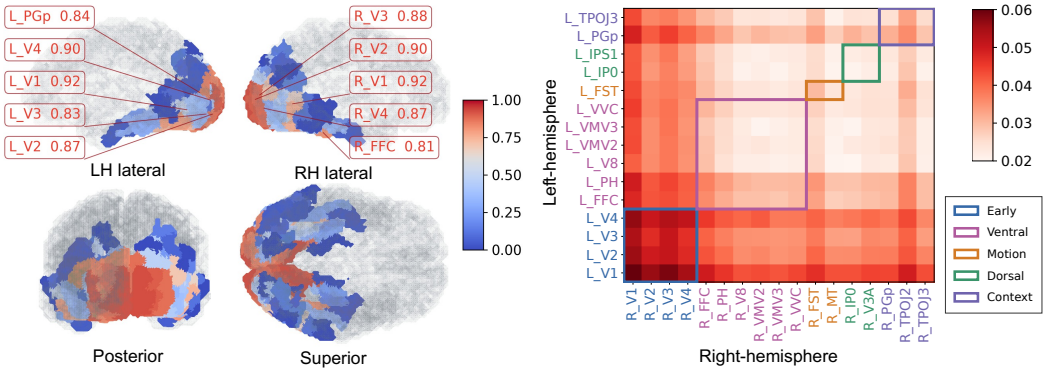


Figure 3: Results are macro-averaged across Subjects 1, 2, 5, and 7. (a) Performance retention and GFLOPs reduction across evidence budgets K. $K = 32$ is the smallest budget that preserves near-Dense performance, retaining 97.89% with only 39.15% of the available parcels; reducing to $K = 16$ causes a substantial performance drop. (b) Reductions in GFLOPs, inference latency, and peak memory relative to the Dense baseline. At $K = 32$, these decrease by 58.5%, 30.4%, and 17.8%, respectively. The red shading and outlines indicate $K = 32$.



(a) Selection frequency across parcels (b) Conditional attention among selected parcels

Figure 4: SANE learns anatomically structured evidence selection and post-selection integration. (a) For each HCP-MMP1 parcel, selection frequency is the fraction of test responses in which it is included in SANE's Top-32 set. Values use a $[0, 1]$ scale, with low and high frequencies in blue and red, respectively; the ten most frequently selected parcels are annotated. (b) The heat map shows conditional post-selection attention in the contextual encoder between left- and right-hemisphere parcels. Each cell is averaged across encoder layers, attention heads, and co-selected trials, with directions symmetrized; darker red indicates stronger attention. The 32 parcels are ordered into five functional groups: early visual, ventral visual, motion, dorsal visual, and contextual. These values reflect model-internal interactions (see Appendix B) rather than biological functional connectivity.

donor target, with random and voxel-count-matched substitutions as controls. This probes whether selected evidence carries stimulus-specific content rather than only generic predictive information. We additionally evaluate alternative substitution strategies in Appendix D.4.3.

3 Experiments

3.1 Experimental Setup

Dataset. We evaluate SANE on NSD Subjects 1, 2, 5, and 7. Each subject has 10,000 images with three single-trial fMRI responses per image. We use the HCP-MMP1 parcellation of visual cortex, with 9,000 subject-specific images for training/validation and the official 1,000 shared images held out for testing. Repeated measurements remain separate trials during evidence learning.

Evaluation. Unless otherwise stated, we use hard Top-K selection with $K = 32$. We evaluate DINOv3 decoding by cosine similarity and image reconstruction using PixCorr, SSIM, two-way identification, and feature-space correlation metrics. Results are macro-averaged across subjects

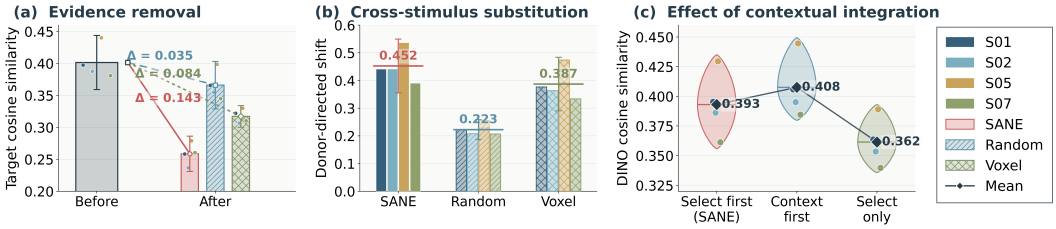


Figure 5: SANE-selected evidence is functionally important, content-specific, and benefits from contextual integration. (a) Removing the 32 SANE-selected parcels causes a larger drop in target cosine similarity ($\Delta = 0.143$) than uniform-random ($\Delta = 0.035$) or voxel-count-matched ($\Delta = 0.084$) removal, indicating disproportionate functional importance. (b) Replacing the selected measurements with those from another stimulus produces a stronger donor-directed shift (0.452) than random (0.223) or voxel-count-matched (0.387) substitution, showing stimulus-specific content. (c) Selecting parcels before cross-parcel integration (Select first) retains most of the benefit of contextualizing all parcels (Context first; 0.393 vs. 0.408), while selecting parcels without any cross-parcel integration (Select only) performs worse (0.362). See Appendix C.1 and D.4 for more details.

Table 1: SANE transfers compact neural evidence across five pretrained reconstruction backbones. We evaluate transfer to MindEye2 (Scotti et al., 2024), NeuroPictor (Huo et al., 2024), NeuroAdapter (Feng et al., 2026), Brain-IT (Beliy et al., 2026), and NeuroFlow (Mai et al., 2026). The same frozen, stimulus-adaptive evidence representation is adapted to each backbone through a lightweight adapter. † denotes our reproduced results; reproduced baselines and SANE variants report mean ± standard deviation over three runs. Results across all five backbones demonstrate the portability of SANE evidence beyond its training decoder.

	Low-Level				High-Level			
Method	**PixCorr ↑**	**SSIM ↑**	**Alex(2) ↑**	**Alex(5) ↑**	**Incep ↑**	**CLIP ↑**	**Eff ↓**	**SwAV ↓**
NeuroPictor (ECCV 2024)	0.229	0.375	96.5%	98.4%	94.5%	93.3%	0.639	0.350
NeuroPictor (ECCV 2024)†	0.228 ± 0.091	0.398 ± 0.020	$96.7 \pm 2.0\%$	$98.4 \pm 1.0\%$	$94.6 \pm 3.3\%$	$93.5 \pm 3.1\%$	0.637 ± 0.055	0.349 ± 0.031
SANE + NeuroPictor	0.157 ± 0.093	0.414 ± 0.022	$83.6 \pm 7.5\%$	$91.3 \pm 4.2\%$	$88.1 \pm 7.0\%$	$87.5 \pm 5.1\%$	0.752 ± 0.060	0.436 ± 0.037
MindEye2 (ICML 2024)	0.322	0.431	96.1%	98.6%	95.4%	93.0%	0.619	0.344
MindEye2 (ICML 2024)†	0.284 ± 0.069	0.373 ± 0.020	$96.5 \pm 2.4\%$	$98.7 \pm 1.0\%$	$95.2 \pm 3.0\%$	$92.9 \pm 3.5\%$	0.625 ± 0.057	0.348 ± 0.032
SANE + MindEye2	0.223 ± 0.069	0.387 ± 0.017	$89.8 \pm 4.9\%$	$95.1 \pm 2.6\%$	$89.7 \pm 5.4\%$	$87.7 \pm 5.1\%$	0.718 ± 0.054	0.408 ± 0.032
NeuroAdapter (ICLR 2026)	0.124	0.307	84.54%	93.48%	90.79%	90.97%	0.716	0.408
NeuroAdapter (ICLR 2026)†	0.080 ± 0.124	0.276 ± 0.035	$77.7 \pm 15.4\%$	$88.1 \pm 9.5\%$	$86.9 \pm 10.0\%$	$86.7 \pm 9.8\%$	0.767 ± 0.079	0.496 ± 0.060
SANE + NeuroAdapter	0.095 ± 0.127	0.300 ± 0.029	$79.1 \pm 14.2\%$	$89.3 \pm 8.0\%$	$87.4 \pm 8.7\%$	$89.0 \pm 7.7\%$	0.737 ± 0.075	0.439 ± 0.051
Brain-IT (ICLR 2026)	0.386	0.486	98.4%	99.5%	97.3%	96.4%	0.564	0.320
Brain-IT (ICLR 2026)†	0.375 ± 0.056	0.476 ± 0.013	$98.7 \pm 0.9\%$	$99.4 \pm 0.4\%$	$97.3 \pm 1.7\%$	$96.3 \pm 2.1\%$	0.565 ± 0.054	0.323 ± 0.030
SANE + Brain-IT	0.307 ± 0.059	0.465 ± 0.013	$96.6 \pm 2.1\%$	$98.3 \pm 1.0\%$	$94.3 \pm 3.2\%$	$93.5 \pm 2.9\%$	0.636 ± 0.055	0.369 ± 0.032
NeuroFlow (CVPR 2026)	-	-	-	-	98.6%	98.7%	0.590	0.341
NeuroFlow (CVPR 2026)†	0.271 ± 0.017	0.210 ± 0.006	$93.4 \pm 2.0\%$	$98.6 \pm 0.6\%$	$98.7 \pm 0.7\%$	$99.2 \pm 0.5\%$	0.583 ± 0.027	0.346 ± 0.012
SANE + NeuroFlow	0.170 ± 0.019	0.204 ± 0.006	$85.2 \pm 3.6\%$	$93.0 \pm 2.0\%$	$90.0 \pm 3.7\%$	$90.0 \pm 3.6\%$	0.742 ± 0.025	0.432 ± 0.012

over three runs (see Appendix C.2). For fair backbone comparisons, native and SANE-driven variants use the same frozen downstream components and evaluation settings.

Implementation. SANE is trained separately for each subject and evidence budget, initialized from the corresponding dense decoder, with model selection based on validation cosine similarity. Detailed preprocessing, optimization, evaluation, statistical, and reproducibility protocols are provided in Appendix C.2. Below, we present our experimental results and analyses, with additional evaluations provided in Appendix D.1, D.2, D.4, and E.1.

3.2 Compactness of Neural Evidence

Parcel-budget ablation. Fig. 3 shows a clear performance-compression trade-off. Performance remains close to Dense as the evidence budget decreases to $K = 32$, where SANE retains 97.89% of Dense performance using only 39.15% of the available parcels. Reducing the budget further to $K = 16$ causes a substantial performance drop. Thus, dense fMRI responses contain substantial redundancy for decoding, but this redundancy becomes limited under more aggressive compression. We therefore use $K = 32$ as the default evidence budget in subsequent experiments.

Computational efficiency. This compact representation also provides substantial computational savings. Relative to Dense, $K = 32$ reduces GFLOPs, inference latency, and peak memory by

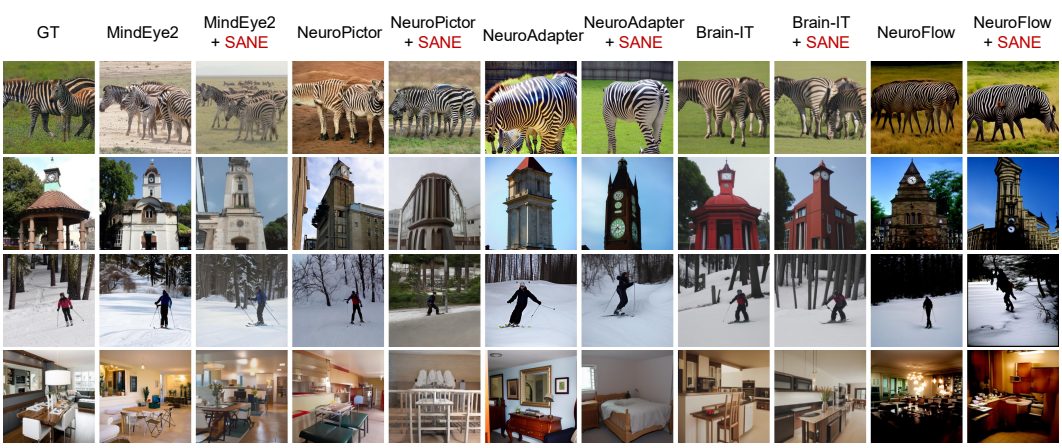


Figure 6: SANE transfers compact neural evidence across reconstruction backbones. Across five pretrained backbones, SANE-driven reconstructions preserve the main content of the ground-truth images, while differences in fine detail reflect backbone-dependent reconstruction fidelity.

58.5%, 30.4%, and 17.8%, respectively. Hence, SANE retains near-Dense decoding performance while substantially reducing the amount of neural evidence that should be processed.

3.3 Structure and Stimulus Adaptivity

Anatomically structured evidence. Fig. 4 shows that SANE does not select parcels uniformly. Selection is concentrated in early visual cortex, while also extending to ventral, motion-sensitive, dorsal, and contextual regions. The selected parcels therefore span a distributed visual hierarchy rather than a single cortical subsystem. After selection, the contextual encoder exhibits structured interactions among these parcels, indicating that SANE first identifies anatomically localized evidence and then models their interactions.

Stimulus-dependent selection. The selected evidence also contains a reproducible stimulus-dependent component. As detailed in Appendix D.2.3, repeated measurements of the same image produce more similar Top-32 sets than different images, with mean Jaccard similarity of 0.464 versus 0.427 and mean overlap of 20.2 versus 19.0 parcels. Moreover, 78.8% of stimuli show higher within-stimulus than cross-stimulus Jaccard similarity. This indicates that selection is neither fixed across stimuli nor dominated by trial-to-trial variability.

Adaptivity under limited evidence. Appendix D.2.4 further shows that stimulus adaptivity matters most when the evidence budget is tight. Relative to a static training-derived selection, dynamic selection improves DINO cosine by 0.068 at $K = 4$ and 0.038 at $K = 8$, but the difference falls to 0.002 at $K = 16$ and becomes negligible for larger budgets. Thus, SANE's stimulus adaptivity is most consequential when only a small number of neural measurements can be retained.

3.4 Functional and Content-Specific Evidence

Functional importance. Retaining only the selected $K = 32$ parcels preserves near-Dense decoding performance (Fig. 3(a)), but this sufficiency does not imply that the selected parcels are interchangeable. Removing SANE-selected evidence from the independently trained dense decoder reduces target cosine similarity by 0.143, compared with 0.035 for uniform-random removal and 0.084 for voxel-count-matched removal. The selected evidence therefore has a disproportionate effect on the decoded representation beyond the amount of neural input removed.

Content specificity. Replacing the selected measurements with those from another stimulus produces a donor-directed shift of 0.452, compared with 0.223 for random and 0.387 for voxel-count-matched substitution (Fig. 5(b)). Thus, the selected measurements do not merely support generic decoding; changing them redirects the decoded representation toward the donor stimulus.



Figure 7: Qualitative interventions reveal stimulus-specific neural evidence. The first two rows use Brain-IT and the last two use MindEye2. SANE retention preserves the main content of Dense reconstructions, removal disrupts the decoded content, and cross-stimulus substitution redirects the reconstruction toward the donor stimulus. Blue and orange borders identify the two stimuli, with arrows indicating swap direction. See Appendix E.2.8 for additional visualizations.

Identify before integrate. The contextual-integration ablation (Fig. 5(c)) explains the role of SANE's selection order. Selecting evidence before contextualization achieves a DINOv3 cosine similarity of 0.393, close to contextualizing all parcels before selection (0.408), whereas selecting without cross-parcel integration drops to 0.362. SANE preserves most of the predictive benefit of contextual integration while keeping each selected token explicitly tied to its anatomical source.

3.5 Transferability of Neural Evidence

Cross-backbone transfer. For each backbone, only a lightweight adapter maps SANE evidence to its native conditioning space, while the downstream reconstruction pipeline remains frozen. Table 1 shows that SANE-driven variants retain meaningful reconstruction performance across all five backbones despite their heterogeneous conditioning interfaces. Fig. 6 compares native and SANE-driven reconstructions. SANE-driven reconstructions preserve the dominant visual content across all five backbones, while fine-grained fidelity varies by architecture.

Backbone-dependent fidelity. The effect of SANE evidence varies across reconstruction backbones. NeuroAdapter is the only case where SANE improves all reported reconstruction metrics. For MindEye2 and NeuroPictor, SANE improves SSIM while the remaining metrics generally decrease. Brain-IT and NeuroFlow retain high identification accuracy but show lower fidelity on most reconstruction metrics. These results show that the portability of compact neural evidence depends on the conditioning interface and downstream representation of each backbone.

Intervention consistency. Fig. 7 further shows qualitative interventions using Brain-IT and MindEye2. Retaining the selected $K = 32$ parcels preserves the main content of the Dense reconstruction, whereas removing them disrupts the decoded content. Replacing the selected measurements with those from a paired stimulus redirects the reconstruction toward the donor stimulus, consistent with the quantitative removal and substitution results.

4 Conclusion

We introduced SANE, a framework for identifying compact, stimulus-adaptive neural evidence from anatomically defined fMRI measurements before cross-parcel integration. Rather than treating compression as sufficient, SANE tests selected evidence through retention, removal, and cross-stimulus substitution, revealing a key asymmetry in fMRI decoding: dense cortical responses are highly redundant for sufficiency, yet the compact evidence identified by SANE remains functionally structured and stimulus-dependent.

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, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. *Advances in Neural Information Processing Systems*, 32, 2019.

Roman Beliy, Amit Zalcher, Jonathan Kogman, et al. Brain-it: Image reconstruction from fmri via brain-interaction transformer. In *International Conference on Learning Representations*, volume 2026, pp. 124487–124521, 2026.

Abdulkadir Canatar, Jenelle Feather, Albert Wakhloo, and SueYeon Chung. A spectral theory of neural prediction and alignment. *Advances in Neural Information Processing Systems*, 36:47052–47080, 2023.

Marlène Careil, Yohann Benchetrit, and Jean-Rémi King. Dynadiff: Single-stage decoding of images from continuously evolving fmri. *arXiv preprint arXiv:2505.14556*, 2025.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Minkyu Choi, Kuan Han, Xiaokai Wang, Yizhen Zhang, and Zhongming Liu. A dual-stream neural network explains the functional segregation of dorsal and ventral visual pathways in human brains. *arXiv preprint arXiv:2310.13849*, 2023.

Yuqin Dai, Zhouheng Yao, Chunfeng Song, Qihao Zheng, Weijian Mai, Kunyu Peng, Shuai Lu, Wanli Ouyang, Jian Yang, and Jiamin Wu. Mindaligner: Explicit brain functional alignment for cross-subject visual decoding from limited fmri data. *arXiv preprint arXiv:2502.05034*, 2025.

Pinyuan Feng, Hossein Adeli, Wenxuan Guo, Fan Cheng, Ethan Hwang, and Nikolaus Kriegeskorte. Towards interpretable visual decoding with attention to brain representations. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 42867–42894, 2026.

Matteo Ferrante, Tommaso Boccato, and Nicola Toschi. Semantic brain decoding: from fmri to conceptually similar image reconstruction of visual stimuli. *arXiv preprint arXiv:2212.06726*, 2022.

Matteo Ferrante, Tommaso Boccato, Furkan Ozcelik, Rufin VanRullen, and Nicola Toschi. Multimodal decoding of human brain activity into images and text. In *UniReps: the First Workshop on Unifying Representations in Neural Models*, 2023.

Matteo Ferrante, Tommaso Boccato, Furkan Ozcelik, Rufin VanRullen, and Nicola Toschi. Through their eyes: multi-subject brain decoding with simple alignment techniques. *Imaging Neuroscience*, 2:imag–2, 2024.

Matteo Ferrante, Tommaso Boccato, Nicola Toschi, and Rufin VanRullen. Evidence for compositionality in fmri visual representations via brain algebra. *Communications Biology*, 8(1):1263, 2025.

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, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

Zixuan Gong, Qi Zhang, Guangyin Bao, Lei Zhu, Rongtao Xu, Ke Liu, Liang Hu, and Duoqian Miao. Mindtuner: Cross-subject visual decoding with visual fingerprint and semantic correction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 14247–14255, 2025.

Umut Güçlü and Marcel AJ Van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *The Journal of neuroscience*, 35(27):10005–10014, 2015.

Yağmur Güçlütürk, Umut Güçlü, Katja Seeliger, Sander Bosch, Rob van Lier, and Marcel A van Gerven. Reconstructing perceived faces from brain activations with deep adversarial neural decoding. *Advances in neural information processing systems*, 30, 2017.

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

Wanqi He, Jin Wang, Hui Li, Hanyang Chi, and Bingfeng Zhang. Balancing semantic and structural decoding for fmri-to-image reconstruction. *Expert Systems with Applications*, pp. 129836, 2025.

Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):15037, 2017.

Liwei Huang, Zhengyu Ma, Liutao Yu, Huihui Zhou, and Yonghong Tian. Long-range feedback spiking network captures dynamic and static representations of the visual cortex under movie stimuli. *Advances in Neural Information Processing Systems*, 37:11432–11455, 2024.

Zheng Huang, Enpei Zhang, Weikang Qiu, Yinghao Cai, Carl Yang, Elynn Chen, Xiang Zhang, Rex Ying, Dawei Zhou, and Yujun Yan. Seeing through the brain: New insights from decoding visual stimuli with fmri. In *International Conference on Learning Representations*, volume 2026, pp. 88145–88169, 2026.

Jingyang Huo, Yikai Wang, Yun Wang, Xuelin Qian, Chong Li, Yanwei Fu, and Jianfeng Feng. Neuropictor: Refining fmri-to-image reconstruction via multi-individual pretraining and multi-level modulation. In *European Conference on Computer Vision*, pp. 56–73. Springer, 2024.

Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012.

Shuai Jiang, Zhu Meng, Haiwen Li, Delong Liu, Fei Su, and Zhicheng Zhao. Mindshot: A few-shot brain decoding framework via transferring cross-subject prior and distilling frequency domain knowledge. *Knowledge-Based Systems*, pp. 114867, 2025.

Haodong Jing, Dongyao Jiang, Yongqiang Ma, Haibo Hua, Bo Huang, and Nanning Zheng. Beyond brain decoding: Visual-semantic reconstructions to mental creation extension based on fmri. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19258–19268. IEEE, 2025.

Jaehoon Joo, Taejin Jeong, and Seongjae Hwang. Brain-streams: fmri-to-image reconstruction with multi-modal guidance. *arXiv preprint arXiv:2409.12099*, 2024.

Fatemeh Kalantari, Karim Faez, Hamidreza Amindavar, and Soheila Nazari. Improved image reconstruction from brain activity through automatic image captioning. *Scientific Reports*, 15(1): 4907, 2025a.

Fatemeh Kalantari, Karim Faez, Hamidreza Amindavar, and Soheila Nazari. A novel approach for visual and semantic image reconstruction from brain activity. *Array*, pp. 100556, 2025b.

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

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.

Reese Kneeland, Paul S Scotti, Ghislain St-Yves, Jesse Breedlove, Kendrick Kay, and Thomas Naselaris. Nsd-imagery: A benchmark dataset for extending fmri vision decoding methods to mental imagery. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28852–28862. IEEE, 2025.

Reese Kneeland, Cesar Kadir Torrico Villanueva, Tong Chen, Jordyn Ojeda, Shubh Khanna, Jonathan Xu, Paul S Scotti, and Thomas Naselaris. Mirage: Robust multi-modal architectures translate fmri-to-image models from vision to mental imagery. *PLoS computational biology*, 22 (5):e1014263, 2026.

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

Mei Kuang, Zongyi Zhan, and Shaobing Gao. Natural image reconstruction from fmri based on node–edge interaction and multi–scale constraint. *Brain Sciences*, 14(3):234, 2024.

Hongmi Lee and Brice A Kuhl. Reconstructing perceived and retrieved faces from activity patterns in lateral parietal cortex. *The Journal of Neuroscience*, 36(22):6069–6082, 2016.

Haoyu Li, Hao Wu, and Badong Chen. Neuraldiffuser: Neuroscience-inspired diffusion guidance for fmri visual reconstruction. *IEEE Transactions on Image Processing*, 34:552–565, 2025.

Jinke Li, Yuxiao Yang, Yanyan Huang, Kaiqiang Xu, Yannan Chen, Lequan Yu, Zhijun Yao, and Yu Fu. Reconstructing shared visual experiences from human brain activity across individuals. *Medical Image Analysis*, pp. 104157, 2026a.

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, 2026b.

Drew Linsley, Ivan F Rodriguez Rodriguez, Thomas Fel, Michael Arcaro, Saloni Sharma, Margaret Livingstone, and Thomas Serre. Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex. *Advances in Neural Information Processing Systems*, 36:28873–28891, 2023.

Yulong Liu, Yongqiang Ma, Guibo Zhu, Haodong Jing, and Nanning Zheng. See through their minds: Learning transferable brain decoding models from cross-subject fmri. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5730–5738, 2025.

Yizhuo Lu, Changde Du, Qiongyi Zhou, Dianpeng Wang, and Huiguang He. Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 5899–5908, 2023.

Yizhuo Lu, Changde Du, Qingyu Shi, Hang Chen, Jie Peng, Liuyun Jiang, Shuangchen Zhao, and Huiguang He. Mind-omni: A unified multi-task framework for brain-vision-language modeling via discrete diffusion. *arXiv preprint arXiv:2605.29591*, 2026.

Yongqiang Ma, Yulong Liu, Liangjun Chen, Guibo Zhu, Badong Chen, and Nanning Zheng. Brainclip: Brain representation via clip for generic natural visual stimulus decoding. *IEEE Transactions on Medical Imaging*, 44(10):3962–3972, 2025.

Weijian Mai, Mu Nan, Yu Zhu, Jiahang Cao, Rui Zhang, Yuqin Dai, Chunfeng Song, Andrew Luo, and Jiamin Wu. Neuroflow: Toward unified visual encoding and decoding from neural activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12966–12976, 2026.

David Mayo, Christopher Wang, Asa Harbin, Abdulrahman Alabdulkareem, Albert Shaw, Boris Katz, and Andrei Barbu. Brainbits: how much of the brain are generative reconstruction methods using? *Advances in Neural Information Processing Systems*, 37:54396–54420, 2024.

Lu Meng and Chuanhao Yang. Dual-guided brain diffusion model: Natural image reconstruction from human visual stimulus fmri. *Bioengineering*, 10(10):1117, 2023.

Lu Meng and Chuanhao Yang. Semantics-guided hierarchical feature encoding generative adversarial network for visual image reconstruction from brain activity. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:1267–1283, 2024.

Lu Meng, Zhenxuan Tang, and Yangqian Liu. Reconstruction of natural images from human fmri using a three-stage multi-level deep fusion model. *Journal of Neuroscience Methods*, 411:110269, 2024.

Lu Meng, Zengyu Song, and Jiayang Lu. Brain-imager: a multimodal framework for image reconstruction and captioning from human brain activity. *Brain Informatics*, 12(1):32, 2025.

Eleni Miliotou, Panagiotis Kyriakis, Jason D Hinman, Andrei Irimia, and Paul Bogdan. Generative decoding of visual stimuli. In *International Conference on Machine Learning*, pp. 24775–24784. PMLR, 2023.

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

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.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

Nina Nellen, Polina Turishcheva, Michaela Vystrčilová, Shashwat Sridhar, Tim Gollisch, Andreas Tolias, and Alexander Ecker. Learning to cluster neuronal function. *Advances in Neural Information Processing Systems*, 38:137863–137890, 2026.

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(1):15666, 2023.

Furkan Ozcelik, Bhavin Choksi, Milad Mozafari, Leila Reddy, and Rufin VanRullen. Reconstruction of perceived images from fmri patterns and semantic brain exploration using instance-conditioned gans. In *2022 international joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE, 2022.

Galen Pogoncheff, Jacob Granley, and Michael Beyeler. Explaining v1 properties with a biologically constrained deep learning architecture. *Advances in Neural Information Processing Systems*, 36: 13908–13930, 2023.

Xuelin Qian, Yikai Wang, Xinwei Sun, Yanwei Fu, Xiangyang Xue, and Jianfeng Feng. Lea: Learning latent embedding alignment model for fmri decoding and encoding. *Transactions on Machine Learning Research*, 2024.

Xuelin Qian, Yun Wang, Yikai Wang, Shijie Zhao, Jianfeng Feng, and Yanwei Fu. Cross-subject fmri-to-image with visual-cortex 2d representation and pre-training. *IEEE Journal of Biomedical and Health Informatics*, 2026.

Ruijie Quan, Wenguan Wang, Zhibo Tian, Fan Ma, and Yi Yang. Psychometry: An omnifit model for image reconstruction from human brain activity. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 233–243. IEEE, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Zarina Rakhimberdina, Quentin Jodelet, Xin Liu, and Tsuyoshi Murata. Natural image reconstruction from fmri using deep learning: A survey. *Frontiers in neuroscience*, 15:795488, 2021.

Haonan Ren and Mengxing Huang. Decoding visual neural representations by fmri-based brain activity through contrastive learning. *Biomedical Signal Processing and Control*, 123:110596, 2026.

Gabriel H Sarch, Michael J Tarr, Katerina Fragkiadaki, and Leila Wehbe. Brain dissection: fmri-trained networks reveal spatial selectivity in the processing of natural images. *bioRxiv*, pp. 2023–05, 2023.

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36:24705–24728, 2023.

Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. *arXiv preprint arXiv:2403.11207*, 2024.

Katja Seeliger, Umut Güçlü, Luca Ambrogioni, Yagmur Güçlütürk, and Marcel AJ Van Gerven. Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, 181:775–785, 2018.

Guohua Shen, Kshitij Dwivedi, Kei Majima, Tomoyasu Horikawa, and Yukiyasu Kamitani. End-to-end deep image reconstruction from human brain activity. *Frontiers in computational neuroscience*, 13:21, 2019a.

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

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Ghislain St-Yves and Thomas Naselaris. Generative adversarial networks conditioned on brain activity reconstruct seen images. In *2018 IEEE international conference on systems, man, and cybernetics (SMC)*, pp. 1054–1061. IEEE, 2018.

Vighnesh Subramaniam, Colin Conwell, Christopher Wang, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Revealing vision-language integration in the brain with multimodal networks. *ArXiv*, pp. arXiv–2406, 2024.

Guoying Sun, Weiyu Guo, Tong Shao, Yang Yang, Haijin Zeng, Jie Liu, and Jingyong Su. Braincognizer: Brain decoding with human visual cognition simulation for fmri-to-image reconstruction. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2831–2838. IEEE, 2025.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

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

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PmLR, 2019.

Zhibo Tian, Ruijie Quan, Fan Ma, Kun Zhan, and Yi Yang. Brainguard: Privacy-preserving multisubject image reconstructions from brain activities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14414–14422, 2025.

Polina Turishcheva, Max F Burg, Fabian H Sinz, and Alexander S Ecker. Reproducibility of predictive networks for mouse visual cortex. *Advances in Neural Information Processing Systems*, 37: 7930–7956, 2024.

Haonan Wang, Qixiang Zhang, Lehan Wang, Xuanqi Huang, and Xiaomeng Li. Neurons: Emulating the human visual cortex improves fidelity and interpretability in fmri-to-video reconstruction. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18367–18376. IEEE, 2025a.

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

Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Mindbridge: A cross-subject brain decoding framework. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11333–11342. IEEE, 2024.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Ziyu Wang, Tengyu Pan, Zhenyu Li, Ji Wu, Xiuxing Li, and Jianyong Wang. Troi: Cross-subject pretraining with sparse voxel selection for enhanced fmri visual decoding. *arXiv preprint arXiv:2502.00412*, 2025b.

Haitao Wu, Qirui Zhang, Zhouheng Yao, Shangquan Sun, Qihao Zheng, Mianxin Liu, Chi Zhang, Wanli Ouyang, Chunfeng Song, Changqing Zhang, et al. Brainjanus: A unified model for understanding and generation across brain, vision, and language. *arXiv preprint arXiv:2606.30319*, 2026.

Tian Xia, Zihan Ma, Yawen Zhang, Xinlong Wang, Qing Liu, Xiaowei He, and Yudan Ren. Brainmclip: Brain image decoding with multi-layer feature fusion of clip. *Meta-Radiology*, pp. 100219, 2026.

Weihao Xia and Cengiz Oztireli. Multigranular evaluation for brain visual decoding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17868–17876, 2026.

Weihao Xia, Raoul De Charette, Cengiz Oztireli, and Jing-Hao Xue. Dream: Visual decoding from reversing human visual system. In *2024 IEEE/CVF winter conference on applications of computer vision (WACV)*, pp. 8211–8220. IEEE, 2024a.

Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. Umbrae: Unified multimodal brain decoding. In *European Conference on Computer Vision*, pp. 242–259. Springer, 2024b.

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

Lingxiao Yang, Hui Zhen, Le Li, Yuanning Li, Han Zhang, Xiaohua Xie, and Ru-Yuan Zhang. Functional diversity of visual cortex improves constraint-free natural image reconstruction from human brain activity. *Fundamental Research*, 5(6):2639–2648, 2025.

Junliang Ye, Lei Wang, and Md Zakir Hossain. Freqselect: frequency-aware fmri-to-image reconstruction. *arXiv preprint arXiv:2505.12552*, 2025.

Xu Yin, Zhengping Wu, and Haixian Wang. A novel drl-guided sparse voxel decoding model for reconstructing perceived images from brain activity. *Journal of Neuroscience Methods*, 412: 110292, 2024.

Muquan Yu, Mu Nan, Hossein Adeli, Jacob Prince, John A Pyles, Leila Wehbe, Maggie Henderson, Michael Tarr, and Andrew Luo. Meta-learning an in-context transformer model of human higher visual cortex. *Advances in Neural Information Processing Systems*, 38:168122–168180, 2026.

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 *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 6248–6257. IEEE, 2025.

Bohan Zeng, Shanglin Li, Xuhui Liu, Sicheng Gao, Xiaolong Jiang, Xu Tang, Yao Hu, Jianzhuang Liu, and Baochang Zhang. Controllable mind visual diffusion model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 6935–6943, 2024.

Guowei Zhang, Yun Zhao, Kai Sun, Moein Khajehnejad, Adeel Razi, Dinh Phung, and Levin Kuhlmann. Hi-dream: Brain-inspired hierarchical diffusion for fmri-to-image reconstruction via roi encoder and visual mapping. *arXiv preprint arXiv:2511.11437*, 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*, volume 2026, pp. 68490–68515, 2026.

Yu Zhu, Chunfeng Song, Wanli Ouyang, Shan Yu, and Tiejun Huang. Neural representational consistency emerges from probabilistic neural-behavioral representation alignment. *arXiv preprint arXiv:2505.04331*, 2025.

APPENDIX CONTENTS

A Related Work **18**

B Anatomical Interpretation of SANE-Selected Evidence **19**

B.1 HCP-MMP1 as the Anatomical Basis of SANE 19

B.2 Group-Level Selection Frequency and Top-32 Ranking 19

B.3 Functional Organization of Frequently Selected Evidence 21

C Implementation and Experimental Details **21**

C.1 SANE Architecture and Optimization Details 21

C.2 Full Experimental Protocol 24

D Evidence Selection and Intervention Analyses **26**

D.1 Evidence Selection Analyses and Budget Ablations 26

D.2 Validating Adaptive Evidence Selection 29

D.3 Stimulus-Specific Evidence Selection and Interaction Patterns 35

D.4 Intervention Definitions, Controls, Results, and Interpretation 35

E Backbone Transfer and Reconstruction Analyses **43**

E.1 Backbone-Agnostic Evidence Transfer 43

E.2 Backbone-Specific Alignment and Reconstruction Details 44

E.3 Backbone-Dependent Effects of SANE Interface Replacement 49

E.4 Neural-Side Efficiency Across Reconstruction Backbones 60

A Related Work

fMRI visual decoding and generative reconstruction. Early work established that distributed fMRI activity can be mapped to visual content using explicit encoding models and pretrained visual representations (Shen et al., 2019a;b; Miyawaki et al., 2008; Kay et al., 2008; Naselaris et al., 2009; Nishimoto et al., 2011; Horikawa & Kamitani, 2017). The emergence of large vision models and generative priors substantially improved reconstruction fidelity (Güçlütürk et al., 2017; Seeliger et al., 2018; St-Yves & Naselaris, 2018; Beliy et al., 2019; Ozcelik et al., 2022; Jing et al., 2025). Brain-Diffuser (Ozcelik & VanRullen, 2023) combines low-level VDVAE features with CLIP-based semantic guidance and latent diffusion, while Takagi and Nishimoto (Takagi & Nishimoto, 2023) map fMRI into Stable Diffusion latent and conditioning spaces. MindEye (Scotti et al., 2023) combines contrastive retrieval with a diffusion prior, and DREAM (Xia et al., 2024a) separates semantic, color, and depth information. More recent systems further enrich the neural-visual interface through multi-individual pretraining and multi-level conditioning (Huo et al., 2024; Quan et al., 2024; Li et al., 2026b; Jiang et al., 2025; Qian et al., 2026), functionally specialized visual representations (Xia et al., 2026; Yang et al., 2025), hierarchical autoregressive reconstruction (Zhang et al., 2026; Li et al., 2025), shared neural-visual latent spaces (Mai et al., 2026; Meng et al., 2025; He et al., 2025; Kalantari et al., 2025b), functional neural clusters (Beliy et al., 2026; Yu et al., 2025; Sun et al., 2025; Wang et al., 2025a), and direct brain-token conditioning (Feng et al., 2026; Lu et al., 2026; Wu et al., 2026; Li et al., 2026a).

These methods primarily focus on improving or extending *what can be decoded or reconstructed* from fMRI (Joo et al., 2024; Dai et al., 2025; Liu et al., 2025; Ferrante et al., 2024; Xu et al., 2025; Wang et al., 2026; Ferrante et al., 2023; Meng et al., 2024; Kuang et al., 2024; Kalantari et al., 2025a; Kneeland et al., 2026; Careil et al., 2025; Zhang et al., 2025; Huang et al., 2026; Tian et al., 2025; Ma et al., 2025; Qian et al., 2024; Lu et al., 2023; Zeng et al., 2024; Sun et al., 2025; Qian et al., 2026; Ren & Huang, 2026; He et al., 2025; Li et al., 2026a; Zhang et al., 2026; Ye et al., 2025; Yin et al., 2024; Ferrante et al., 2022; Meng & Yang, 2023; Kamitani et al., 2025). They do not directly isolate which measured neural signals support an individual prediction. Powerful visual and generative priors can compensate for redundant or missing neural measurements, making reconstruction fidelity alone insufficient for identifying neural evidence. SANE addresses this complementary question.

Structured and anatomically organized neural representations. A parallel line of work explicitly models anatomical, functional, or subject-specific structure in neural representations (Huth et al., 2012; Güçlü & Van Gerven, 2015; Naselaris et al., 2011; Miliotou et al., 2023; Meng & Yang, 2024; Sarch et al., 2023; Choi et al., 2023; Canatar et al., 2023; Linsley et al., 2023; Pogoncheff et al., 2023; Huang et al., 2024; Subramaniam et al., 2024; Turishcheva et al., 2024; Yu et al., 2026; Zhu et al., 2025; Nellen et al., 2026). MindBridge (Wang et al., 2024), MindEye2 (Scotti et al., 2024), UMBRAE (Xia et al., 2024b), NeuroPictor (Huo et al., 2024), and MindTuner (Gong et al., 2025) address subject variation through cross-subject alignment, adaptation, or personalization. Other approaches preserve functional organization within the neural representation itself. NeuroAdapter (Feng et al., 2026) represents regional brain activity as parcel-wise tokens and analyzes their interaction with image features; Brain-IT (Beliy et al., 2026) organizes functionally related voxels into interacting clusters; and BrainMCLIP (Xia et al., 2026) aligns functionally distinct visual regions with different levels of visual representation.

SANE also uses anatomically identifiable neural units, but for evidence identification rather than representation organization alone. HCP-MMP1 parcels serve as candidate evidence units, and SANE selects a fixed-budget subset *for each fMRI response*. Crucially, selection occurs before cross-parcel contextualization, preserving a direct correspondence between selected evidence and its anatomical source. Anatomical structure therefore defines the units on which stimulus-adaptive evidence is identified, rather than only providing structure for downstream representation learning or analysis.

Neural compression and sparse selection. Recent work has shown that accurate visual decoding does not require all measured neural information. BrainBits (Mayo et al., 2024) uses an information bottleneck to characterize how much neural information is sufficient to reproduce the performance of modern reconstruction systems, highlighting the role of strong generative priors. TROI (Wang et al., 2025b) instead learns sparse voxel masks for cross-subject visual decoding, moving from predefined regions toward data-driven voxel selection.

These approaches motivate compact neural representations but leave a different question open: *which* measurements constitute evidence for a particular prediction? A bottleneck characterizes how much neural information is sufficient without identifying which measurements are functionally important, while sparse selection does not by itself establish stimulus dependence or content specificity. SANE therefore combines fixed-capacity, stimulus-adaptive selection with anatomically identifiable parcels and subsequently tests the selected measurements through controlled interventions.

Intervention and evaluation of neural information. Recent work has also moved beyond reconstruction quality as the sole measure of neural information use (Lee & Kuhl, 2016; Rakhimberdina et al., 2021; Guo et al., 2025; Kneeland et al., 2025). ROI-based analyses examine what can be reconstructed from predefined cortical regions (Ozcelik & VanRullen, 2023), while NeuroAdapter (Feng et al., 2026) analyzes image-brain cross-attention to relate neural tokens to generated content. Brain Algebra (Ferrante et al., 2025) provides a more direct intervention-based analysis by constructing concept-specific fMRI perturbations and showing that they can induce predictable changes in reconstructed images. BrainBits (Mayo et al., 2024) further evaluates neural information use through bottleneck curves and model-specific baselines, while BASIC (Xia & Oztireli, 2026) advocates multigranular evaluation beyond conventional reconstruction metrics.

SANE differs in what is selected, when it is selected, and how it is tested. Its evidence set is learned without using intervention outcomes as training objectives, and is then evaluated post hoc using three complementary tests: retention tests *sufficiency*; removal tests *functional importance* against matched controls; and cross-stimulus substitution tests *content specificity* by replacing selected measurements with anatomically corresponding measurements from another stimulus. These are counterfactual tests of the decoding model rather than claims of biological causality. The resulting framework combines stimulus-adaptive anatomical evidence selection, identify-before-integrate processing, and intervention-based validation to distinguish compact neural representations from neural evidence that is functionally important and stimulus dependent.

B Anatomical Interpretation of SANE-Selected Evidence

B.1 HCP-MMP1 as the Anatomical Basis of SANE

Anatomically interpretable evidence selection requires cortical units whose identities are biologically meaningful and comparable across individuals. We therefore adopt the Human Connectome Project Multi-Modal Parcellation version 1.0 (HCP-MMP1), introduced by Glasser et al. (Glasser et al., 2016). Cortical parcellation has a long history in neuroscience, from classical cytoarchitectonic maps such as Brodmann's atlas to modern in-vivo maps that combine anatomical and functional measurements. HCP-MMP1 represents a multimodal extension of this tradition: rather than defining cortical boundaries from a single property, it integrates complementary information from cortical architecture, task-evoked function, resting-state functional connectivity, and cortical topography. The resulting atlas delineates 180 cortical areas in each hemisphere, including both previously established areas and newly refined subdivisions, and provides corresponding anatomical identities across subjects. This makes HCP-MMP1 particularly suitable for SANE because each candidate evidence unit corresponds to an externally defined and named cortical region rather than an arbitrary latent component learned by the decoder.

In SANE, only HCP-MMP1 parcels containing measured voxels within the subject-specific visual-cortex mask are retained as candidate evidence units. Each parcel is independently encoded before cross-parcel contextualization, so its evidence score remains attributable to measurements originating from that anatomical region. The atlas itself is fixed and is never optimized by SANE. Consequently, areas such as V1, V4, FFC, IPS1, or TPOJ3 preserve their anatomical identities throughout evidence scoring and selection, allowing the learned evidence distribution to be interpreted in relation to known cortical organization without imposing a predefined functional hierarchy on the model.

B.2 Group-Level Selection Frequency and Top-32 Ranking

SANE performs evidence selection independently for every fMRI response. Given parcel-wise evidence scores a_p, the $K = 32$ highest-scoring parcels form the response-specific evidence set $\mathcal{S}_K(\mathbf{x})$. Thus, the model does not use a fixed anatomical subset: the selected parcels may change from one

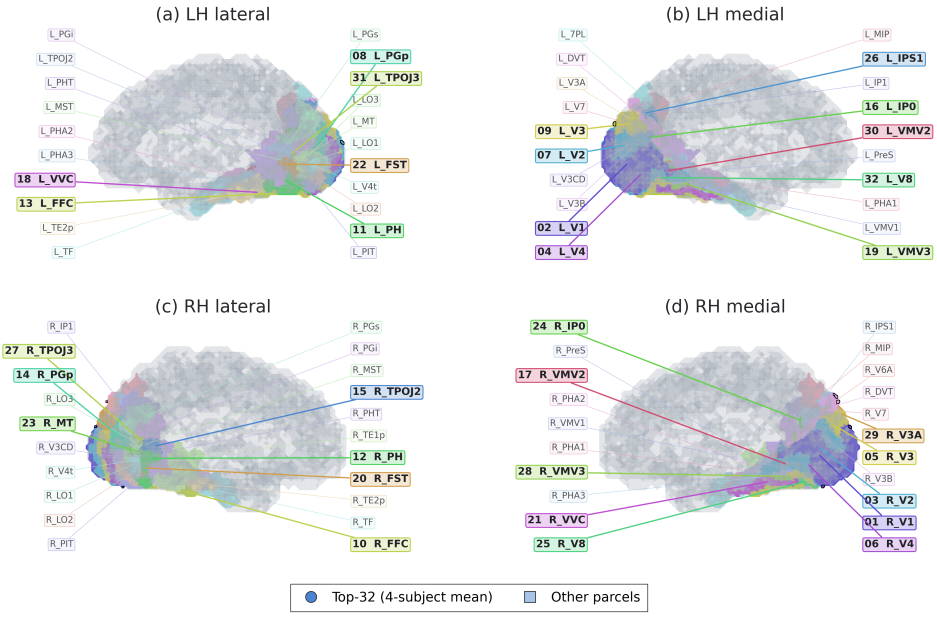


Figure 8: Anatomical distribution of group-level SANE parcel selection. The cortical map visualizes the group-level SANE selection pattern averaged across Subjects 01, 02, 05, and 07. The Top-32 parcels are determined by ranking their group-mean selection frequencies across the four subjects and are highlighted with more saturated colours and prefixed by their group-level rank numbers. For anatomical visualization, we use the HCP-MMP1 cortical geometry of Subject 02, which contains the largest measured parcel set (83 parcels) among the evaluated subjects and therefore provides the most complete spatial reference. The remaining measured parcels are shown using softer colours, while cortical regions outside this measured set are retained as a light-gray, semi-transparent anatomical context. Colours distinguish anatomical parcels and highlighting categories rather than encoding selection-frequency magnitude. Black contours delineate parcel boundaries, and each parcel is identified by its official HCP-MMP1 name, with leader lines used for small or spatially crowded regions. Complementary left and right posterior-oblique views are shown to improve the visibility of parcels distributed across both hemispheres.

stimulus response to another. To determine which cortical regions are nevertheless recruited consistently across the test distribution, we compute the selection frequency of parcel p for subject s as

$$f_{s,p} = \frac{1}{N_s} \sum_{i=1}^{N_s} \mathbb{I}\left[p \in \mathcal{S}_K(\mathbf{x}_{s,i})\right], \tag{11}$$

where $N_s = 3000$ test trials correspond to the official 1,000 shared NSD images and their three independently measured fMRI repetitions. Parcels are then aligned across Subjects 01, 02, 05, and 07 using their HCP-MMP1 identities, and the group-level frequency is obtained by averaging the available subject-level frequencies,

$$\bar{f}_p = \frac{1}{|\mathcal{S}_p|} \sum_{s \in \mathcal{S}_p} f_{s,p}, \tag{12}$$

where $\mathcal{S}_p$ denotes the subjects for which parcel p is available. Parcels are ranked according to $\bar{f}_p$, and Table 2 summarizes the cortical areas represented among the 32 highest-ranked parcels. To avoid redundant descriptions, left- and right-hemisphere homologues are merged into a single functional entry in the table while their selection frequencies are reported separately.

B.3 Functional Organization of Frequently Selected Evidence

The group-level ranking reveals a clear anatomical structure. Early and intermediate visual areas dominate the highest-frequency evidence: bilateral V1 occupies the first two positions, with mean selection frequencies above 92%, while V2, V3, and V4 are also retained in approximately 83–90% of test responses. These areas encode progressively richer information about retinotopic position, local contrast, orientation, contours, color, form, and spatial structure, providing the low- and intermediate-level visual information required for accurate decoding. The highly selected set nevertheless extends well beyond early visual cortex. Ventral high-level regions including FFC, VVC, V8, VMV2, and VMV3 contribute representations associated with complex object form, faces, color, appearance, and scenes; MT and FST extend the evidence set into motion-sensitive visual processing; V3A, IP0, and IPS1 contribute depth, spatial organization, and visuospatial attention; and PGp together with TPOJ2 and TPOJ3 extends the selected evidence toward higher-order spatial and contextual integration. The frequencies of these higher-order regions are generally lower and more subject-dependent than those of V1–V4, indicating a graded hierarchy rather than a fixed division between "relevant" and "irrelevant" cortex.

Taken together, the Top-32 parcels are therefore not a scattered collection of unrelated cortical regions. Instead, SANE preferentially selects a distributed visual hierarchy spanning early visual cortex, ventral high-level visual areas, motion-sensitive regions, dorsal visuospatial regions, and parietal/contextual areas. The frequently selected evidence consequently extends from early visual encoding toward progressively richer representations of object appearance, scenes, spatial organization, and contextual information, rather than relying on a single localized cortical subsystem. This organization is broadly consistent with the known hierarchical structure of human visual processing and suggests that SANE's compact evidence budget preserves complementary information distributed across multiple stages of the visual system. Importantly, however, this analysis describes the anatomical organization of *model-assigned decoding evidence*: selection frequency is neither a measure of neural activation magnitude nor evidence that a cortical region is biologically necessary or causally responsible for visual perception or reconstruction. A parcel may also be selected infrequently because its information is redundant with other measurements or less strongly represented by the current decoding objective. We therefore interpret these results as showing that SANE repeatedly identifies anatomically structured measurements as useful decoding evidence, rather than as establishing biological causality.

C Implementation and Experimental Details

C.1 SANE Architecture and Optimization Details

This section provides the architectural and optimization details omitted from the main text for clarity. The main paper emphasizes the separation between parcel-local evidence identification and

Table 2: Functional interpretation of cortical areas represented in the Top-32 SANE-selected parcels. Scores denote group-mean selection frequency across Subjects 01, 02, 05, and 07. Bilateral homologous parcels are merged for compactness.

Area	Selection Freq.	Anatomical / medical name	Approx. biological function	Relevance to visual reconstruction
V1	L: 92.08%, R: 92.39%	Primary visual cortex	Edges, orientation, contrast, basic retinotopic structure.	Fine spatial structure, contours, and low-level geometry.
V2	L: 86.76%, R: 89.96%	Secondary visual cortex	Integrates V1 signals into contours, local shape, disparity, and color.	Local shape, boundaries, and intermediate visual structure.
V3	L: 82.53%, R: 87.52%	Visual area 3	Intermediate form, spatial structure, and dynamic visual processing.	Connects low-level features with larger-scale visual form.
V4	L: 89.52%, R: 87.17%	Visual area 4	Color, curvature, complex shape, and visual form.	Object color, shape, and appearance fidelity.
PGp	L: 84.23%, R: 69.53%	Posterior area PG	Higher-order visuospatial and contextual integration.	Global scene context and spatial relationships.
FFC	L: 72.78%, R: 80.73%	Fusiform face complex	High-level face and complex object-form processing.	Face, object identity, and category-related appearance.
PH	L: 78.68%, R: 74.56%	Area PH	Higher-order visual and contextual processing.	Complex object/scene and contextual information.
TPOJ2	R: 65.25%	Temporo-parieto-occipital junction area 2	Multimodal visual and contextual integration.	High-level contextual and cross-regional visual information.
IP0	L: 59.18%, R: 49.43%	Intraparietal area 0	Visuospatial attention and spatial integration.	Object location and spatial relationships.
VMV2	L: 47.12%, R: 57.89%	Ventromedial visual area 2	High-level ventral visual and scene processing.	Scene structure and complex visual appearance.
VVC	L: 57.77%, R: 53.27%	Ventral visual complex	High-level color, texture, object, and scene representation.	Rich object appearance, texture, and semantic visual content.
VMV3	L: 54.41%, R: 48.94%	Ventromedial visual area 3	Higher-order scene and visual-form processing.	Global scene appearance and complex visual structure.
FST	L: 52.68%, R: 53.72%	Fundus of superior temporal area	Complex motion and dynamic visual processing.	Motion-related and dynamic visual information.
MT	R: 49.91%	Middle temporal visual area (MT/V5)	Motion direction and speed processing.	Motion and dynamic-scene cues.
V8	L: 45.56%, R: 49.34%	Visual area 8	Higher-order color and object-form processing.	Color-rich and high-level object appearance.
IPS1	L: 49.31%	Intraparietal sulcus area 1	Spatial attention and visuospatial working memory.	Spatial layout and attentional selection.
TPOJ3	L: 46.15%, R: 49.29%	Temporo-parieto-occipital junction area 3	High-level multimodal and contextual integration.	Global contextual and cross-regional visual information.
V3A	R: 47.33%	Visual area 3A	Depth, motion, stereoscopic and spatial processing.	Depth, 3D structure, and spatial layout.

post-selection evidence integration, whereas the implementation below specifies the exact parameterization used in all experiments.

C.1.1 Parcel Definition and Local Tokenization

We use the HCP-MMP1 atlas as a fixed, externally defined cortical parcellation with consistent parcel identities across subjects. The atlas is chosen independently of decoding performance and serves only to define anatomically attributable candidate units. We retain parcels that contain measured voxels within the subject-specific visual-cortex mask.

The resulting numbers of non-empty parcels are 81, 83, 81, and 82 for NSD Subjects 1, 2, 5, and 7, respectively. These values arise from subject-specific measured cortical coverage and are not model hyperparameters.

For parcel p, the voxel response is projected independently into a 512-dimensional common token space:

$$\mathbf{h}_p = \mathrm{LN}\left[\mathrm{GELU}\left(\mathbf{W}_p \mathbf{x}[\mathcal{V}_{s,p}] + \mathbf{b}_p\right)\right] \in \mathbb{R}^{512}. \tag{13}$$

The parcel-specific projection matrix $\mathbf{W}_p$ accommodates the different numbers of voxels contained in different HCP-MMP1 parcels. No cross-parcel operation is applied before the evidence score is computed.

C.1.2 Evidence Scorer and Budget-Constrained Selection

A shared two-layer scorer maps each local parcel token to a scalar evidence score:

$$a_p = \mathbf{w}_2^\top \,\mathrm{Dropout}\left[\mathrm{GELU}\left(\mathbf{W}_1\,\mathrm{LN}(\mathbf{h}_p) + \mathbf{b}_1\right)\right] + b_2, \tag{14}$$

where the hidden dimension is 256. The scorer parameters are shared across parcels, whereas the score itself depends on the parcel token from the current fMRI response.

During training, hard Top-K selection is approximated by a differentiable budget-constrained gate,

$$m_p = \sigma\left(\frac{a_p - \tau}{T}\right), \qquad \sum_{p=1}^{P_s} m_p \simeq K, \tag{15}$$

where T is the relaxation temperature and the sample-specific threshold τ is determined by bisection. The temperature is annealed from 0.5 to 0.1 during training, and the threshold is solved using 32 bisection steps. The local token supplied to the contextual encoder is

$$\widetilde{\mathbf{h}}_p = m_p \mathbf{h}_p. \tag{16}$$

At validation and test time, the relaxation is discarded and evidence is selected by exact score ranking:

$$\mathcal{S}_K(\mathbf{x}) = \mathrm{TopK}_{p\in\{1,\dots,P_s\}}(a_p, K). \tag{17}$$

Thus, inference always retains exactly K parcel tokens. The default $K = 32$ is used only as the representative operating point for intervention and reconstruction experiments; the complete $K \in \{64, 48, 32, 16, 8, 4, 2, 1\}$ budget analysis is reported in Sec. 3.2.

C.1.3 Post-Selection Contextual Encoder

After evidence identification, the retained parcel tokens are integrated using a lightweight Transformer. The contextual encoder contains two Transformer layers, each with eight attention heads and a 2048-dimensional feed-forward block. The token dimensionality is 512 throughout the contextual encoder.

For hard inference, the selected tokens are collected as

$$\mathbf{H}_K = [\mathbf{h}_p]_{p\in\mathcal{S}_K(\mathbf{x})}, \tag{18}$$

and contextualized as

$$\mathbf{E}_K = T_\theta(\mathbf{H}_K) \in \mathbb{R}^{K\times 512}. \tag{19}$$

During training, the same contextual encoder operates on the softly gated parcel tokens.

The Transformer is introduced only after evidence scoring and selection. Its role is therefore evidence integration rather than evidence identification.

C.1.4 Visual Prediction Head

The contextualized evidence tokens are mean-pooled and projected into the 768-dimensional frozen DINOv3 feature space:

$$\widehat{\mathbf{v}} = \mathrm{Norm}\left[q_\theta\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{E}_{K,k}\right)\right], \tag{20}$$

where q_θ is a $512 \to 1024 \to 768$ prediction head. Both the predicted representation and the frozen DINOv3 target are L_2-normalized because evidence learning is optimized with cosine distance. Normalization therefore removes feature magnitude from the objective and makes the loss depend only on directional agreement in the DINOv3 representation space.

C.1.5 Dense Initialization

Before learning sparse evidence, we train a subject-specific dense parcel decoder using all available HCP-MMP1 parcels. The dense model uses the same parcel tokenizer, contextual encoder, and DINOv3 prediction head as SANE but does not apply evidence selection.

The trained parcel tokenizer, contextual encoder, and visual prediction head are then used to initialize the corresponding SANE components, while the stimulus-adaptive evidence scorer is initialized from scratch. SANE therefore begins from an established brain-to-vision mapping and learns which parcel-level measurements can preserve that mapping under a restricted evidence budget.

This initialization is used to stabilize evidence learning rather than to define the selected evidence itself. All evidence scores remain stimulus dependent and are learned after the evidence gate is introduced.

C.1.6 EVIDENCE-LEARNING OPTIMIZATION

The dense initialization model is trained for 20 epochs with batch size 128 using AdamW, a learning rate of 10^{-4}, weight decay of 10^{-4}, and gradient clipping at 1.0.

SANE is subsequently trained for 20 epochs with batch size 128. The newly introduced evidence scorer uses a learning rate of 10^{-4}, while parameters inherited from the dense model use a smaller learning rate of 2×10^{-5}. The differentiable selection temperature is annealed from 0.5 to 0.1 over training. Model selection is based on validation performance.

For a mini-batch of B samples, SANE minimizes

$$\mathcal{L}_{\text{evidence}} = \frac{1}{B} \sum_{b=1}^{B} \left[1 - \cos\left(\widehat{\mathbf{v}}_b, \mathbf{v}_b\right)\right]. \tag{21}$$

No auxiliary sparsity regularizer is used because the evidence capacity is controlled directly by the budget-constrained selector.

C.2 FULL EXPERIMENTAL PROTOCOL

C.2.1 DATASET AND PREPROCESSING.

We use the Natural Scenes Dataset (NSD) and follow the standard four-participant protocol, evaluating Subjects 1, 2, 5, and 7, for whom all scanning sessions are available. Each participant contributes 30,000 single-trial fMRI responses to 10,000 natural images, with every image presented three times. We use the provided renormalized single-trial beta estimates in the subject's native 1.8-mm functional space. Responses are restricted to the subject-specific `nsdgeneral` voxel set and standardized per voxel within each session following the MindEye2 preprocessing pipeline; importantly, we retain individual trials rather than averaging repetitions before evidence selection. This yields 15,724, 14,278, 13,039, and 12,682 input voxels for Subjects 1, 2, 5, and 7, respectively.

To obtain anatomically grounded candidate evidence units, we begin with `nsdgeneral`, the standard NSD cortical ROI mask that defines the visually responsive cortical voxels commonly used for NSD-based decoding analyses (Allen et al., 2022). We intersect the ordered `nsdgeneral` voxels with the subject-specific left- and right-hemisphere HCP-MMP1 label volumes. Every nonempty atlas region becomes one parcel, and its indices are stored in the exact column order of the beta matrix. Because HCP-MMP1 coverage within `nsdgeneral` is subject dependent, this procedure produces 81, 83, 81, and 82 parcels for Subjects 01, 02, 05, and 07, respectively, rather than forcing all participants into an artificial fixed-size layout. Voxels without an HCP-MMP1 label are not assigned to a parcel. We additionally verify that the ROI volumes share the same shape and affine and that the number of `nsdgeneral` voxels matches the corresponding beta-matrix width, preventing silent spatial or subject misalignment.

The 9,000 participant-specific, non-shared stimuli (27,000 trials) form the development set. Using a fixed seed, we hold out 5% of its *stimulus identities* for validation, giving 8,550 training images (25,650 trials) and 450 validation images (1,350 trials). All three repetitions of an image remain in the same partition, eliminating repetition leakage. The official 1,000 images shared across NSD participants are reserved for final testing and contribute exactly 3,000 trials per participant; in the released MindEye2 metadata this complete split is named `new_test`, whereas the legacy `test` split is not used. During evaluation, each repetition is decoded independently, the three predictions are averaged in representation space, and the result is re-normalized before scoring. Visual supervision is fixed across participants: each stimulus is represented by the L2-normalized 768-dimensional CLS token of a frozen DINOv3 ViT-B/16 encoder. Trial-to-image pairing is performed only through the recorded trial identifiers and zero-based COCO stimulus identifiers, which are checked against both the beta matrix and the feature cache.

C.2.2 EVALUATION PROTOCOL.

All evaluations use the complete shared-1,000 test set and exact Top-K selection. For representation-space evaluation, the three fMRI repetitions are decoded independently; we report both single-trial performance and, as the primary measure, performance after averaging and re-normalizing the three predicted representations of each stimulus. Intervention conditions are paired

at the stimulus level, with random controls averaged over ten draws and confidence intervals obtained from 10,000 paired bootstrap resamples. For image reconstruction, the three repetitions yield one conditioning representation per stimulus according to each backbone's native interface. Native and SANE-driven variants use identical test stimuli, sampling settings, and three prespecified generation seeds, with no best-of-N selection. For each test stimulus, reconstruction metrics are computed separately for the three generation seeds. Reported mean performance is averaged over stimuli and seeds, while the reported standard deviation is the mean of the per-stimulus standard deviations across the three generation seeds.

C.2.3 EVALUATION METRICS.

We evaluate SANE at both the neural-representation and image-reconstruction levels. At the representation level, we report cosine similarity between predicted and target DINOv3 features and 1,000-way Top-1 and Top-10 retrieval accuracy. For repeated stimuli, predictions from the three fMRI repetitions are averaged before evaluation. Intervention effects are measured by the decrease in target similarity after evidence removal and by the recipient-to-donor directional shift under cross-stimulus substitution.

For final reconstructions, we follow the evaluation protocol commonly used in NSD image-reconstruction studies (Ozcelik & VanRullen, 2023; Scotti et al., 2023). Pixel-level fidelity is assessed using pixel-wise correlation (PixCorr) and structural similarity (SSIM) (Wang et al., 2004). PixCorr computes the Pearson correlation between corresponding pixels of the reconstructed and ground-truth images and therefore measures agreement in absolute color and spatial layout, although it is sensitive to small spatial misalignments. SSIM compares local luminance, contrast, and structural patterns, providing a complementary measure of image-level structural fidelity.

We evaluate feature-level correspondence using two-way identification accuracy. For each reconstruction, its similarity to the corresponding ground-truth image is compared with its similarity to each nonmatching test image in a pretrained feature space; an identification is counted as correct when the matched image is more similar than the nonmatching image. We compute this accuracy using features from AlexNet layers 2 and 5 (Krizhevsky et al., 2012), Inception-V3 (Szegedy et al., 2015), and CLIP (Radford et al., 2021). AlexNet layer 2 primarily captures relatively low-level visual properties such as edges and local textures, whereas layer 5 reflects more abstract visual structure. Inception and CLIP provide progressively higher-level measures of object and semantic correspondence.

We additionally report correlation distance, defined as one minus the feature correlation, in EfficientNet-B1 (Tan & Le, 2019) and SwAV-ResNet50 (Caron et al., 2020) feature spaces. These distances measure perceptual similarity using supervised and self-supervised visual representations, respectively. Higher values indicate better performance for PixCorr, SSIM, Alex(2), Alex(5), Inception, and CLIP identification, whereas lower values indicate better performance for EfficientNet-B1 (Eff) and SwAV-ResNet50 (SwAV) distances. All reconstruction backbones are evaluated using identical image preprocessing, feature extractors, and metric implementations.

C.2.4 BASELINES AND CONTROLS.

For representation-level evaluation, the *dense* baseline retains all available HCP-MMP1 parcels and provides the reference decoding performance for each subject. SANE uses exact hard Top-K selection at inference, with $K = 32$ unless otherwise stated.

For intervention experiments, selected-parcel removal and substitution are compared with uniform-random controls using the same number of parcels. Because HCP-MMP1 parcels contain different numbers of voxels, we additionally use voxel-count-matched random controls to account for non-specific effects caused by removing or replacing different amounts of fMRI input. Random-control results are averaged over ten independently sampled control sets.

For final image reconstruction, we evaluate representation portability with five pretrained reconstruction backbones: MindEye2, NeuroPictor, NeuroAdapter, NeuroFlow, and Brain-IT. For each backbone, the SANE-driven system is compared directly with its corresponding native pretrained reconstruction pathway. Within each backbone, the generative model, sampling configuration,

post-processing, evaluation stimuli, and metric implementation are kept fixed; only the neural-to-conditioning pathway is replaced.

C.2.5 IMPLEMENTATION DETAILS.

Unless stated otherwise, we train a separate subject-specific model for each evidence budget K using PyTorch in FP32. We first optimize the dense HCP-parcel encoder for 20 epochs with a batch size of 128 and a cosine-distance objective. We use AdamW with a learning rate of 10^{-4}, weight decay of 10^{-4}, and a maximum gradient norm of 1.0. The checkpoint with the highest validation cosine similarity initializes SANE's parcel tokenizer, contextual Transformer, and prediction head; only the evidence gate is newly initialized. This dense warm start gives every budget the same learned representation basis and prevents differences in selection quality from being confounded by unstable representation learning from scratch.

We then fine-tune SANE for 20 epochs with a batch size of 128. AdamW uses separate learning rates of 10^{-4} for the new gate and $2{\times}10^{-5}$ for the inherited modules, with weight decay 10^{-4} and gradient clipping at 1.0. During training, the budget-constrained soft mask is computed independently for each sample by 32 bisection steps, and its temperature is geometrically annealed from 0.5 to 0.1. The initially smooth mask distributes learning signal across parcels, whereas progressive sharpening reduces the train-inference discrepancy. Validation and inference always use exact hard Top-K selection. A plateau scheduler halves both learning rates when the hard-selection validation loss ceases to improve (patience 4; minimum learning rate 10^{-6}), and the final checkpoint is chosen solely by hard Top-K validation cosine similarity; soft-mask validation is diagnostic only. The shared-1,000 test set is never used for checkpoint or hyperparameter selection.

We retain the same stimulus-disjoint split, initialization protocol, optimizer, and training schedule across parcel budgets, while training each $(\text{subject}, K)$ model independently. Results are computed per subject and macro-averaged with equal subject weight. Stochastic reconstructions use paired generation seeds for each native/SANE comparison, without best-of-N selection. Intervention confidence intervals are estimated using 10,000 paired bootstrap resamples over stimuli, and size-matched random controls are averaged over ten draws. These choices ensure that training, stochastic generation, and statistical aggregation remain consistent across compared conditions.

D EVIDENCE SELECTION AND INTERVENTION ANALYSES

D.1 EVIDENCE SELECTION ANALYSES AND BUDGET ABLATIONS

We further analyze how SANE distributes its selected evidence anatomically and report additional subject-level and qualitative results. These analyses complement the intervention experiments by characterizing where evidence is repeatedly selected and by documenting the robustness of the aggregate findings across subjects.

D.1.1 ANATOMICAL DISTRIBUTION OF SELECTED EVIDENCE

Because evidence scoring is performed before cross-parcel contextualization, every selected token remains associated with an identifiable HCP-MMP1 parcel. We therefore examine how frequently each parcel is selected under the $K = 32$ budget. For each subject, selection frequency is defined as the fraction of test trials for which a parcel appears in the exact hard Top-32 set. Frequencies are aligned by HCP-MMP1 parcel identity and averaged across subjects for cortical visualization.

As shown in Fig. 9, early visual areas including V1–V4 are among the most consistently selected parcels, while higher-order regions such as FFC and PGp also appear frequently. This distribution shows that SANE repeatedly draws evidence from multiple levels of the visual hierarchy while the selected subset remains stimulus dependent. Selection frequency reflects model selection consistency and should not be interpreted as neural activation magnitude or biological importance.

D.1.2 PARCEL-BUDGET ABLATION

We evaluated how the number of selected cortical parcels affects decoding performance using four NSD participants (subj01, subj02, subj05, and subj07). Because the number of

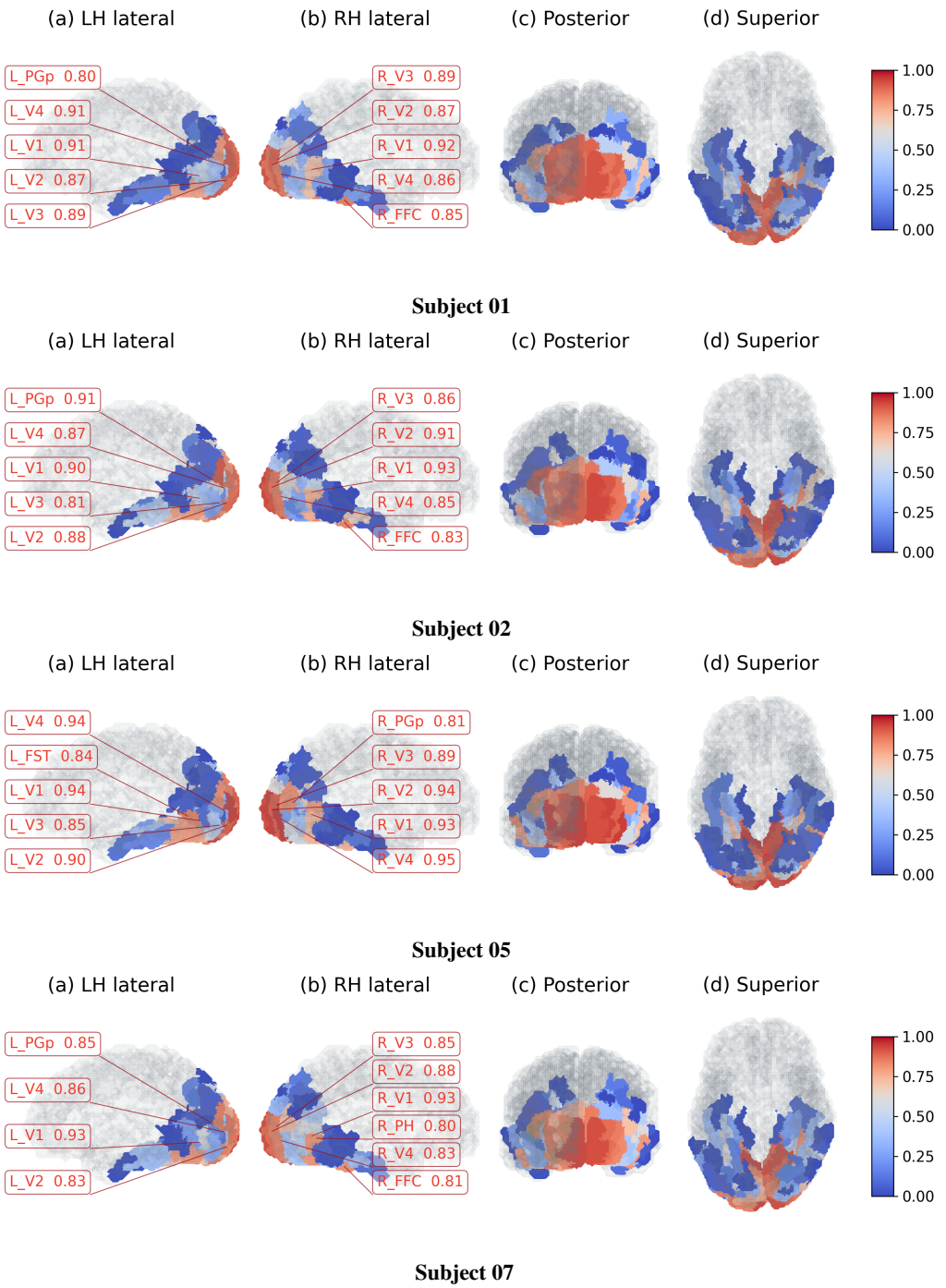


Figure 9: $K = 32$ parcel-selection frequencies for the group mean and individual subjects. Rows show the group mean and Subjects 01, 02, 05, and 07, respectively, with four cortical views per row.

Table 3: Parcel-budget ablation. Accuracy-efficiency trade-off as the evidence budget is reduced from the full HCP-MMP1 parcel set to a single parcel. Results are macro-averaged across NSD Subjects 1, 2, 5, and 7.

Metric	**Dense**	64	48	32	16	8	4	2	1
Retained (%)	100.00	78.30	58.72	**39.15**	19.57	9.79	4.89	2.45	1.22
3-Repeat Cos. ↑	0.4015	0.4044	0.4016	**0.3931**	0.3023	0.2438	0.1957	0.1676	0.1452
Dense Ret. (%)	100.00	100.69	100.00	**97.89**	75.28	60.72	48.73	41.76	36.21
GFLOPs ↓	1.073	0.860	0.652	**0.445**	0.241	0.139	0.089	0.063	0.051
B32 Lat. (ms) ↓	21.20	18.02	16.24	**14.75**	13.01	12.20	11.75	11.51	11.37
Peak Mem. (MiB) ↓	109.78	106.10	98.18	**90.26**	82.15	81.88	81.88	81.88	81.88

valid HCP-MMP1 parcels varied slightly across participants, the dense models retained $P_s \in \{81, 83, 81, 82\}$ parcels, respectively. For each participant, we independently trained SANE with $K \in \{64, 48, 32, 16, 8, 4, 2, 1\}$ under the same data split, initialization, optimization schedule, and evaluation protocol. Each sparse model was initialized from its corresponding subject-specific dense checkpoint. During training, the evidence gate used a differentiable budget-constrained soft Top-K approximation, whereas validation and held-out testing used exact hard Top-K selection. Performance was evaluated on the standard held-out set of 1,000 images, each with three fMRI repetitions. We report the cosine similarity after averaging predictions across the three repetitions. Relative performance was computed within each participant as

$$R_s(K) = \frac{\mathrm{Cosine}_s(K)}{\mathrm{Cosine}_s(\mathrm{Dense})} \times 100\%, \tag{22}$$

and then averaged equally across participants.

The dense models achieved a mean three-repeat cosine similarity of 0.4015 ± 0.0266 across participants. Retaining 64 parcels achieved 0.4044 ± 0.0282, corresponding to $100.69\% \pm 1.25\%$ of dense performance, while 48 parcels achieved 0.4016 ± 0.0273 or $100.00\% \pm 1.08\%$. Importantly, selecting only 32 parcels still achieved 0.3931 ± 0.0282 and retained $97.89\% \pm 2.20\%$ of dense performance. This behavior was consistent across participants, with subject-level retention ranging from 94.88% to 99.61%. In contrast, reducing the budget to 16 parcels caused a marked drop to 0.3023 ± 0.0228, corresponding to only $75.28\% \pm 2.49\%$ of dense performance. Performance decreased further to 60.72%, 48.73%, 41.76%, and 36.21% at $K = 8, 4, 2$, and 1, respectively. These results reveal a clear capacity-performance transition between 16 and 32 parcels: most of the decodable visual information is preserved at $K = 32$, whereas more aggressive compression produces substantial degradation.

We therefore selected $K = 32$ as the default operating point because it provides the strongest balance between decoding fidelity, sparsity, and computational efficiency. Across participants, $K = 32$ retains only 39.15% of the available parcels, reducing the number of tokens entering the Transformer by approximately 60.85%. Because the selected parcel tokens are physically gathered before contextual processing, token-linear Transformer operations are reduced proportionally. Moreover, the self-attention matrix contains only approximately $(32/P_s)^2 = 15.33\%$ of the dense pairwise interactions on average, corresponding to an approximately 84.67% reduction in the quadratic attention component. Compared with $K = 48$, the $K = 32$ model uses one-third fewer tokens and approximately 55.6% fewer attention interactions, while sacrificing only 2.11 percentage points of relative decoding performance. Conversely, reducing the budget from 32 to 16 halves the token count but incurs a substantially larger 22.61 percentage-point performance loss. Thus, $K = 32$ lies near the empirical knee of the capacity-performance curve, preserving nearly all dense decoding performance while providing a substantially more compact neural evidence representation.

Table 4 provides the complete subject-level budget curve underlying these aggregate results.

D.1.3 Subject-Level Computational Efficiency

We evaluated the efficiency of SANE under different parcel budgets ($K = 64, 48, 32, 16, 8, 4, 2, 1$) using Subjects 01, 02, 05, and 07. Inference latency and GPU memory were measured for the complete hard-selection pathway from fMRI input to the predicted visual feature, excluding data loading

Table 4: Subject-level parcel-budget ablation. Three-repeat decoding performance as the evidence budget is reduced from the full parcel set to $K = 1$. Dense retention is computed relative to the dense model of each subject.

Subject	Metric	Dense	64	48	32	16	8	4	2	1
01 (81)	Retained (%)	100.00	79.01	59.26	39.51	19.75	9.88	4.94	2.47	1.23
	3-Repeat Cos. ↑	0.3975	**0.4066**	0.4020	0.3954	0.3069	0.2483	0.2002	0.1688	0.1483
	Dense Ret. (%)	100.00	**102.30**	101.14	99.46	77.21	62.47	50.37	42.47	37.30
02 (83)	Retained (%)	100.00	77.11	57.83	38.55	19.28	9.64	4.82	2.41	1.20
	3-Repeat Cos. ↑	0.3877	**0.3902**	0.3895	0.3862	0.2992	0.2477	0.1923	0.1702	0.1455
	Dense Ret. (%)	100.00	**100.63**	100.47	99.61	77.17	63.88	49.61	43.89	37.52
05 (81)	Retained (%)	100.00	79.01	59.26	39.51	19.75	9.88	4.94	2.47	1.23
	3-Repeat Cos. ↑	0.4401	**0.4428**	0.4392	0.4296	0.3292	0.2621	0.2145	0.1800	0.1531
	Dense Ret. (%)	100.00	**100.60**	99.78	97.60	74.79	59.56	48.75	40.89	34.78
07 (82)	Retained (%)	100.00	78.05	58.54	39.02	19.51	9.76	4.88	2.44	1.22
	3-Repeat Cos. ↑	**0.3808**	0.3779	0.3756	0.3613	0.2740	0.2169	0.1759	0.1516	0.1342
	Dense Ret. (%)	**100.00**	99.24	98.62	94.88	71.96	56.96	46.18	39.81	35.23
Mean	Retained (%)	100.00	78.30	58.72	**39.15**	19.57	9.79	4.89	2.45	1.22
	3-Repeat Cos. ↑	0.4015	0.4044	0.4016	**0.3931**	0.3023	0.2438	0.1957	0.1676	0.1452
	Dense Ret. (%)	100.00	100.69	100.00	**97.89**	75.28	60.72	48.73	41.76	36.21

and downstream image generation. Latency was averaged over 100 runs after warm-up, and FLOPs were estimated from the dominant linear and matrix-multiplication operations. Results were averaged across the four subjects. Reducing K from 64 to 32 decreased the estimated computation from 0.860 to 0.445 GFLOPs and reduced batch-32 latency from 18.022 to 14.747 ms, while retaining 97.89% of the dense decoding performance. These results identify $K = 32$ as a favorable trade-off between decoding accuracy and computational efficiency. Complete subject-level measurements are reported in Table 5.

D.2 Validating Adaptive Evidence Selection

SANE assigns an evidence score to every cortical parcel before retaining a small subset for contextualization and decoding. We performed a sequence of controlled analyses to determine whether this selection is merely a noisy version of a fixed global parcel preference or contains reproducible stimulus-specific information. The validation follows four steps. First, we test whether the selected parcels are more useful than random subsets. Second, we test whether the mask selected for the current stimulus is more useful than masks selected for other stimuli. Third, we determine whether stimulus-dependent selection is reproducible across independent fMRI repeats. Finally, we vary the inference-time evidence budget to identify the regime in which dynamic selection is most beneficial.

D.2.1 Experimental Setup

Experiments use NSD Subjects 01, 02, 05, and 07 and the 1,000 shared held-out test stimuli available for every subject. Each test stimulus has three fMRI repetitions. Unless stated otherwise, we use each subject's trained $K = 32$ SANE checkpoint and its frozen parcel tokenizer, EvidenceGate, Transformer, and DINOv3 decoder. For performance evaluation, gate logits are first averaged over the three repetitions of stimulus i, producing one stimulus-level ranking over parcels. The corresponding Top-K set is denoted by S_i^K. The same mask is applied independently to all three fMRI repetitions; the three unit-normalized predictions are then averaged and renormalized before computing cosine similarity with the target DINOv3 feature.

We compare four mask definitions:

- **Matched Dynamic-K:** stimulus i is decoded using its own mask S_i^K.
- **Mismatched Dynamic-K:** the fMRI remains that of stimulus i, but it is decoded using S_j^K from a different stimulus $j \neq i$. We use ten independently sampled derangements,

Table 5: Subject-level computational efficiency. Computational cost, latency, throughput, and memory usage as the evidence budget is reduced from the dense model to $K = 1$.

Subject	Metric	Dense	64	48	32	16	8	4	2	1
01	Total GFLOPs ↓	1.065	0.862	0.653	**0.447**	0.242	0.141	0.090	0.065	0.053
	Transformer GFLOPs ↓	1.046	0.822	0.613	**0.407**	0.202	0.101	0.050	0.025	0.013
	B1 Lat. (ms) ↓	11.24	10.96	10.93	**10.88**	10.83	10.86	10.83	10.87	10.91
	B32 Lat. (ms) ↓	21.41	18.07	16.27	**14.78**	13.00	12.26	11.77	11.56	11.41
	B32 Throughput ↑	1494.6	1771.1	1966.5	**2164.6**	2462.5	2610.6	2717.7	2768.4	2804.4
	Peak Mem. (MiB) ↓	112.76	109.87	101.86	**93.87**	85.85	85.42	85.42	85.42	85.42
	Incr. Mem. (MiB) ↓	41.50	38.11	30.10	**22.11**	14.09	13.66	13.66	13.66	13.66
02	Total GFLOPs ↓	1.090	0.861	0.652	**0.446**	0.241	0.140	0.089	0.064	0.052
	Transformer GFLOPs ↓	1.073	0.822	0.613	**0.407**	0.202	0.101	0.050	0.025	0.013
	B1 Lat. (ms) ↓	11.36	11.26	11.27	**11.27**	11.21	11.19	11.23	11.29	11.26
	B32 Lat. (ms) ↓	20.64	18.13	16.32	**14.88**	13.12	12.32	11.86	11.63	11.50
	B32 Throughput ↑	1550.4	1765.0	1960.7	**2151.1**	2439.6	2597.3	2699.0	2751.6	2782.4
	Peak Mem. (MiB) ↓	110.62	106.42	98.41	**90.40**	82.40	82.18	82.18	82.18	82.18
	Incr. Mem. (MiB) ↓	42.94	38.23	30.23	**22.22**	14.21	14.00	14.00	14.00	14.00
05	Total GFLOPs ↓	1.062	0.859	0.651	**0.444**	0.240	0.138	0.088	0.062	0.050
	Transformer GFLOPs ↓	1.046	0.822	0.613	**0.407**	0.202	0.101	0.050	0.025	0.013
	B1 Lat. (ms) ↓	11.05	10.97	11.06	**10.98**	11.06	10.99	11.04	11.01	11.04
	B32 Lat. (ms) ↓	21.22	17.83	16.13	**14.56**	12.84	12.01	11.63	11.31	11.16
	B32 Throughput ↑	1508.3	1795.0	1983.9	**2197.8**	2491.3	2664.0	2751.8	2829.7	2866.8
	Peak Mem. (MiB) ↓	108.00	104.30	96.64	**88.63**	80.28	79.85	79.85	79.85	79.85
	Incr. Mem. (MiB) ↓	42.31	38.11	30.45	**22.44**	14.09	13.66	13.66	13.66	13.66
07	Total GFLOPs ↓	1.075	0.859	0.651	**0.444**	0.239	0.138	0.087	0.062	0.050
	Transformer GFLOPs ↓	1.059	0.822	0.613	**0.407**	0.202	0.101	0.050	0.025	0.013
	B1 Lat. (ms) ↓	11.29	11.13	11.16	**11.10**	11.14	11.13	11.10	11.07	11.11
	B32 Lat. (ms) ↓	21.53	18.06	16.22	**14.77**	13.09	12.19	11.75	11.53	11.39
	B32 Throughput ↑	1486.3	1771.6	1973.1	**2167.1**	2445.1	2624.3	2724.1	2776.0	2809.0
	Peak Mem. (MiB) ↓	107.75	103.80	95.79	**88.11**	80.08	80.08	80.08	80.08	80.08
	Incr. Mem. (MiB) ↓	42.62	38.17	30.16	**22.48**	14.45	14.45	14.45	14.45	14.45
Mean	Total GFLOPs ↓	1.073	0.860	0.652	**0.445**	0.241	0.139	0.089	0.063	0.051
	Transformer GFLOPs ↓	1.056	0.822	0.613	**0.407**	0.202	0.101	0.050	0.025	0.013
	B1 Lat. (ms) ↓	11.24	11.08	11.10	**11.06**	11.06	11.04	11.05	11.06	11.08
	B32 Lat. (ms) ↓	21.20	18.02	16.24	**14.75**	13.01	12.20	11.75	11.51	11.37
	B32 Throughput ↑	1509.9	1775.7	1971.1	**2170.1**	2459.6	2624.1	2723.1	2781.4	2815.6
	Peak Mem. (MiB) ↓	109.78	106.10	98.18	**90.26**	82.15	81.88	81.88	81.88	81.88
	Incr. Mem. (MiB) ↓	42.34	38.15	30.23	**22.31**	14.21	13.94	13.94	13.94	13.94

Table 6: Stimulus-level paired comparison of four parcel-selection strategies at $K = 32$. The matched dynamic condition uses each stimulus's own Top-K mask and serves as the reference. Δ is the DINO cosine difference between matched dynamic selection and the indicated control. Mismatched and random scores are first averaged over ten control masks within each stimulus. Win rate is the fraction of the 4,000 paired differences above zero.

Comparison	Mean Δ	Median Δ	Win rate	Q25	Q75
Dynamic-Static	-0.003	-0.001	48.6%	-0.028	$+0.024$
Dynamic-Mismatched	$+0.008$	$+0.007$	57.6%	-0.018	$+0.034$
Dynamic-Random	$+0.067$	$+0.056$	78.3%	$+0.007$	$+0.118$

ensuring that no stimulus receives its own mask, and average the ten control scores within each stimulus before paired analysis.

- **Static-K:** all test stimuli use the same K parcels. For each budget K, we apply the frozen gate to the training split, count the frequency with which each parcel appears in the trial-level Top-K, and retain the K most frequent parcels. The held-out test set is therefore never used to define the static mask.
- **Random-K:** each stimulus is decoded with a uniformly sampled subset of K parcels, without replacement. As for the mismatched control, ten random masks are averaged within each stimulus.

All primary comparisons are paired at the stimulus level. With four subjects and 1,000 stimuli per subject, every comparison contains 4,000 paired differences. We report the mean and median paired difference, the 25th and 75th percentiles, and the fraction of stimuli for which the matched dynamic condition performs better. Unless explicitly labeled as a cross-subject summary, variability in this section is calculated across the 4,000 subject-stimulus summaries rather than across only four subject means.

D.2.2 Stimulus-Matched versus Fixed and Mismatched Selection

We first evaluate whether a stimulus's own SANE mask is functionally better matched to its fMRI response than plausible controls at $K = 32$. For each stimulus, let c_i^{dyn} denote the DINO cosine obtained with its own dynamic mask and c_i^b the cosine under control b. We analyze the paired difference

$$\Delta_i^{\mathrm{dyn}-b} = c_i^{\mathrm{dyn}} - c_i^b, \qquad b \in \{\mathrm{static}, \mathrm{mismatch}, \mathrm{random}\}. \tag{23}$$

This paired formulation preserves stimulus-level heterogeneity that would be lost by comparing only condition means.

Matched dynamic selection substantially outperforms uniform random parcel subsets (Table 6), with a mean gain of 0.067, a positive median gain of 0.056, and a stimulus-level win rate of 78.3%. Thus, the gate learns non-random parcel structure that is relevant to DINO decoding. More importantly, matched masks outperform dynamic masks selected for other stimuli by 0.008 DINO cosine on average. The median difference is also positive (0.007), the stimulus-level win rate is 57.6%, and the mean difference has the same positive direction in all four subjects. These results demonstrate a measurable correspondence between the selected mask and the current stimulus rather than only a generic preference for useful parcels.

In contrast, matched Dynamic-32 and Static-32 have nearly identical average performance. Static-32 is higher by 0.003 DINO cosine on average, and the Dynamic-32 win rate is 48.6%. The paired distribution is centered close to zero rather than concealing a widespread dynamic advantage: its median is -0.001 and its interquartile range is $[-0.028, 0.024]$. Therefore, the evidence supports a stimulus-specific component in SANE selection, but not a universal average-accuracy advantage over a strong training-derived fixed mask at the relatively broad budget of $K = 32$.

D.2.3 Repeat Consistency of Stimulus-Dependent Selection

The preceding matched-mask result could arise from structured neural evidence or from unstable gate fluctuations. We distinguish these explanations using the three independent fMRI measure-

Table 7: Repeat consistency of stimulus-dependent Top-32 selection. Values are mean ± sample standard deviation across 4,000 subject-stimulus summaries. "Same > Different" is the fraction of stimuli whose mean within-stimulus Jaccard exceeds its mean cross-stimulus Jaccard.

Pair type	Jaccard ↑	Top-K overlap ↑	Same > Different
Same stimulus, different repeats	$\mathbf{0.464 \pm 0.061}$	$\mathbf{20.2 \pm 1.8}$	$\mathbf{78.8\%}$
Different stimuli	0.427 ± 0.026	19.0 ± 0.8	–

ments available for every test image. In this analysis only, we retain a separate trial-level Top-32 set for each repetition instead of averaging gate logits across repetitions. For each image, the same-stimulus score is the mean over its three repetition pairs. The different-stimulus score is the exact mean between its three repetitions and the 2,997 trials belonging to the other 999 images (8,991 comparisons per image). We measure both Jaccard similarity,

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}, \tag{24}$$

and the raw number of shared parcels, $|A \cap B|$. Because the pairwise comparisons are dependent, statistical inference is performed at the stimulus level. As an additional control, we construct 10,000 randomized pseudo-repetition triplets from three distinct stimuli and compare their selection consistency with that of the true repeated measurements.

Repeated measurements of the same image produce more similar parcel sets than measurements of different images (Table 7). Mean Jaccard increases from 0.427 to 0.464, while the mean intersection increases from 19.0 to 20.2 parcels. At the stimulus level, 78.8% of comparisons show higher within-stimulus than between-stimulus Jaccard. The direction is consistent for all four subjects: their same-stimulus Jaccard means are 0.457, 0.467, 0.506, and 0.427, compared with different-stimulus means of 0.422, 0.424, 0.458, and 0.403 for Subjects 01, 02, 05, and 07, respectively. All subject-level 95% bootstrap confidence intervals for the same-versus-different stimulus contrast exclude zero. The randomized pseudo-repetition tests are also significant for every subject and both similarity measures ($p < 10^{-4}$). Hence, the stimulus-dependent component of selection is reproducible across repeated fMRI measurements and cannot be explained solely by random trial-to-trial mask fluctuations.

D.2.4 Effect of Evidence Budget on Stimulus Adaptivity

The similarity between Dynamic-32 and Static-32 suggests that a fixed set of 32 globally informative parcels may already cover much of the evidence needed across stimuli. We therefore perform an inference-time evidence-budget intervention. Each subject's trained $K = 32$ SANE model is held completely fixed, and only the number of tokens supplied to its Transformer and DINO decoder is varied over

$$K \in \{4, 8, 16, 24, 32, 48, 64\}. \tag{25}$$

For every K, Dynamic-K, Static-K, and ten Mismatched Dynamic-K derangements are recomputed using the definitions in Section D.2.1. The same donor derangements are reused across budgets within a subject, so changes with K are not caused by different mismatch pairings. Importantly, no model is retrained or selected at any budget.

Under severe evidence constraints, stimulus-adaptive selection provides a clear and cross-subject-consistent advantage over a fixed global subset (Table 8; Fig. 10). Dynamic exceeds Static by 0.068 at $K = 4$ and by 0.038 at $K = 8$, with stimulus-level win rates of 67.7% and 60.2%, respectively. All four subjects show a positive Dynamic-Static difference at both budgets. The mean advantage falls to 0.002 at $K = 16$ and remains close to zero thereafter. Thus, a fixed global mask becomes sufficient to match average dynamic performance once the retained set is broad enough.

Matched dynamic selection also consistently outperforms mismatched dynamic selection, but its advantage decreases monotonically from 0.023 at $K = 4$ to 0.001 at $K = 64$. This behavior connects the matched-mask and budget results: stimulus identity is most consequential when the model must allocate a scarce evidence budget, whereas many stimulus-specific masks become functionally similar when most parcels are retained. It also explains the strong Static-32 baseline. The choice

Table 8: Effect of inference-time evidence budget on stimulus-adaptive selection. All rows use the same frozen $K = 32$-trained SANE model for each subject; only the number of retained parcel tokens changes. Dynamic and Static are mean DINO cosine across 4,000 subject-stimulus pairs. The two difference columns are stimulus-level paired means; mismatched scores are first averaged over ten derangements.

K	Dynamic	Static	Dyn.-Static	Dyn.-Mismatched	Dyn.>Static
4	0.232	0.164	+0.068	+0.023	67.7%
8	0.288	0.250	+0.038	+0.022	60.2%
16	0.350	0.348	+0.002	+0.016	49.5%
24	0.380	0.384	−0.005	+0.012	47.8%
32	0.395	0.398	−0.003	+0.008	48.6%
48	0.408	0.409	−0.001	+0.004	49.6%
64	0.410	0.410	+0.000	+0.001	50.5%

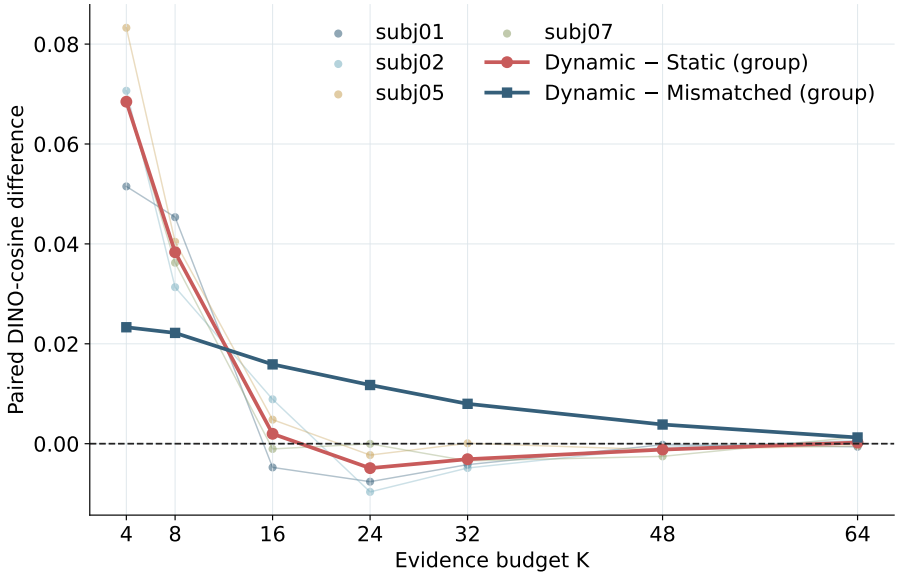


Figure 10: Stimulus adaptivity across evidence budgets. The red curve shows the paired Dynamic-Static DINO cosine difference, with faint curves and points denoting the four individual subjects, while the blue curve shows Dynamic-Mismatched. Dynamic selection provides the largest advantage under tight evidence budgets ($K = 4$ and $K = 8$), whereas both differences progressively approach zero as the evidence budget increases. The horizontal dashed line denotes equal performance.

$K = 32$ is an accuracy-efficiency operating point, not the budget that maximizes the measurable benefit of adaptivity.

Because all $K \neq 32$ settings are inference-time interventions on a model trained at $K = 32$, this experiment isolates sensitivity to the amount of retained evidence. It should not be interpreted as a comparison between separately optimized models trained at different budgets. As a consistency check, the $K = 32$ outputs reproduce the independently computed matched, static, and mismatched stimulus-level scores for every subject to a maximum absolute error below 3.2×10^{-7}.

Dynamic-Static heterogeneity at $K = 32$. The near-zero mean at $K = 32$ does not imply identical predictions for every stimulus. Dynamic selection exceeds Static-32 by more than 0.05 DINO cosine for 11.5% of stimuli, whereas Static-32 exceeds Dynamic by the same margin for 14.0%. The paired distribution is therefore heterogeneous but approximately balanced rather than hiding a general dynamic advantage. We also find no association between the Jaccard similarity of S_i^{32} and the Static-32 mask and the corresponding Dynamic-Static performance difference ($\rho = 0.003$,

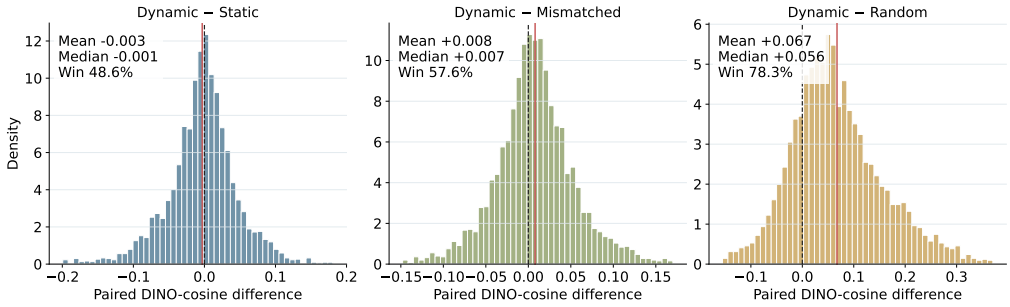


Figure 11: Stimulus-level paired differences between matched Dynamic selection and three controls at $K = 32$. Each panel shows the distribution over 4,000 subject-stimulus pairs. Left: Dynamic-Static differences are centered near zero (mean -0.003, median -0.001, win rate 48.6%), indicating that Dynamic-32 and the training-derived Static-32 mask have similar average decoding performance. Middle: Dynamic-Mismatched is shifted positively (mean $+0.008$, median $+0.007$, win rate 57.6%), showing a modest but consistent advantage when the selection mask is matched to the current stimulus. Right: Dynamic-Random exhibits a substantially larger positive shift (mean $+0.067$, median $+0.056$, win rate 78.3%), confirming that learned SANE selection is far more informative than arbitrary parcel subsets of the same size. Dashed black lines denote zero difference, while solid vertical lines indicate the corresponding distribution means.

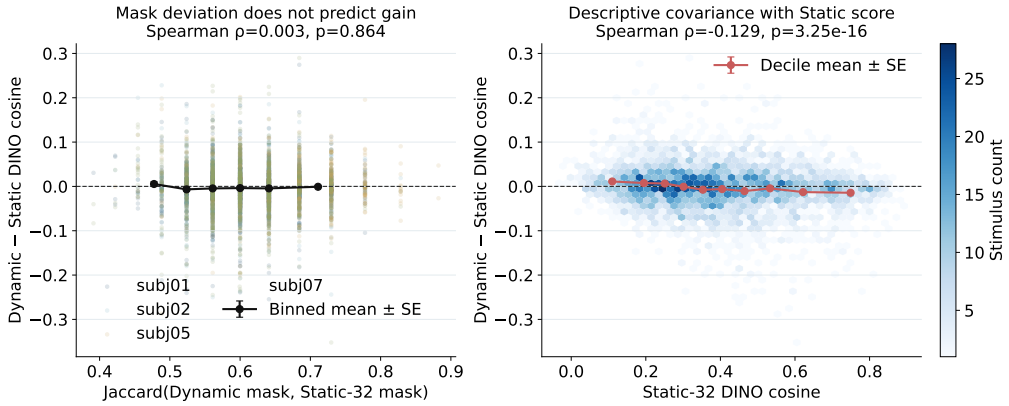


Figure 12: Structure of stimulus-level Dynamic-Static differences at $K = 32$. Left: Each point represents one subject-stimulus pair and compares the Jaccard similarity between its Dynamic-32 mask and the Static-32 mask with its paired DINO-cosine difference. Colored points denote individual subjects, and the black curve reports binned means with standard errors. Mask deviation is not associated with dynamic benefit (Spearman $\rho = 0.003$, $p = 0.864$). Right: Hexagonal bins show the joint distribution of Static-32 DINO cosine and the paired Dynamic-Static difference; the red curve reports decile means with standard errors. The negative association ($\rho = -0.129$, $p = 3.25 \times 10^{-16}$) is descriptive, because the Static-32 score also enters the definition of the vertical-axis difference.

$p = 0.864$). This provides no evidence that simply deviating further from the global mask systematically yields a larger performance gain.

D.2.5 Summary

Together, these analyses identify two complementary components of SANE selection. First, SANE learns a strong global preference for consistently informative parcels, explaining why a training-derived Static-32 mask is competitive under a relatively generous evidence budget. Second, the selected set retains a smaller but reproducible stimulus-dependent component: matched masks outperform masks from other stimuli, repeated measurements of the same stimulus yield more consistent selections than different stimuli, and the advantage of dynamic over static allocation becomes

pronounced when only four or eight parcels can be retained. SANE's stimulus adaptivity should therefore be understood as a mechanism for allocating limited neural evidence to the measurements most relevant to the current fMRI response, rather than as a universal improvement in average decoding accuracy at every budget.

These conclusions concern representation-space DINO decoding by the frozen SANE model. They do not by themselves establish improvements in generated image quality or image-space perceptual metrics, and the inference-only budget intervention does not replace comparisons between separately trained, budget-specific models.

D.3 Stimulus-Specific Evidence Selection and Interaction Patterns

To make SANE's stimulus adaptivity concrete, we visualize 32 held-out examples from NSD Subject 01. These examples were drawn from a diversity-oriented pool of 100 stimuli from the subject's `new_test` split. This pool was constructed using deterministic farthest-point sampling in the cosine space of ℓ_2-normalized DINOv3 image features. We selected 32 examples from this pool for visualization. Their selection was independent of all SANE-derived quantities, including evidence scores, parcel identities, attention weights, and decoding outcomes. The resulting gallery is therefore intended to provide a broad qualitative view of stimulus-adaptive behavior rather than to emphasize particular neural patterns.

Each row has the same four-part organization. Panel (a) shows the visual stimulus. Panel (b) shows the pre-contextualization parcel evidence score a_p, averaged over the three independently measured fMRI repetitions of that stimulus and mapped back to the Subject 01 HCP-MMP1 cortical surface. The score color limits are shared across the complete 100-example batch, so color differences can be compared across rows. Panel (c) highlights the exact hard Top-K evidence set with $K = 32$, obtained from the repetition-averaged parcel scores. Red parcels are retained and gray parcels are measured but not selected. Because scoring occurs before cross-parcel contextualization, every highlighted region remains tied to an identifiable anatomical source.

Panel (d) summarizes interactions after selection. The 32 retained parcels are arranged on a circular layout, with left-hemisphere parcels on the left arc and right-hemisphere parcels on the right arc; within each hemisphere, nodes follow atlas parcel order. Node color encodes the corresponding evidence score. Edge weights are obtained from post-selection Transformer self-attention, averaged over three fMRI repetitions, two Transformer layers, and eight attention heads, and then symmetrized as $(A_{ij} + A_{ji})/2$. For readability, the 128 strongest undirected parcel-parcel interactions are shown, with darker and thicker lines denoting larger attention weight. The box below the graph identifies the six highest-scoring selected parcels. Across stimuli, the changing cortical score maps, retained anatomical subsets, and interaction graphs illustrate that SANE uses a fixed evidence budget while adapting both evidence location and post-selection integration to the measured response.

D.4 Intervention Definitions, Controls, Results, and Interpretation

We performed evidence-removal and cross-stimulus evidence-swap interventions to test whether the stimulus-dependent parcels selected by SANE are merely correlated with decoding performance or carry information that is necessary and content specific. Experiments were conducted independently for Subjects 01, 02, 05, and 07 on the 1,000 held-out test stimuli. All interventions were performed after SANE training and were evaluated with a separately trained, frozen dense HCP-parcel decoder, thereby avoiding circular evaluation through the sparse SANE prediction head; intervention outcomes were never used as SANE training objectives.

For each stimulus, gate logits were averaged across its three fMRI repetitions, and the 32 highest-scoring parcels defined the stimulus-level evidence set S_i. A seed-42 circular shift of a random permutation paired each recipient stimulus i with one non-self donor stimulus j from the same subject; repetition r of the recipient was paired with repetition r of the donor. Predictions from the three repetitions were averaged before evaluation against the 768-dimensional DINOv3 targets. Uniform-random and voxel-count-matched controls were each averaged over ten independently sampled control sets.

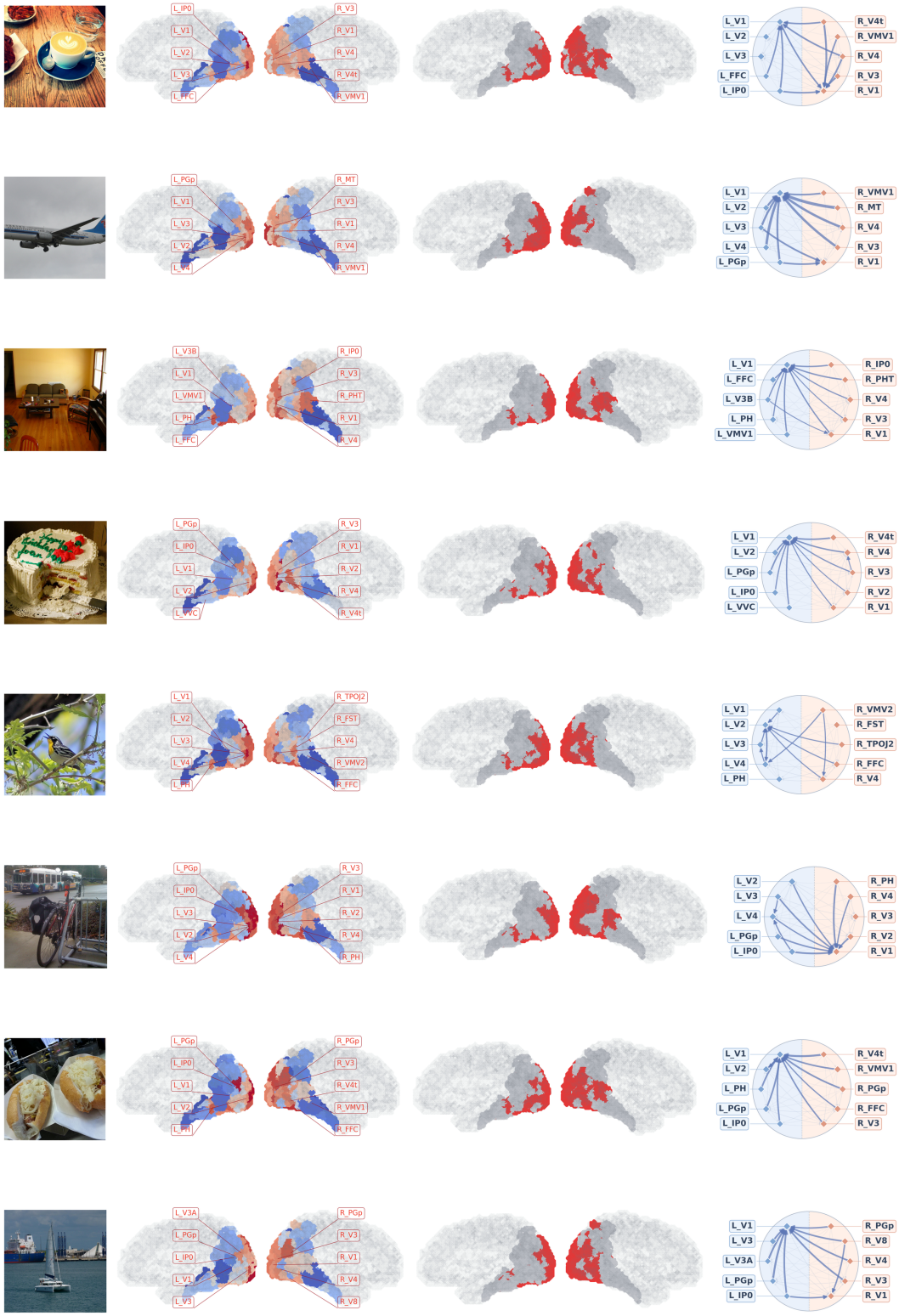


Figure 13: Subject 01 stimulus-specific SANE examples 1–8.

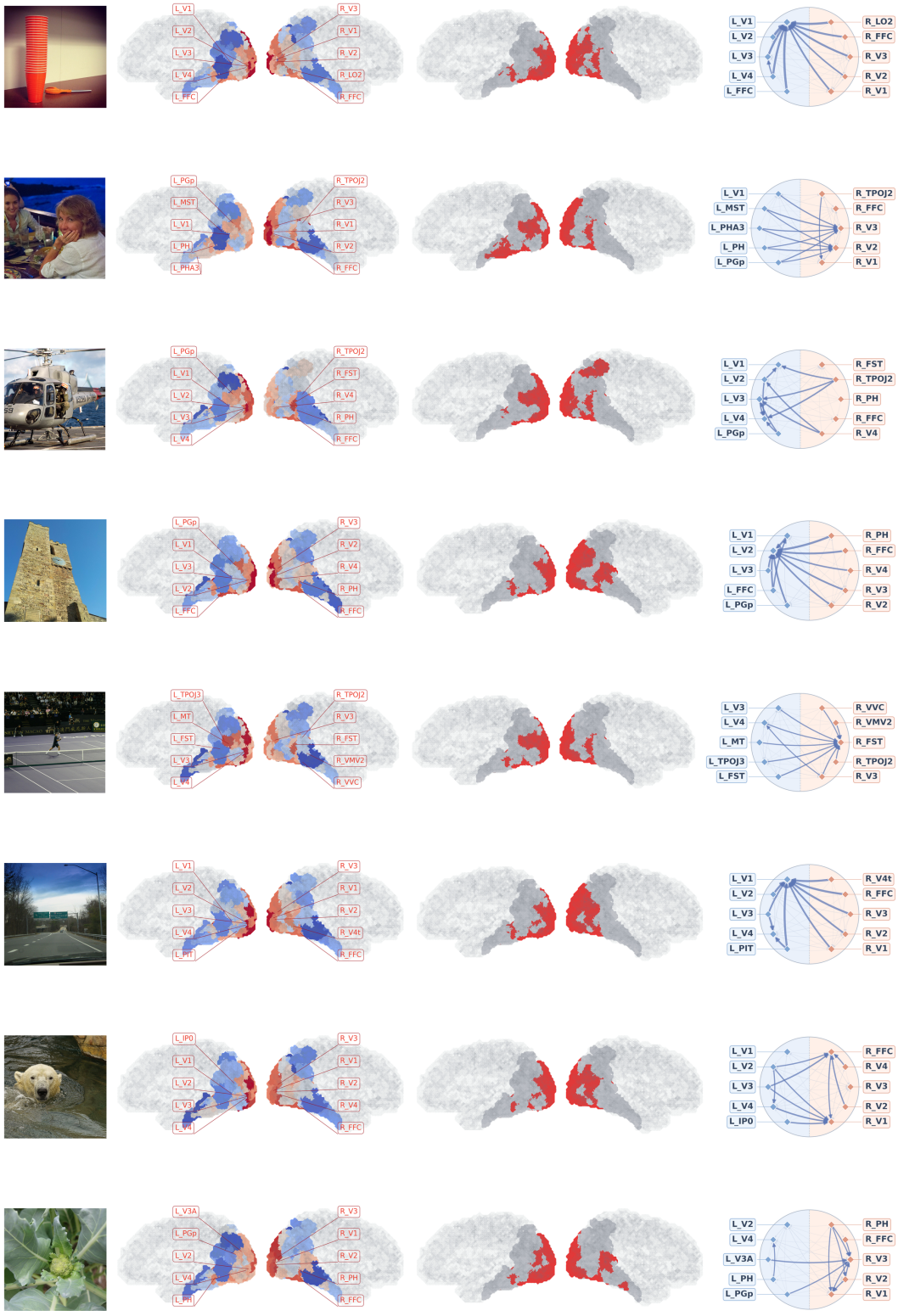


Figure 14: Subject 01 stimulus-specific SANE examples 9–16.

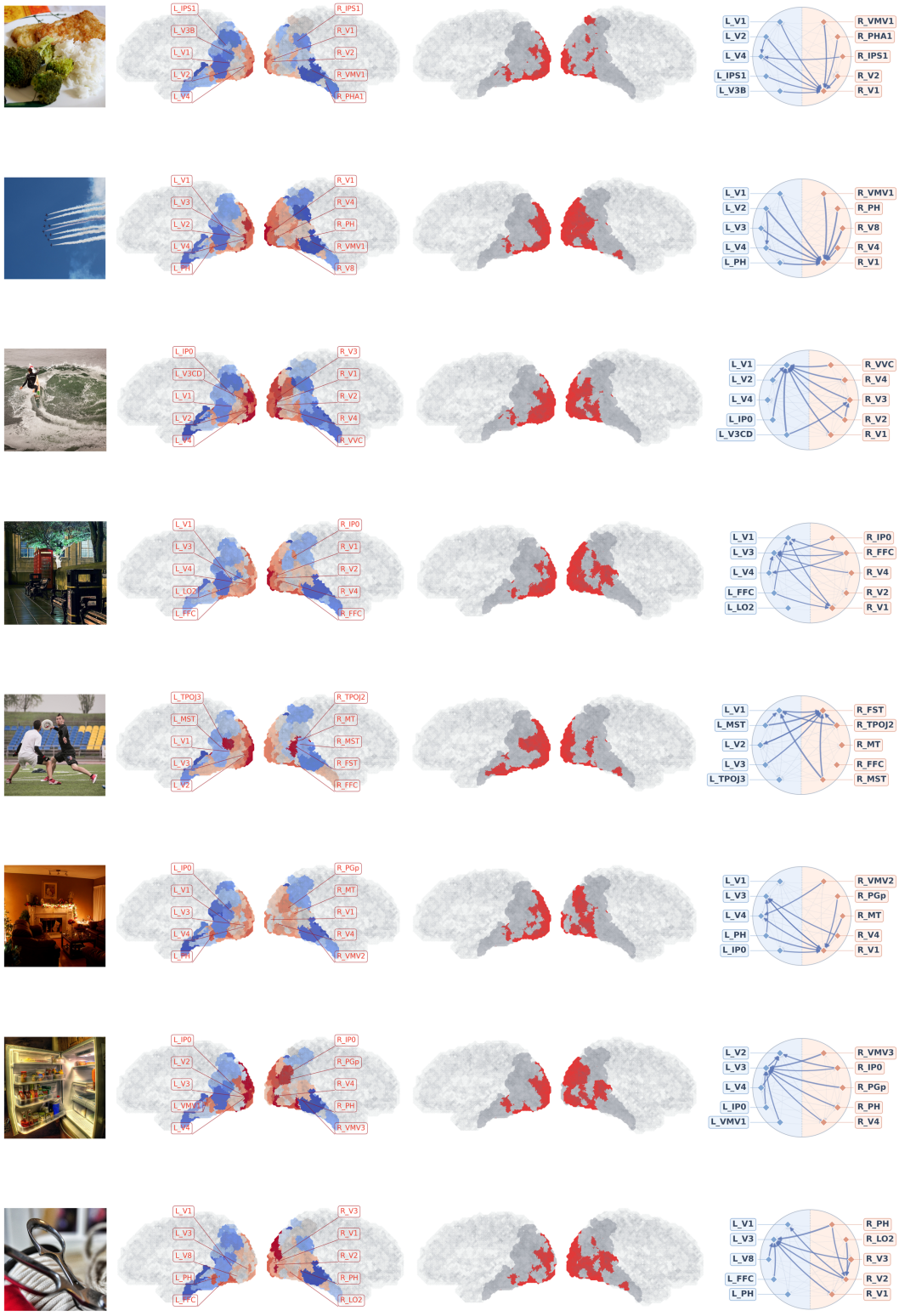


Figure 15: Subject 01 stimulus-specific SANE examples 17–24.

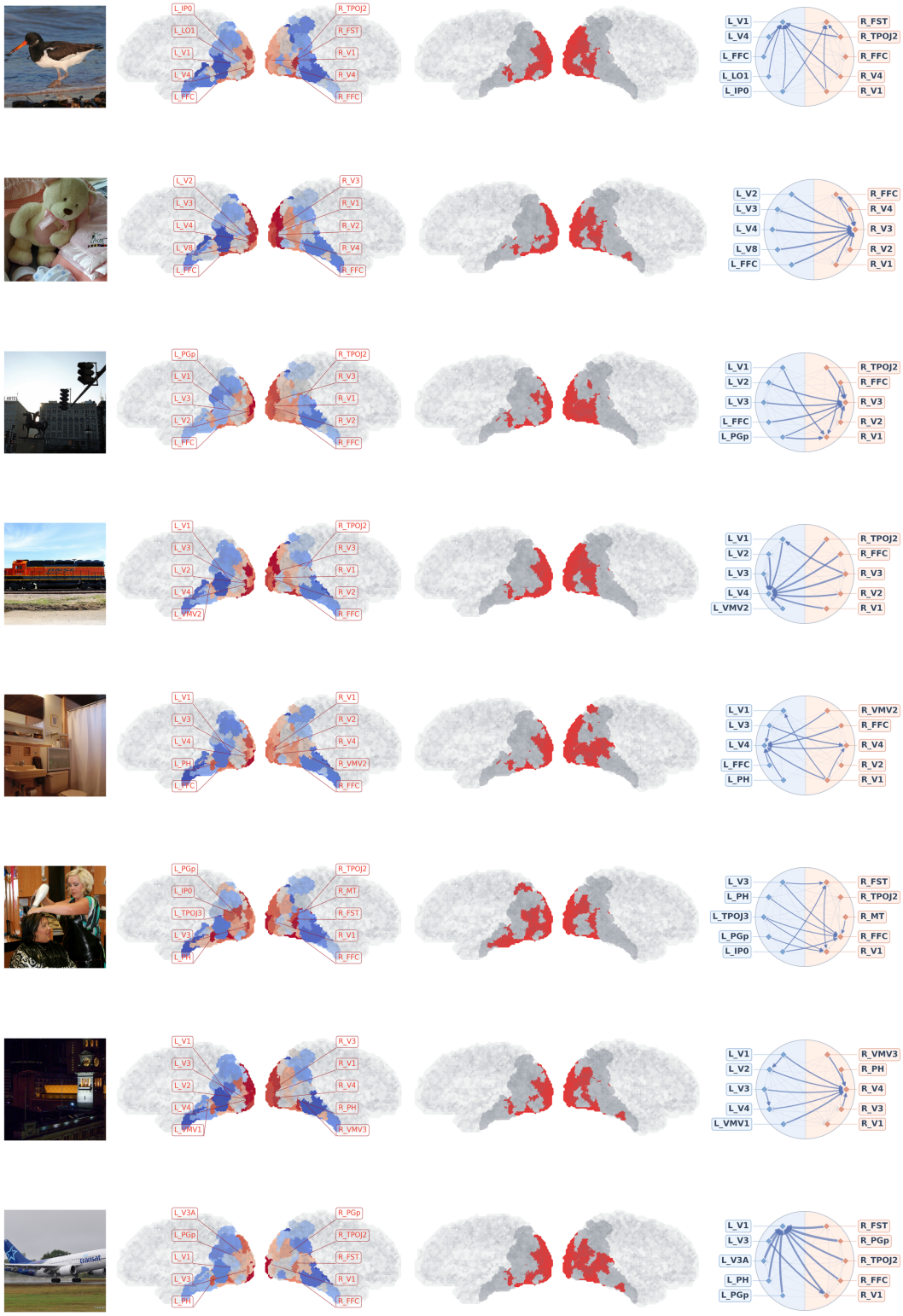


Figure 16: Subject 01 stimulus-specific SANE examples 25–32.

Table 9: Subject-level evidence-removal results at $K = 32$. Full denotes intact dense-decoder cosine similarity; Sel., Rand., and Size denote removal of SANE-selected parcels, uniform-random parcels, and voxel-count-matched random parcels, respectively. Sel. Drop, Rand. Drop, and Size Drop quantify the corresponding decreases from the intact prediction, while Sel.-Rand. and Sel.-Size report the excess degradation caused by removing SANE-selected evidence. T1 and T10 denote 1,000-way Top-1 and Top-10 retrieval accuracy after removal.

Subject	Full	Sel. Rem.	Sel. Drop	Rand. Drop	Size Drop	Sel.-Rand.	Sel.-Size	Sel. T1	Rand. T1	Size T1	Sel. T10	Rand. T10	Size T10
Subj01	0.3975	0.2584	0.1391	0.0318	0.0755	0.1073	0.0636	3.50%	5.06%	4.29%	18.30%	30.19%	23.37%
Subj02	0.3877	0.2368	0.1509	0.0326	0.0810	0.1183	0.0699	2.20%	5.28%	3.66%	12.90%	27.42%	21.34%
Subj05	0.4401	0.2791	0.1610	0.0414	0.1102	0.1196	0.0508	3.90%	8.71%	5.16%	22.30%	40.53%	28.25%
Subj07	0.3808	0.2605	0.1204	0.0360	0.0706	0.0844	0.0498	2.70%	4.21%	3.64%	16.80%	25.03%	22.39%
Mean	**0.4015**	**0.2587**	**0.1428**	**0.0354**	**0.0843**	**0.1074**	**0.0585**	**3.08%**	**5.82%**	**4.19%**	**17.58%**	**30.79%**	**23.84%**

D.4.1 Evidence Removal

For stimulus i, let $\hat{\mathbf{z}}_i$ denote the intact dense-decoder prediction, $\mathbf{z}_i$ its target DINO feature, and $\hat{\mathbf{z}}_i^{-S_i}$ the prediction after physically removing S_i before contextualization. The removal effect is

$$\Delta_{\mathrm{rem},i} = \cos(\hat{\mathbf{z}}_i, \mathbf{z}_i) - \cos(\hat{\mathbf{z}}_i^{-S_i}, \mathbf{z}_i), \tag{26}$$

where a positive value indicates impaired decoding after removal. Selected removal was compared with uniform-random removal of 32 parcels and with parcel-size-matched random removal, the latter controlling more stringently for nonspecific loss of fMRI input because HCP parcels differ in voxel count.

As shown in Table 9, removing the selected parcels reduced cosine similarity by 0.143 ± 0.015 across subjects, compared with 0.035 ± 0.004 for uniform-random removal and 0.084 ± 0.015 for size-matched removal. The selected effect was therefore approximately $4.0\times$ the uniform-random effect and $1.69\times$ the size-matched effect; its excess over the latter was 0.059 ± 0.009, with a strictly positive paired-bootstrap confidence interval in every subject. Retrieval showed the same pattern: after selected removal, mean Top-1 and Top-10 accuracy were 3.08% and 17.58%, compared with 4.19% and 23.84% under size-matched removal. Thus, the selected parcels are disproportionately necessary for accurate decoding beyond effects explained by their number or voxel size.

D.4.2 Full Cross-Stimulus Evidence Swap

We next tested whether the selected parcels carry information about stimulus content rather than only generally useful signal. Let $\mathcal{P}_s = \{1, \ldots, P_s\}$ denote the subject-specific parcel index set and $\mathbf{h}_i(p)$ the local Dense-HCP token of recipient i at parcel p. In the primary donor-selected overwrite (D1), the donor set S_j determines the 32 intervention locations, and the recipient token at each selected anatomical location is replaced by the donor token from that same location:

$$\widetilde{\mathbf{h}}_{i\leftarrow j}^{(1)}(p) = \begin{cases} \mathbf{h}_j(p), & p \in S_j, \\ \mathbf{h}_i(p), & p \notin S_j, \end{cases} \qquad p \in \mathcal{P}_s. \tag{27}$$

The remaining $P_s - 32$ recipient tokens are unchanged, and all P_s tokens enter the same dense evaluator. Donor values are always transferred at the same anatomical parcel index and are never matched to recipient parcels by Top-32 rank. This produces the swapped prediction $\hat{\mathbf{z}}_{i\leftarrow j}$. Movement away from the recipient and toward the donor is quantified as

$$\Delta_{\mathrm{rec},i} = \cos(\hat{\mathbf{z}}_i, \mathbf{z}_i) - \cos(\hat{\mathbf{z}}_{i\leftarrow j}, \mathbf{z}_i), \tag{28}$$

$$\Delta_{\mathrm{don},i} = \cos(\hat{\mathbf{z}}_{i\leftarrow j}, \mathbf{z}_j) - \cos(\hat{\mathbf{z}}_i, \mathbf{z}_j), \tag{29}$$

$$\Delta_{\mathrm{dir},i} = \Delta_{\mathrm{rec},i} + \Delta_{\mathrm{don},i}. \tag{30}$$

A positive directional shift means that the decoded representation moved from recipient i toward donor j. D1 was compared with ten uniform-random and ten parcel-size-matched random swap sets at the same intervention size ($q = 32$).

Full selected-evidence swapping caused a recipient similarity decrease of 0.228 ± 0.031 and a donor similarity increase of 0.224 ± 0.022, yielding a directional shift of 0.452 ± 0.053 (Table 10). Both directions were successful for $92.4\% \pm 2.3\%$ of stimuli. In 1,000-way retrieval, the donor improved

by an average of 381.8 rank positions, whereas the recipient worsened by 249.4 positions. The selected directional shift exceeded the uniform-random and size-matched controls, which yielded shifts of 0.223 ± 0.020 and 0.387 ± 0.053, respectively. The excess over the more stringent size-matched control was 0.065 ± 0.008, with a strictly positive paired-bootstrap confidence interval for each subject. The effect was largest for Subject 05 and smallest for Subject 07, but its direction and statistical support were consistent across all four subjects.

D.4.3 Sensitivity to the Cross-Stimulus Swap Definition

Cross-stimulus substitution can instantiate different counterfactual questions depending on which stimulus determines the intervention locations and whether selected tokens are overwritten or physically removed. We therefore compared three anatomically aligned swap definitions. Definition 1 (D1) is the donor-selected overwrite used in the main intervention analysis, Definition 2 (D2) is a recipient-selected overwrite evaluated in the same full subject-specific Dense-HCP sequence, and Definition 3 (D3) removes the recipient-selected set before inserting the donor-selected set. These variants are not competing algorithms to be ranked by effect magnitude: they differ in intervention support and in the amount of recipient evidence retained. The purpose of this analysis is to test whether donor-directed redirection remains positive under three precisely specified counterfactual constructions.

Common evaluation protocol. The three definitions use the common pairing and repetition-averaging protocol above and the same recipient-drop, donor-gain, and directional-shift metrics. Their subject-specific HCP mappings contain $P_s = 81$, 83, 81, and 82 valid parcels for Subjects 01, 02, 05, and 07, respectively. D1 is the primary donor-selected overwrite defined in Eq. 27; D2 changes only which stimulus selects the overwrite support, whereas D3 changes both the support and the number of tokens entering the evaluator.

Definition 2: recipient-selected overwrite in the full Dense-HCP sequence. The recipient set S_i determined the 32 intervention locations, while the donor's own Top-32 set was not used. Donor content from the corresponding anatomical locations overwrote the recipient tokens at S_i:

$$\widetilde{\mathbf{h}}^{(2)}_{i\leftarrow j}(p) = \begin{cases} \mathbf{h}_j(p), & p \in S_i, \\ \mathbf{h}_i(p), & p \notin S_i, \end{cases} \qquad p \in \mathcal{P}_s. \tag{31}$$

As in Definition 1, the other $P_s - 32$ recipient tokens remained unchanged and all P_s tokens entered the same Dense-HCP evaluator. This definition asks whether donor content redirects the prediction when inserted specifically at the recipient's most strongly selected sites. Its advantage is a direct control of recipient-defined intervention support; its limitation is that the transferred locations need not be those most strongly selected by the donor.

Definition 3: recipient-set removal followed by donor-set insertion. We first removed the recipient set S_i, retained the other recipient tokens, and then inserted the donor's own Top-32 tokens at their native anatomical positions. If a donor-selected position was already present among the retained recipient positions, the donor token overwrote the recipient token. The final support was

$$\mathcal{A}^{(3)}_{i\leftarrow j} = (\mathcal{P}_s \setminus S_i) \cup S_j, \qquad \left|\mathcal{A}^{(3)}_{i\leftarrow j}\right| = P_s - 32 + |S_i \cap S_j|, \tag{32}$$

with active tokens

$$\widetilde{\mathbf{h}}^{(3)}_{i\leftarrow j}(p) = \begin{cases} \mathbf{h}_j(p), & p \in S_j, \\ \mathbf{h}_i(p), & p \in \mathcal{P}_s \setminus (S_i \cup S_j), \end{cases} \qquad p \in \mathcal{A}^{(3)}_{i\leftarrow j}. \tag{33}$$

This definition provides an explicit transfer of the donor-selected set while preserving recipient evidence outside the recipient-selected set. Its trade-off is that it combines token substitution with physical removal and produces a variable-length evaluator input, so its effect cannot be attributed to overwrite alone.

Results and scope. All three definitions produced positive recipient-to-donor redirection for every subject (Table 10). Relative to D1's directional shift of 0.452 ± 0.053 reported above, D2 yielded a closely comparable 0.460 ± 0.055, comprising a recipient decrease of 0.243 ± 0.030 and a donor increase of 0.217 ± 0.025. Its subject-level shifts were 0.451, 0.446, 0.548, and 0.397 for Subjects 01, 02, 05, and 07, showing that the effect was not driven by one subject. D3 yielded a larger

Table 10: Subject-level comparison of three cross-stimulus swap definitions at $K = 32$ and $q = 32$. D1 is donor-selected overwrite in the full subject-specific Dense-HCP sequence, D2 is recipient-selected overwrite in the same full sequence, and D3 removes the recipient Top-32 before inserting the donor Top-32 at their native anatomical positions. In D1 and D2, donor content replaces recipient content at the selected anatomical positions while all nonselected recipient tokens remain unchanged. Rec. Drop, Don. Gain, and their sum (Dir. Shift) quantify donor-directed redirection over $N = 1000$ test stimuli. Input denotes the number of tokens entering the evaluator and reflects the subject-specific parcel count; D3 additionally varies with recipient-donor Top-32 overlap. D1 uses the original values reported in the main intervention analysis. The three definitions evaluate different counterfactual interventions, so effect magnitude is not a method-quality ranking.

Def.	Subject	Base Rec.	Swap Rec.	Rec. Drop	Base Don.	Swap Don.	Don. Gain	Dir. Shift	Succ.	Input
D1	Subj01	0.3975	0.1804	0.2170	0.0909	0.3148	0.2239	0.4410	93.1%	81
	Subj02	0.3877	0.1663	0.2214	0.0878	0.3078	0.2200	0.4414	92.8%	83
	Subj05	0.4401	0.1606	0.2795	0.0766	0.3335	0.2569	0.5364	95.0%	81
	Subj07	0.3808	0.1862	0.1946	0.0899	0.2857	0.1958	0.3904	88.7%	82
	Mean	**0.4015**	**0.1734**	**0.2281**	**0.0863**	**0.3105**	**0.2242**	**0.4523**	**92.4%**	**81.8**
D2	Subj01	0.3975	0.1632	0.2343	0.0909	0.3073	0.2164	0.4507	92.2%	81
	Subj02	0.3877	0.1541	0.2336	0.0878	0.2998	0.2120	0.4456	92.0%	83
	Subj05	0.4401	0.1472	0.2929	0.0766	0.3312	0.2547	0.5476	95.3%	81
	Subj07	0.3808	0.1692	0.2116	0.0899	0.2750	0.1851	0.3967	88.4%	82
	Mean	**0.4015**	**0.1584**	**0.2431**	**0.0863**	**0.3033**	**0.2170**	**0.4601**	**92.0%**	**81.8**
D3	Subj01	0.3975	0.1357	0.2618	0.0909	0.3483	0.2575	0.5193	94.5%	70.7
	Subj02	0.3877	0.1249	0.2628	0.0878	0.3421	0.2543	0.5171	93.9%	72.9
	Subj05	0.4401	0.1183	0.3219	0.0765	0.3753	0.2988	0.6206	96.7%	71.8
	Subj07	0.3808	0.1397	0.2411	0.0899	0.3185	0.2286	0.4697	90.8%	71.4
	Mean	**0.4015**	**0.1297**	**0.2719**	**0.0863**	**0.3461**	**0.2598**	**0.5317**	**94.0%**	**71.7**

descriptive shift of 0.532 ± 0.055, comprising a recipient decrease of 0.272 ± 0.030 and a donor increase of 0.260 ± 0.025. Both directions were successful for $92.4\% \pm 2.3\%$, $92.0\% \pm 2.4\%$, and $94.0\% \pm 2.1\%$ of stimuli for D1–D3, respectively.

D1 and D2 both passed the complete subject-specific parcel sequence to the same dense evaluator: 81, 83, 81, and 82 tokens for Subjects 01, 02, 05, and 07, corresponding to 81.75 ± 0.83 tokens across subjects. D3 instead passed mean sequence lengths of 70.705, 72.918, 71.807, and 71.363 tokens, respectively, or 71.70 ± 0.81 across subjects, because its support depends on $|S_i \cap S_j|$. The small difference between the D1 and D2 mean directional shifts (0.008) shows that donor-directed redirection persists whether the intervention support is selected by the donor or the recipient. D3 produces a larger numerical shift, but it also removes recipient-selected positions and changes the evaluator input length. Its magnitude therefore cannot be compared with D1 and D2 as if all three definitions applied the same intervention. The robust conclusion is the shared positive redirection across subjects and definitions, while each definition answers a different counterfactual question.

D.4.4 Interpretation and Statistical Reporting

Interpretation. The two interventions provide complementary evidence. Removal establishes importance: deleting the SANE-selected evidence harms decoding substantially more than matched control removals. Swap establishes content specificity: replacing those tokens with donor evidence predictably redirects both cosine similarity and retrieval ranking toward the donor. The size-matched controls rule out intervention magnitude and parcel voxel count as sufficient explanations for these effects. Together, the results support a stimulus-dependent neural evidence set with a model-level counterfactual effect on the decoded DINO representation. This conclusion is restricted to representation-space decoding and retrieval; the present intervention does not evaluate generated images or image-space CLIP similarity.

Statistical reporting. Within each subject, 95% confidence intervals were obtained using 10,000 paired bootstrap resamples over the 1,000 test stimuli. Random-control metrics were first averaged over ten independently sampled control masks. Cross-subject values are reported as the unweighted

mean ± population standard deviation across the four subjects and should not be interpreted as a group-level confidence interval.

E Backbone Transfer and Reconstruction Analyses

E.1 Backbone-Agnostic Evidence Transfer

To evaluate whether the learned neural evidence remains useful beyond the decoder used to train SANE, we freeze the trained SANE model and transfer its selected evidence representation to the native conditioning space of each pretrained reconstruction backbone. Let $\boldsymbol{E}^{(r)}_{s,i,K} \in \mathbb{R}^{K\times d}$ denote the contextualized representation of the K selected parcels for stimulus i and subject s. For backbone m, let $\boldsymbol{C}^{(m)}$ denote its native neural conditioning representation. Because these representations generally have different dimensions and structures, we introduce a lightweight, backbone-specific adapter A_m that maps the SANE evidence representation into the corresponding conditioning space:

$$\widehat{\boldsymbol{C}}^{(m)} = A_m\left(\boldsymbol{E}^{(r)}_{s,i,K}\right). \tag{34}$$

For each backbone, the target conditioning representation $\boldsymbol{C}^{(m)}$ is obtained from its native neural pathway using the same training stimulus. These representations are treated as fixed supervision targets, and the adapter is optimized with a backbone-specific alignment objective:

$$\mathcal{L}^{(m)}_{\text{align}} = D_m\left(\widehat{\boldsymbol{C}}^{(m)}, \boldsymbol{C}^{(m)}\right), \tag{35}$$

where D_m is chosen according to the structure of the corresponding conditioning space. The specific conditioning targets, adapter parameterizations, and alignment objectives for each backbone are provided below.

During adapter training, both the SANE evidence pathway and the pretrained reconstruction backbone are frozen; only the adapter A_m is optimized. After alignment, the predicted conditioning representation is provided to the unchanged downstream reconstruction pipeline:

$$\widehat{I}^{(m)} = G^{\text{frozen}}_m\left(\widehat{\boldsymbol{C}}^{(m)}\right), \tag{36}$$

where G^{frozen}_m denotes the pretrained downstream reconstruction and generation components. Thus, the transfer experiment changes only the neural representation supplied to the backbone while leaving its pretrained downstream processing unchanged.

We apply this procedure to five pretrained reconstruction backbones: MindEye2, NeuroPictor, NeuroAdapter, NeuroFlow, and Brain-IT. This experiment therefore evaluates the portability of SANE evidence across heterogeneous neural conditioning interfaces: the same frozen, stimulus-adaptive evidence representation is translated through a lightweight adapter and used to condition different pretrained reconstruction systems.

E.1.1 Experimental Protocol

Dataset and preprocessing. We use the Natural Scenes Dataset (NSD) for Subjects 1, 2, 5, and 7. Each subject has responses to 10,000 natural images, with each image presented three times, yielding 30,000 single-trial fMRI responses per subject. We use the provided single-trial beta estimates at 1.8-mm isotropic resolution. Voxel responses are standardized within each session before being organized according to the HCP-MMP1 visual-cortex parcellation.

The 9,000 subject-specific images are used for training and validation, while the official 1,000 images shared across subjects are held out exclusively for final testing. Thus, the test set contains 3,000 single-trial responses per subject. During evidence learning, the three repetitions of each image are kept as separate fMRI trials rather than averaged, allowing SANE to learn evidence from individual measured responses.

Evidence budgets and model training. We train a separate subject-specific model for each evidence budget K. Unless otherwise stated, the main experiments use $K = 32$. The dense decoder is first trained using all available parcels for each subject. It is trained for 20 epochs with batch

size 128 using AdamW, a learning rate of 10^{-4}, and weight decay of 10^{-4}. The best validation checkpoint is then used to initialize the parcel encoder, contextual encoder, and prediction head of SANE, while the evidence scorer is newly initialized.

SANE is subsequently trained for 20 epochs. The evidence gate uses a learning rate of 10^{-4}, while the inherited modules use a learning rate of 2×10^{-5}. Models are trained in FP32, and checkpoints are selected solely according to validation cosine similarity. The held-out test set is never used for checkpoint selection or hyperparameter tuning.

Inference and DINOv3 evaluation. At validation and test time, SANE uses exact hard Top-K selection rather than the soft relaxation used during training. For feature-space evaluation, the three fMRI repetitions of each test stimulus are decoded independently. Their predicted DINOv3 representations are then averaged and re-normalized before cosine similarity with the corresponding target representation is computed.

Unless otherwise stated, reported DINOv3 results are macro-averaged across Subjects 1, 2, 5, and 7, giving each subject equal weight. When mean $\pm$ standard deviation is reported for these results, the variability is computed across the four subject-level results rather than across individual stimuli.

Image reconstruction evaluation. For reconstruction experiments, the three repetitions are combined according to the native conditioning interface of each reconstruction backbone to produce one conditioning representation per stimulus. We evaluate reconstruction using pixel correlation (PixCorr), structural similarity (SSIM), and two-way identification accuracy with AlexNet-2, AlexNet-5, Inception, and CLIP. We additionally report correlation distances in the EfficientNet-B1 and SwAV feature spaces. Higher values indicate better performance for PixCorr, SSIM, and the identification metrics, whereas lower correlation distance indicates better performance for EfficientNet-B1 and SwAV.

For comparisons between native and SANE-driven versions of a reconstruction backbone, both variants use the same test stimuli, frozen downstream reconstruction and generation components, post-processing, metric implementations, sampling settings, and prespecified generation seeds. Thus, differences are attributable to the neural conditioning representation rather than downstream reconstruction settings.

Intervention evaluation. All intervention experiments use the independently trained frozen dense decoder rather than the SANE decoder. For evidence removal, the measurements at the recipient's SANE-selected parcels are removed and the perturbed response is decoded with the same dense decoder. We compare this intervention with uniformly random removal and a voxel-count-matched random control. The latter matches the total number of voxels removed by the SANE-selected parcels, accounting for differences in parcel size.

For cross-stimulus substitution, measurements at the donor's selected parcel locations are replaced by the anatomically corresponding measurements from a donor stimulus, while all remaining recipient measurements are kept unchanged. Random substitution and voxel-count-matched controls are evaluated using the same test stimuli. Random controls are repeated and averaged before comparison.

Statistical reporting. Unless otherwise specified, all reported results are first computed separately for each subject and then macro-averaged across the four subjects. Statistical uncertainty and significance testing are performed at the subject or stimulus level according to the corresponding analysis, with the complete resampling and testing procedures reported alongside the relevant experiments.

E.2 Backbone-Specific Alignment and Reconstruction Details

This section gives the implemented interface for each reconstruction system. For a subject with P retained HCP-MMP1 parcels, frozen SANE produces the contextual evidence

$$\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_{32}] \in \mathbb{R}^{32 \times 512}, \tag{37}$$

together with the corresponding parcel indices $\mathbf{p}$ and sorted gate logits $\mathbf{s}$. The subject-specific parcel counts are $P = 81, 83, 81,$ and 82 for Subjects 1, 2, 5, and 7, respectively. In every integration, the SANE tokenizer, evidence gate, and contextual Transformer are loaded from the subject-specific $K = 32$ checkpoint, run with exact hard Top-K selection, placed in evaluation mode, and excluded from optimization. The native teacher encoders and all reconstruction modules downstream of the

stated replacement boundary are also frozen. Only the adapter parameters described below are updated.

For compactness, let

$$d_{\mathrm{tok}}(\mathbf{A},\mathbf{B}) = 1 - \frac{1}{BT}\sum_{b=1}^{B}\sum_{t=1}^{T}\cos(\mathbf{A}_{b,t},\mathbf{B}_{b,t}), \tag{38}$$

and let d_{flat} denote cosine distance after flattening all non-batch dimensions. We use $\mathrm{LN}(\cdot)$ below for non-affine layer normalization over the last feature dimension.

E.2.1 MINDEYE2

Semantic branch. The teacher target is produced by the frozen subject-specific MindEye2 ridge layer followed by `BrainNetwork.backbone`; the boundary is the unprojected tensor immediately before `clip_proj` and `diffusion_prior`, with shape 256×1664. The adapter layer-normalizes the 32×512 SANE sequence and cross-attends 256 learned 512-dimensional queries to it using two pre-norm Transformer-decoder layers (8 heads, feed-forward width 2048, GELU, dropout 0.1). A final layer normalization and linear map $512 \to 1664$ give the predicted backbone tensor.

The code estimates one scalar mean μ and standard deviation σ over all elements of the training targets, rather than separate statistics per token or feature. With $\widetilde{\mathbf{C}} = (\mathbf{C} - \mu)/\sigma$, the semantic objective is

$$\mathcal{L}_{\mathrm{ME2,sem}} = \mathrm{MSE}(\widehat{\widetilde{\mathbf{C}}}, \widetilde{\mathbf{C}}) + 0.1\, d_{\mathrm{tok}}(\widehat{\widetilde{\mathbf{C}}}, \widetilde{\mathbf{C}}). \tag{39}$$

The prediction is transformed back with $\widehat{\mathbf{C}} = \sigma\widehat{\widetilde{\mathbf{C}}} + \mu$ before entering the frozen MindEye2 tail.

Low-level branch. The low-level target is the deterministic mode of the frozen MindEye2 image-variation autoencoder applied to the ground-truth training image, multiplied by the native scale factor 0.18215. Its shape is $4 \times 28 \times 28$. The readout uses 49 learned spatial queries arranged as a 7×7 grid and the same two-layer, 8-head, width-2048 decoder specification as the semantic adapter. The decoded queries are projected to 64 channels, reshaped to $64 \times 7 \times 7$, and converted to $4 \times 28 \times 28$ by two bilinear $\times 2$ upsampling stages interleaved with convolution, GroupNorm, and GELU blocks. It is optimized only with

$$\mathcal{L}_{\mathrm{ME2,low}} = \left\| \widehat{\mathbf{Z}} - \mathbf{Z} \right\|_1. \tag{40}$$

The two heads are trained separately with AdamW (10^{-4} learning rate and 10^{-4} weight decay) on a stimulus-disjoint 90/10 training/validation partition. At inference, every fMRI repetition is passed independently through the nonlinear SANE path and both heads; the resulting 256×1664 tensors and $4 \times 28 \times 28$ latents are then averaged across repetitions. The semantic tensor is consumed by the frozen `clip_proj` and diffusion prior, whereas the low-level latent is decoded by the frozen image-variation autoencoder. The reported final MindEye2 image follows the implemented official path: 20 prior steps, 38 unCLIP steps, 25 enhancement steps (image-to-image time point 13 and guidance scale 5.0), followed by 0.75 enhanced image $+ 0.25$ blurry reconstruction.

E.2.2 NEUROPICTOR

The replacement boundary is the output of NeuroPictor's subject-specific surface-fMRI encoder, a 257×1024 latent used by both its semantic and low-level control branches. The adapter first maps each contextual SANE token from 512 to 1024 dimensions. It then adds three learned terms: an embedding of the selected HCP-MMP1 parcel identity, an embedding of its sorted Top-K rank, and a two-layer MLP embedding of its scalar gate logit ($1 \to 64 \to 1024$). After layer normalization, 257 learned 1024-dimensional queries cross-attend to this sequence through two pre-norm decoder layers with 16 heads, feed-forward width 4096, GELU, and dropout 0.1. The final layer-normalized query sequence is the predicted native latent.

Teacher latents are cached from the frozen subject-matched NeuroPictor fMRI encoder. The frozen `FMRIToSemanticModel` and `FMRIToControlModel` weights are extracted from the same NeuroPictor checkpoint and applied to both teacher and predicted latents. Writing the shared latent,

semantic output, and first ControlNet hint as $\mathbf{H}$, $\mathbf{S}$, and $\mathbf{Q}$, respectively, the implemented objective is

$$\begin{aligned}\mathcal{L}_{\mathrm{NP}} = \; & \mathrm{MSE}(\widehat{\mathbf{H}}, \mathbf{H}) + 0.1 d_{\mathrm{tok}}(\widehat{\mathbf{H}}, \mathbf{H}) + 0.25\,\mathrm{MSE}(\mathrm{LN}(\widehat{\mathbf{H}}), \mathrm{LN}(\mathbf{H})) \\ & + 0.25\,\mathrm{MSE}(\mathrm{LN}(\widehat{\mathbf{S}}), \mathrm{LN}(\mathbf{S})) + 0.1 d_{\mathrm{flat}}(\widehat{\mathbf{S}}, \mathbf{S}) \\ & + \mathrm{MSE}(\widehat{\mathbf{Q}}, \mathbf{Q}) + 0.1 d_{\mathrm{flat}}(\widehat{\mathbf{Q}}, \mathbf{Q}). \end{aligned} \tag{41}$$

Thus, the semantic and control terms are not additional trainable reconstruction modules: gradients pass through the frozen projections only to the latent adapter.

Both native-teacher and SANE repetitions are encoded separately and averaged at the 257×1024 boundary. Subject-specific adapters are fitted for at most 50 epochs with AdamW (batch size 16, learning rate and weight decay 10^{-4}), and checkpoint selection uses the total loss on a stimulus-disjoint 5% validation set. At test time, the averaged predicted latent replaces the native encoder output and is supplied unchanged to the frozen semantic conditioner, ControlNet, latent-diffusion model, and image decoder. Reconstruction uses the same 50-step DDIM sampler, guidance scale 5.0, and $\eta = 0$ as the corresponding native run.

E.2.3 NeuroAdapter

The evaluated native NeuroAdapter first selects 100 Schaefer parcels per hemisphere, producing 200 parcel tokens, but its Transformer guidance decoder maps those tokens to only *50* condition tokens of dimension 768. SANE replaces the complete surface-fMRI-to-guidance path at this 50×768 output boundary; the 200 native parcel tokens are therefore teacher-side inputs, not the SANE adapter output.

For SANE, a learned linear map projects each 512-dimensional evidence token to 768 dimensions. A learned positional embedding is indexed by the token's actual HCP-MMP1 parcel ID and added to the projected token. Gate scores and Top-K ranks are not supplied to this adapter (the scores are saved only for diagnostics). The enriched sequence is consumed by NeuroAdapter's one-layer query-decoder specification: 50 learned queries, 8 attention heads, feed-forward width 1024, and dropout 0.1. The query embeddings and decoder weights are initialized from the frozen subject-matched native guidance teacher; the HCP parcel embeddings remain separate because the native teacher uses Schaefer parcels.

The teacher is the released `transformer_decoder` configuration and returns raw 50×768 guidance tokens. Repetitions are averaged in their respective native fMRI spaces before either teacher or SANE encoding. With no diffusion-loss term enabled in the evaluated checkpoints, the exact distillation loss is

$$\mathcal{L}_{\mathrm{NA}} = \mathrm{MSE}(\widehat{\mathbf{G}}, \mathbf{G}) + 0.1 d_{\mathrm{tok}}(\widehat{\mathbf{G}}, \mathbf{G}). \tag{42}$$

The raw-token MSE is retained because the frozen image projection consumes token magnitude as well as direction. The adapter is optimized in FP32 with AdamW (50-epoch maximum, batch size 16, learning rate 10^{-4}, weight decay 10^{-6}); only the projected SANE guidance module is in the optimizer.

For reconstruction, the teacher checkpoint's image-projection weights and IP-attention processors are copied exactly and then frozen together with the Stable Diffusion 1.5 UNet, VAE, and empty-text encoder state. The predicted 50-token sequence is passed through this frozen path and sampled with the same 50-step DDPM schedule and guidance scale 4.0 used for the native/SANE pair.

E.2.4 NeuroFlow

NeuroFlow is replaced at the deterministic output of its frozen subject-specific NeuroVAE, obtained in code with `encode(..., sample=False)[0]`. This interface has shape 256×1664. The SANE adapter adds a learned HCP parcel-ID embedding and a gate-score MLP embedding ($1 \to 64 \to 512$) to each contextual 512-dimensional evidence token. It layer-normalizes the resulting memory, cross-attends 256 learned queries using two pre-norm decoder layers (8 heads, feed-forward width 2048, GELU, dropout 0.1), and applies a final layer normalization and $512 \to 1664$ projection.

SANE-native and NeuroFlow-native responses are paired by the global zero-based NSD stimulus ID. All repetitions are averaged in fMRI space before the frozen SANE or NeuroVAE encoder. The loss applied to the predicted and teacher latents is

$$\mathcal{L}_{\mathrm{NF}} = \mathrm{MSE}(\widehat{\mathbf{F}}, \mathbf{F}) + \mathrm{MSE}(\mathrm{LN}(\widehat{\mathbf{F}}), \mathrm{LN}(\mathbf{F})) + 0.1 d_{\mathrm{tok}}(\widehat{\mathbf{F}}, \mathbf{F}). \tag{43}$$

The subject-specific adapters are fitted with AdamW for at most 100 epochs (batch size 16, learning rate and weight decay 10^{-4}), using a stimulus-disjoint 10% validation split for checkpoint selection.

At inference, only SANE-native fMRI is read. The 256×1664 prediction is passed to the frozen 12-layer backward SiT and then to NeuroFlow's frozen SDXL unCLIP decoder; the direct backward Euler path uses 20 steps. The target-image-derived `f2i` diagnostic path in the original repository is never called.

E.2.5 BRAIN-IT

Brain-IT has independently trained semantic and low-level graph decoders. In each decoder, the native front-end embeds surface voxels, appends 128 learned centers, applies the voxel-to-center `TransformerConv`, and returns layer-normalized ReLU center tokens of shape 128×512. We replace precisely this voxel-to-center computation. The native center-to-output cross-attention blocks and output projection are retained. For the semantic decoder, the stage-2 GNN weights stored in the combined reconstruction checkpoint are overlaid before teacher extraction, so adapter fitting and final inference use the same semantic tail.

Two separate center adapters consume the same SANE evidence. Each adds learned HCP parcel-ID and gate-score embeddings ($1 \to 64 \to 512$), layer-normalizes the memory, and cross-attends 128 learned center queries through two pre-norm decoder layers with 8 heads, feed-forward width 2048, GELU, and dropout 0.1. The adapters do not share weights because the semantic and low-level centers come from separately trained Brain-IT latent spaces. Their frozen tails produce 256×1664 semantic tokens and 7190×512 VGG tokens, respectively.

SANE repetitions are averaged before SANE, and the resulting vector is paired with Brain-IT's native teacher response using the global NSD stimulus identity. For branch $r \in \{\mathrm{sem}, \mathrm{low}\}$, with center tensors $\mathbf{U}_r$ and final tokens $\mathbf{T}_r$, the loss is

$$\mathcal{L}_r = \mathrm{MSE}(\widehat{\mathbf{U}}_r, \mathbf{U}_r) + 0.25\,\mathrm{MSE}(\mathrm{LN}(\widehat{\mathbf{T}}_r), \mathrm{LN}(\mathbf{T}_r)) + 0.1 d_{\mathrm{tok}}(\widehat{\mathbf{T}}_r, \mathbf{T}_r), \tag{44}$$

and $\mathcal{L}_{\mathrm{BIT}} = \mathcal{L}_{\mathrm{sem}} + \mathcal{L}_{\mathrm{low}}$. All 128 centers and all semantic tokens are supervised. To bound training memory, each training batch samples 1,024 of the 7,190 low-level token positions; validation uses one fixed 1,024-position subset. The two adapters are jointly fitted for at most 20 epochs with AdamW (batch size 4, learning rate and weight decay 10^{-4}), while both teacher decoders remain frozen.

At inference the adapters predict both complete center sequences, and the frozen tails generate the full semantic and VGG tensors. The latter drive Brain-IT's original Deep Image Prior low-level reconstruction; the semantic tokens and low-level image then enter the original stage-2 diffusion path (start step 14), followed by the unchanged SDXL enhancement stage.

E.2.6 CONTROLLED INTEGRATION PROTOCOL

The controlled comparison is defined within each backbone. Its native and SANE-driven variants use the same 1,000 shared NSD test stimuli in the same order, the same three generation seeds (42, 3407, and 10007), the same pretrained downstream checkpoint, sampler and guidance settings, final image type, post-processing, and metric implementation. No best-of-N image selection is performed. Accordingly, within a backbone the changed component is the native neural-to-conditioning path versus frozen SANE plus the adapter specified above.

Adapters and normalization statistics are fitted using training stimuli only and selected using validation stimuli. The shared 1,000-image test set is excluded from gradient updates, normalization-statistics estimation, and checkpoint selection. When the native teacher and SANE use different fMRI layouts (for example, surface vertices versus SANE's native `nsdgeneral` voxel vector), the two arrays are paired solely by NSD stimulus identity; native fMRI is used only to compute a frozen teacher target and is not supplied to the SANE-driven reconstruction at inference.

Table 11: Published and subject-level reconstruction results. We report published results from representative prior reconstruction methods, followed by subject-level native and SANE-driven results across five reconstruction backbones for NSD Subjects 1, 2, 5, and 7. Higher values are better for PixCorr, SSIM, AlexNet, Inception, and CLIP, while lower values are better for EfficientNet and SwAV. For the subject-level results, ± denotes the mean per-stimulus standard deviation across the three prespecified generation seeds.

		Low-Level				High-Level			
Method	**Subject**	**PixCorr ↑**	**SSIM ↑**	**Alex(2) ↑**	**Alex(5) ↑**	**Incep ↑**	**CLIP ↑**	**Eff ↓**	**SwAV ↓**
Published results from prior reconstruction methods									
MindEye (NeurIPS 2023)	–	0.309	0.323	94.7%	97.8%	93.8%	94.1%	0.645	0.367
Brain-Diffuser (Sci. Rep. 2023)	–	0.254	0.356	94.2%	96.2%	87.2%	91.5%	0.775	0.423
Takagi (CVPR 2023)	–	0.246*	0.410*	83.0%	83.0%	76.0%	77.0%	0.811*	0.504*
DREAM (WACV 2024)	–	0.288	0.338	95.0%	97.5%	94.8%	95.2%	0.638	0.413
UMBRAE (ECCV 2024)	–	0.283	0.341	95.5%	97.0%	91.7%	93.5%	0.700	0.393
NeuroVLA (NeurIPS 2024)	–	0.265	0.357	93.1%	97.1%	96.8%	97.5%	0.633	0.321
MindBridge (CVPR 2024)	–	0.151	0.263	87.7%	95.5%	92.4%	94.7%	0.712	0.418
MindTuner (AAAI 2025)	–	0.322	0.421	95.8%	98.8%	95.6%	93.8%	0.612	0.340
Controlled native and SANE-driven evaluation by subject									
MindEye2†	Subject 1	0.326 ± 0.061	0.378 ± 0.020	98.1% ± 1.4%	99.2% ± 0.6%	95.8% ± 2.7%	93.3% ± 3.3%	0.613 ± 0.055	0.340 ± 0.031
	Subject 2	0.290 ± 0.067	0.372 ± 0.020	97.3% ± 1.9%	99.0% ± 0.7%	94.7% ± 3.1%	91.7% ± 3.9%	0.633 ± 0.055	0.351 ± 0.031
	Subject 5	0.268 ± 0.073	0.375 ± 0.020	95.6% ± 2.9%	98.5% ± 1.1%	96.0% ± 2.7%	94.3% ± 2.8%	0.609 ± 0.058	0.341 ± 0.031
	Subject 7	0.253 ± 0.075	0.367 ± 0.021	95.0% ± 3.3%	98.0% ± 1.4%	94.1% ± 3.8%	92.4% ± 3.7%	0.644 ± 0.058	0.362 ± 0.033
SANE + MindEye2	Subject 1	0.253 ± 0.065	0.389 ± 0.017	92.3% ± 4.0%	96.2% ± 2.2%	90.9% ± 5.0%	87.7% ± 5.1%	0.711 ± 0.054	0.404 ± 0.033
	Subject 2	0.227 ± 0.065	0.380 ± 0.018	91.2% ± 4.4%	95.7% ± 2.4%	88.9% ± 5.9%	87.0% ± 5.4%	0.722 ± 0.054	0.406 ± 0.032
	Subject 5	0.216 ± 0.069	0.394 ± 0.017	88.9% ± 5.4%	95.4% ± 2.4%	91.3% ± 5.1%	89.7% ± 4.4%	0.694 ± 0.055	0.394 ± 0.032
	Subject 7	0.197 ± 0.075	0.384 ± 0.018	87.1% ± 5.9%	93.0% ± 3.3%	87.9% ± 5.8%	86.5% ± 5.5%	0.745 ± 0.052	0.429 ± 0.032
NeuroPictor†	Subject 1	0.270 ± 0.090	0.407 ± 0.020	99.0% ± 0.7%	99.3% ± 0.5%	95.7% ± 2.8%	94.4% ± 2.9%	0.616 ± 0.055	0.335 ± 0.031
	Subject 2	0.249 ± 0.089	0.406 ± 0.019	98.6% ± 0.9%	99.0% ± 0.7%	94.4% ± 3.4%	92.9% ± 3.2%	0.636 ± 0.053	0.347 ± 0.029
	Subject 5	0.201 ± 0.093	0.389 ± 0.021	94.6% ± 3.4%	98.1% ± 1.3%	95.1% ± 2.9%	94.8% ± 2.7%	0.623 ± 0.054	0.342 ± 0.032
	Subject 7	0.195 ± 0.093	0.389 ± 0.021	94.5% ± 3.1%	97.1% ± 1.6%	93.0% ± 4.1%	91.7% ± 3.5%	0.673 ± 0.056	0.375 ± 0.032
SANE + NeuroPictor	Subject 1	0.174 ± 0.080	0.393 ± 0.021	85.2% ± 6.3%	92.6% ± 3.7%	88.2% ± 7.0%	88.3% ± 4.7%	0.751 ± 0.060	0.436 ± 0.037
	Subject 2	0.175 ± 0.097	0.443 ± 0.020	84.6% ± 7.2%	91.8% ± 4.2%	87.2% ± 7.8%	85.6% ± 6.1%	0.772 ± 0.061	0.457 ± 0.037
	Subject 5	0.143 ± 0.098	0.406 ± 0.022	84.1% ± 7.8%	92.1% ± 3.9%	91.3% ± 5.1%	90.9% ± 4.0%	0.702 ± 0.059	0.397 ± 0.037
	Subject 7	0.137 ± 0.096	0.412 ± 0.024	80.5% ± 8.7%	88.7% ± 5.1%	85.7% ± 8.0%	85.1% ± 5.6%	0.781 ± 0.060	0.455 ± 0.039
NeuroAdapter†	Subject 1	0.088 ± 0.124	0.273 ± 0.034	80.3% ± 13.8%	90.6% ± 7.7%	89.4% ± 8.3%	89.5% ± 7.7%	0.738 ± 0.080	0.461 ± 0.058
	Subject 2	0.077 ± 0.122	0.280 ± 0.035	77.0% ± 16.2%	88.2% ± 9.6%	86.7% ± 10.1%	86.3% ± 9.9%	0.769 ± 0.078	0.494 ± 0.057
	Subject 5	0.079 ± 0.123	0.269 ± 0.037	78.2% ± 15.4%	87.6% ± 10.0%	87.3% ± 10.1%	86.4% ± 10.4%	0.775 ± 0.080	0.519 ± 0.061
	Subject 7	0.076 ± 0.126	0.280 ± 0.036	75.4% ± 16.3%	86.1% ± 10.8%	84.2% ± 11.6%	84.6% ± 11.0%	0.787 ± 0.079	0.510 ± 0.062
SANE + NeuroAdapter	Subject 1	0.099 ± 0.124	0.300 ± 0.029	81.0% ± 13.2%	90.9% ± 7.0%	88.3% ± 8.1%	89.3% ± 7.2%	0.725 ± 0.073	0.429 ± 0.049
	Subject 2	0.100 ± 0.131	0.310 ± 0.028	78.5% ± 14.3%	89.0% ± 8.4%	86.9% ± 9.4%	88.0% ± 8.2%	0.746 ± 0.078	0.443 ± 0.051
	Subject 5	0.093 ± 0.127	0.299 ± 0.028	79.7% ± 14.1%	89.8% ± 8.0%	90.5% ± 6.6%	91.2% ± 6.4%	0.712 ± 0.075	0.426 ± 0.051
	Subject 7	0.087 ± 0.127	0.292 ± 0.031	77.2% ± 15.3%	87.3% ± 8.8%	84.1% ± 10.5%	87.3% ± 8.9%	0.765 ± 0.075	0.457 ± 0.051
Brain-IT†	Subject 1	0.434 ± 0.048	0.483 ± 0.012	99.7% ± 0.2%	99.8% ± 0.2%	98.0% ± 1.2%	97.0% ± 1.8%	0.544 ± 0.053	0.305 ± 0.029
	Subject 2	0.393 ± 0.057	0.479 ± 0.013	99.6% ± 0.3%	99.8% ± 0.2%	97.8% ± 1.4%	96.3% ± 2.1%	0.550 ± 0.052	0.312 ± 0.029
	Subject 5	0.335 ± 0.059	0.472 ± 0.013	97.6% ± 1.6%	99.1% ± 0.7%	97.2% ± 1.9%	96.8% ± 1.8%	0.569 ± 0.057	0.329 ± 0.031
	Subject 7	0.337 ± 0.059	0.468 ± 0.013	97.7% ± 1.5%	99.0% ± 0.7%	96.2% ± 2.4%	95.2% ± 2.7%	0.596 ± 0.054	0.348 ± 0.033
SANE + Brain-IT	Subject 1	0.366 ± 0.050	0.471 ± 0.012	98.7% ± 0.9%	99.2% ± 0.5%	95.8% ± 2.5%	94.8% ± 2.7%	0.611 ± 0.054	0.348 ± 0.031
	Subject 2	0.320 ± 0.058	0.466 ± 0.013	98.3% ± 1.2%	99.0% ± 0.7%	95.0% ± 3.0%	93.2% ± 3.1%	0.629 ± 0.053	0.361 ± 0.031
	Subject 5	0.272 ± 0.064	0.466 ± 0.013	95.0% ± 3.1%	98.2% ± 1.2%	94.7% ± 3.1%	94.6% ± 2.6%	0.632 ± 0.058	0.369 ± 0.034
	Subject 7	0.270 ± 0.063	0.459 ± 0.014	94.3% ± 3.3%	97.0% ± 1.7%	91.7% ± 4.1%	91.6% ± 3.3%	0.670 ± 0.055	0.397 ± 0.034
NeuroFlow†	Subject 1	0.275 ± 0.016	0.202 ± 0.006	94.0% ± 1.8%	98.6% ± 0.6%	98.5% ± 0.8%	99.1% ± 0.5%	0.584 ± 0.027	0.346 ± 0.012
	Subject 2	0.276 ± 0.017	0.209 ± 0.006	94.6% ± 1.8%	99.1% ± 0.4%	98.8% ± 0.7%	99.3% ± 0.5%	0.579 ± 0.026	0.344 ± 0.013
	Subject 5	0.274 ± 0.017	0.215 ± 0.006	92.2% ± 2.1%	98.4% ± 0.6%	98.7% ± 0.7%	99.2% ± 0.6%	0.584 ± 0.028	0.346 ± 0.013
	Subject 7	0.259 ± 0.017	0.214 ± 0.007	92.5% ± 2.2%	98.4% ± 0.6%	98.6% ± 0.8%	99.1% ± 0.5%	0.584 ± 0.026	0.351 ± 0.012
SANE + NeuroFlow	Subject 1	0.182 ± 0.019	0.211 ± 0.006	88.1% ± 3.4%	94.5% ± 1.8%	91.1% ± 3.6%	90.9% ± 3.6%	0.742 ± 0.026	0.428 ± 0.012
	Subject 2	0.176 ± 0.019	0.212 ± 0.006	87.6% ± 3.4%	94.5% ± 1.7%	89.2% ± 3.9%	89.1% ± 3.8%	0.742 ± 0.025	0.426 ± 0.012
	Subject 5	0.165 ± 0.018	0.203 ± 0.006	83.7% ± 3.7%	92.4% ± 2.1%	91.8% ± 3.3%	92.1% ± 2.7%	0.719 ± 0.025	0.420 ± 0.012
	Subject 7	0.157 ± 0.019	0.192 ± 0.006	81.3% ± 4.0%	90.4% ± 2.3%	87.7% ± 4.1%	87.9% ± 4.2%	0.764 ± 0.024	0.454 ± 0.012

The five adapters are not intended to be parameter-matched to one another: their output ranks and semantics are fixed by five different pretrained systems. We therefore report only controlled native-versus-SANE replacement within each backbone, rather than treating cross-backbone differences as an adapter-capacity comparison.

E.2.7 Published and Subject-Level Reconstruction Results

Table 11 summarizes published results from representative prior reconstruction methods and provides a subject-level breakdown of the controlled native and SANE-driven evaluations across the five reconstruction backbones.

Consistency across subjects. The subject-wise breakdown shows that the aggregate conclusions are not driven by a single participant. The clearest example is NeuroAdapter: replacing its native interface with SANE increases PixCorr, SSIM, AlexNet-2, and AlexNet-5 for each of the four subjects, and decreases both EfficientNet and SwAV distance for every subject. The high-level measures follow the same overall tendency but are not uniformly positive: CLIP improves for Subjects 2, 5, and 7, whereas Inception changes only modestly for Subjects 1, 2, and 7 and improves more clearly for Subject 5. Thus, the NeuroAdapter result reflects a repeatable subject-level pattern rather than one unusually favorable subject or metric.

Backbone-specific trade-offs. The other four interfaces retain useful reconstruction signals but exhibit different metric trade-offs. For MindEye2, SANE raises SSIM for all four subjects, even though PixCorr and perceptual identification decrease, indicating better preservation of coarse structural organization than of exact appearance or feature identity. For NeuroPictor, SSIM increases for Subjects 2, 5, and 7 but PixCorr and identification are lower for all four subjects. SANE-driven Brain-IT also remains highly identifiable at the AlexNet-5 level (97.0–99.2%) and retains Inception and CLIP identification above 91% for every subject, despite a consistent reduction relative to its own native interface. NeuroFlow shows a related retention pattern: its SANE-driven reconstructions retain 87.7–91.8% Inception and 87.9–92.1% CLIP identification, while PixCorr and feature-space distances worsen. These observations support the narrower portability claim that the same compact SANE representation can drive all five frozen reconstruction pipelines; they do not imply that interface replacement is lossless for every backbone.

Across the table, the direction of the native-to-SANE change is generally repeated across Subjects 1, 2, 5, and 7, whereas the magnitude depends on the backbone and metric. This distinction is important: the experiment evaluates whether compact evidence can substitute for each native neural front-end under a controlled downstream pipeline, not whether a single adapter architecture should yield identical behavior across heterogeneous conditioning spaces. The backbone-dependent mechanisms behind these patterns are discussed in Appendix E.3.

E.2.8 Additional Qualitative Reconstructions

Figures 17–20 visualize the same controlled interface-replacement experiment reported quantitatively above. The grids are assembled from final reconstruction outputs by matching the subject and global NSD stimulus identity and, within each backbone, pairing the native and SANE-driven images generated with the same prespecified seed. The native image is produced by the backbone's original neural encoder, whereas its paired SANE image replaces only that neural interface with frozen $K = 32$ evidence and the corresponding adapter. All modules after the replacement boundary, including the sampler, decoder, enhancement stages, and post-processing used by that backbone, remain identical. Ground-truth images, where displayed, serve only as visual references and are never supplied to the SANE-driven inference path.

The figures show final post-processed reconstructions rather than intermediate adapter targets. They should therefore be read pairwise within each backbone: object and scene identity indicate semantic retention, spatial arrangement and dominant color indicate structural retention, and texture or boundary sharpness expose losses in fine detail. Absolute appearance should not be compared across backbones because their generators and native conditioning spaces differ. The four sets broaden the qualitative coverage across held-out stimuli; they are presented as complements to the complete 1,000-image metrics rather than as a substitute for quantitative evaluation, and no best-of-N selection is used.

E.3 Backbone-Dependent Effects of SANE Interface Replacement

The main portability experiments evaluate SANE with five heterogeneous pretrained reconstruction backbones: MindEye2, NeuroPictor, NeuroAdapter, NeuroFlow, and Brain-IT. Across these systems, the effect of replacing the native neural front-end with compact SANE evidence is not uniform. The SANE-driven NeuroAdapter variant improves several reconstruction metrics relative to its native counterpart, whereas MindEye2, Brain-IT, NeuroFlow, and NeuroPictor generally retain useful reconstruction ability but exhibit some loss in fidelity. This section analyzes these backbone-dependent differences and then reports an additional experiment using the comparatively simple Takagi-Nishimoto fMRI-to-Stable-Diffusion pipeline.

The purpose of this analysis is not to argue that SANE should outperform every task-specific neural encoder. Rather, it clarifies when a compact, independently learned neural evidence representation can effectively replace a backbone's native neural interface and when the information bottleneck introduced by SANE becomes limiting.

The downstream image generator is preserved. A central aspect of the portability experiment is that SANE replaces the neural interface rather than the generative model. For each backbone, SANE is frozen and a lightweight backbone-specific adapter maps the contextualized $K = 32$

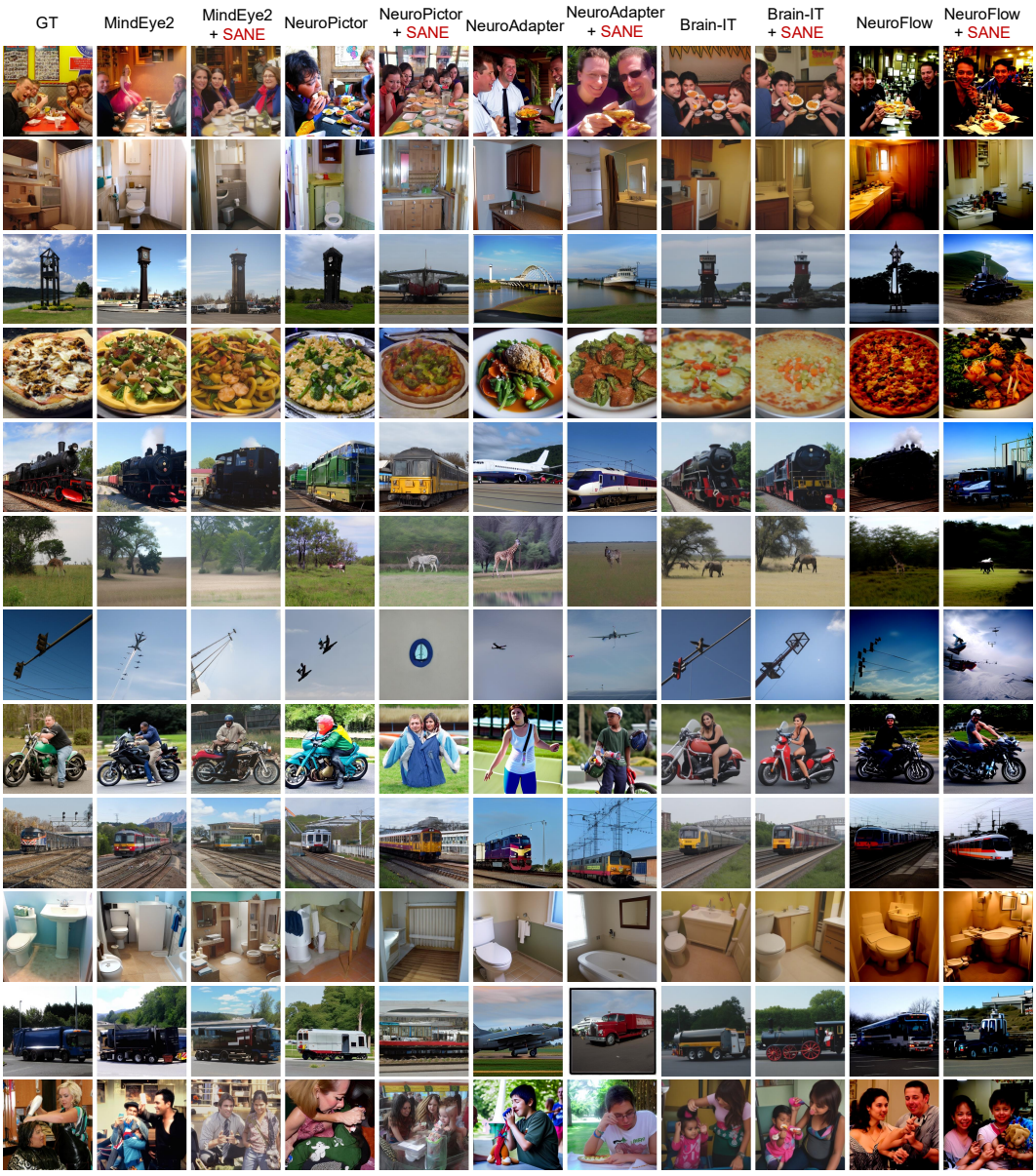


Figure 17: Qualitative backbone comparison, example set 1. Native and SANE-driven reconstructions across MindEye2, NeuroPictor, NeuroAdapter, Brain-IT, and NeuroFlow.

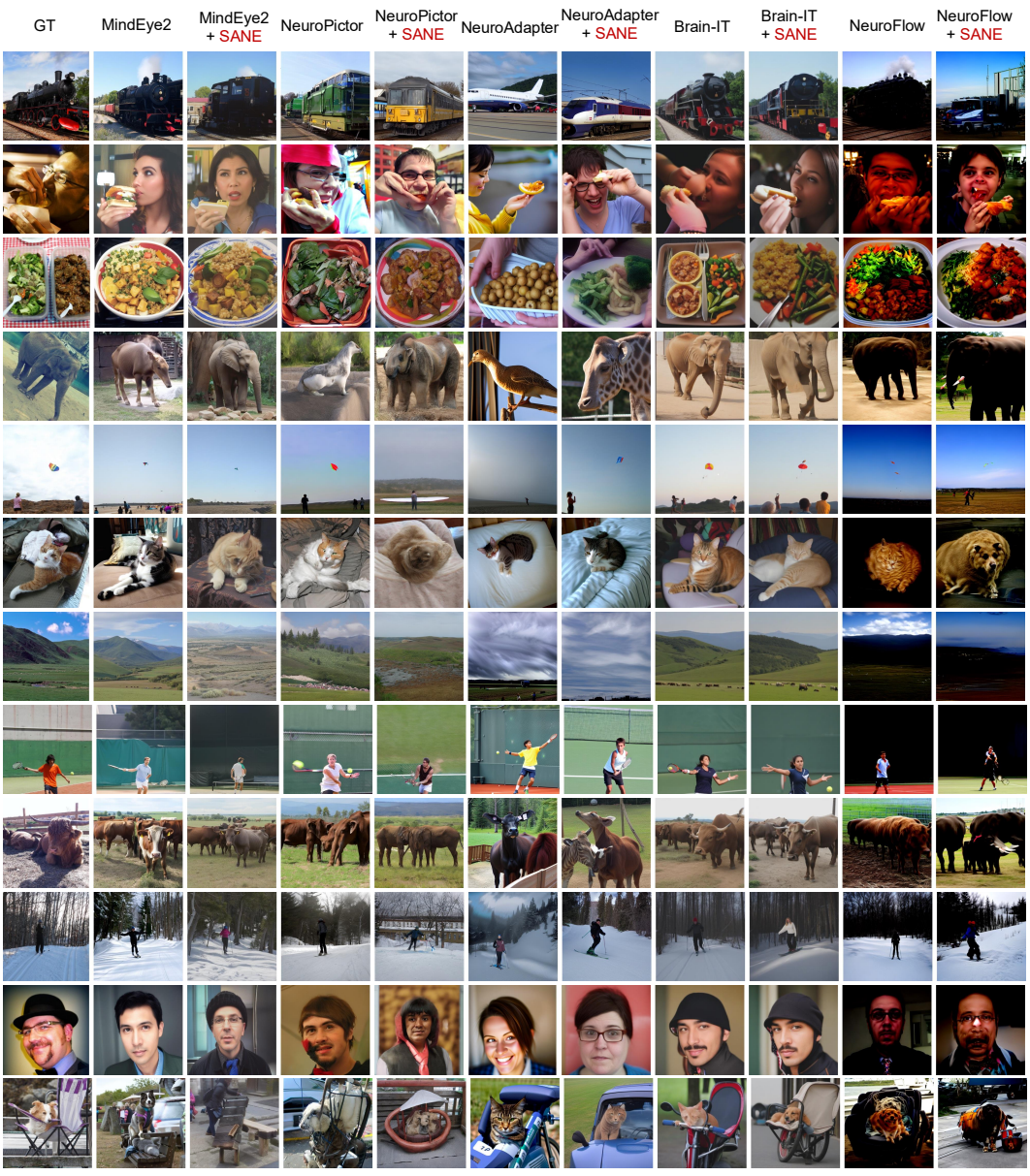


Figure 18: Qualitative backbone comparison, example set 2. Native and SANE-driven reconstructions across MindEye2, NeuroPictor, NeuroAdapter, Brain-IT, and NeuroFlow.

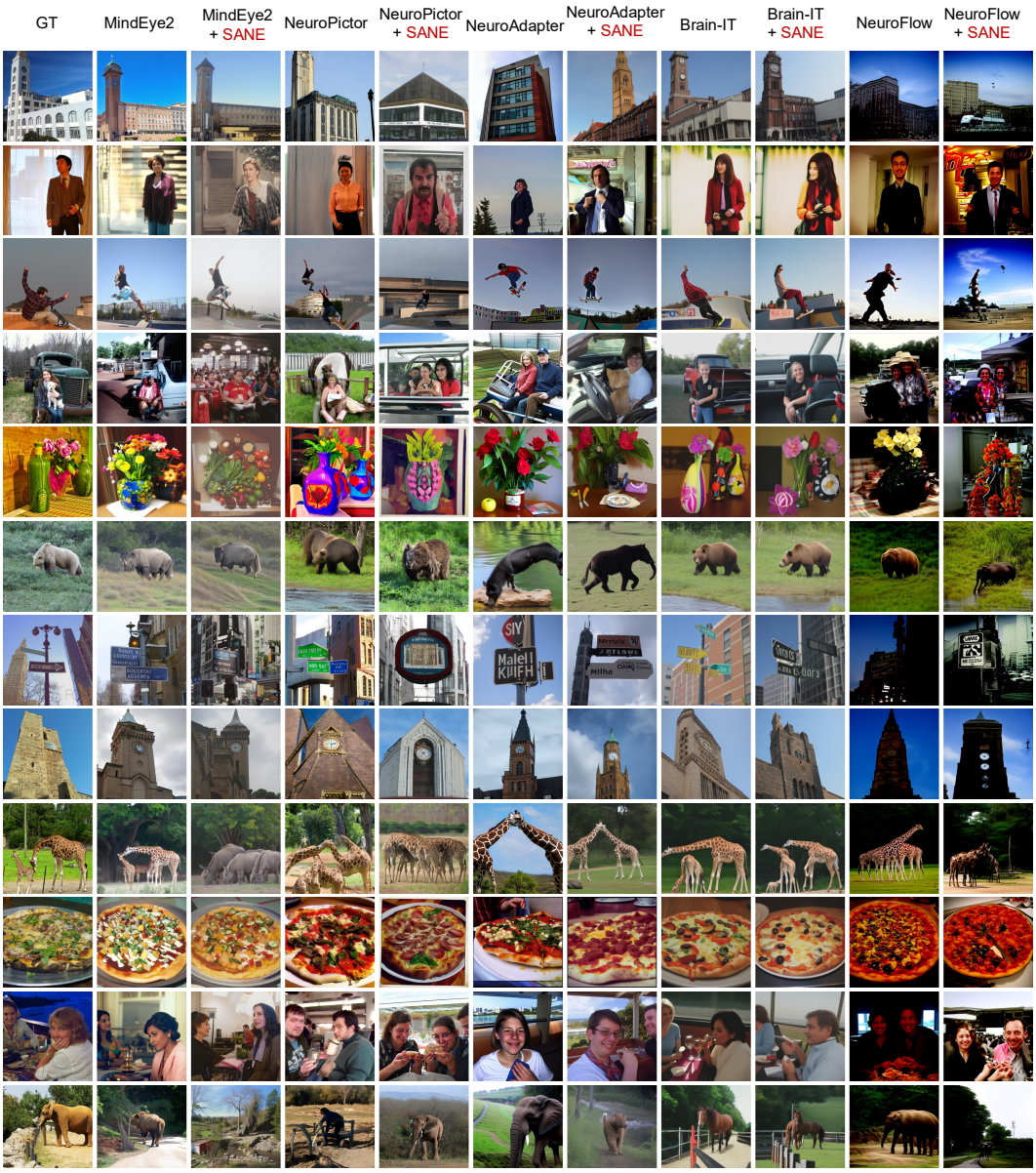


Figure 19: Qualitative backbone comparison, example set 3. Native and SANE-driven reconstructions across MindEye2, NeuroPictor, NeuroAdapter, Brain-IT, and NeuroFlow.

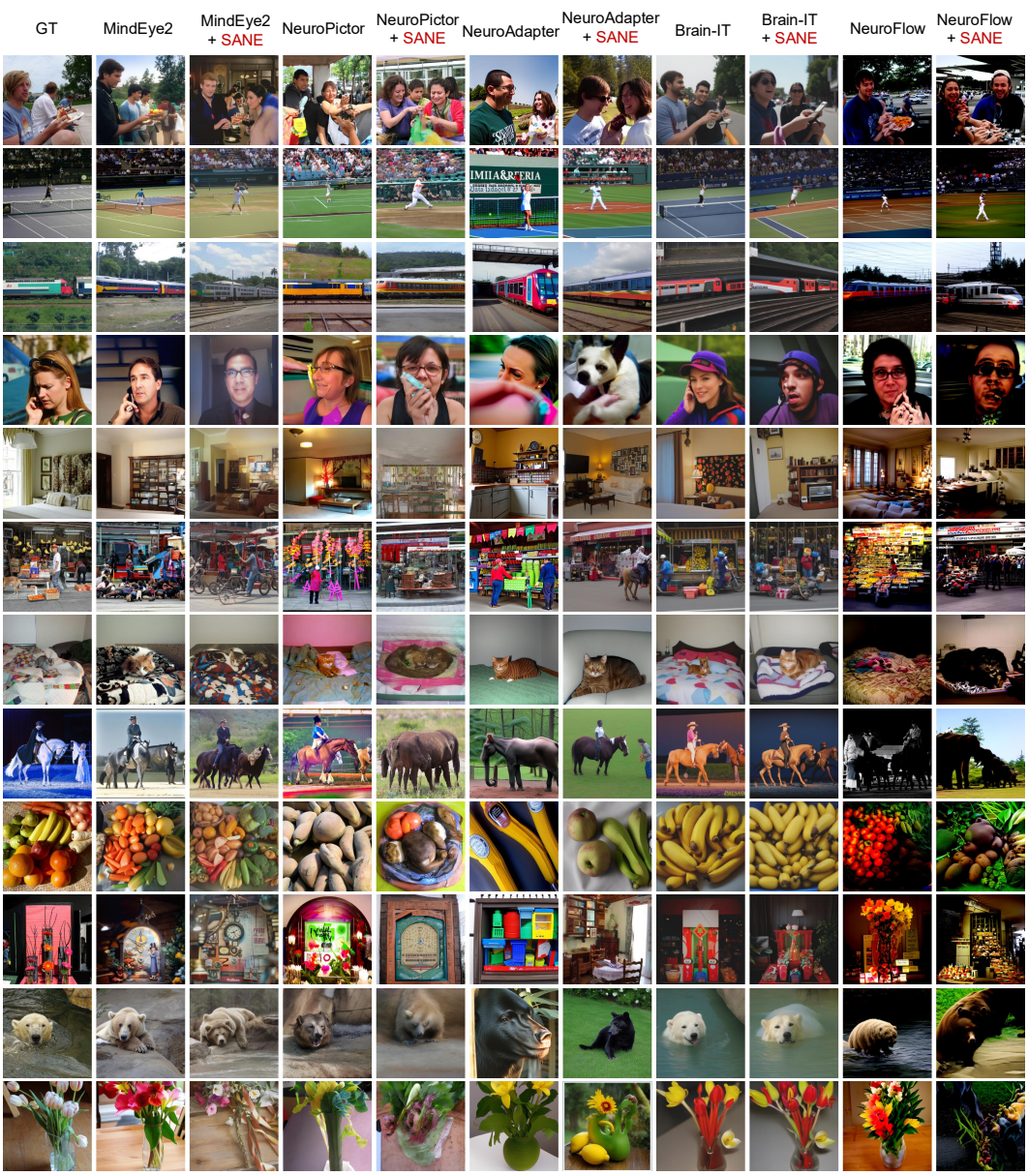


Figure 20: Qualitative backbone comparison, example set 4. Native and SANE-driven reconstructions across MindEye2, NeuroPictor, NeuroAdapter, Brain-IT, and NeuroFlow.

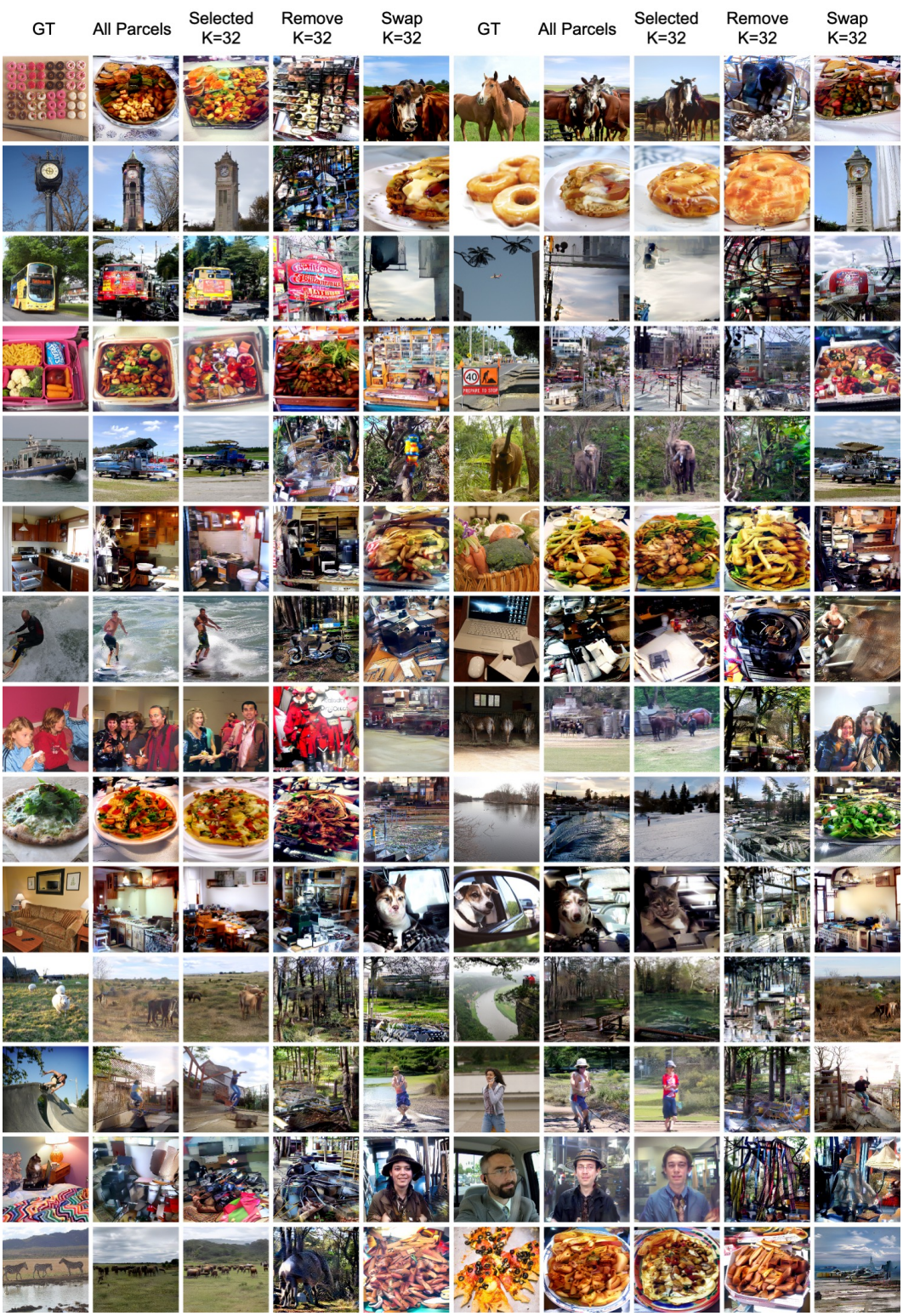


Figure 21: Qualitative retention, removal, and cross-stimulus substitution with MindEye2. Each row shows a left-right stimulus pair, with the GT, reconstruction from all parcels, reconstruction using only the SANE-selected $K = 32$ parcels, reconstruction after their removal, and reconstruction after swapping their responses with those from the paired stimulus.

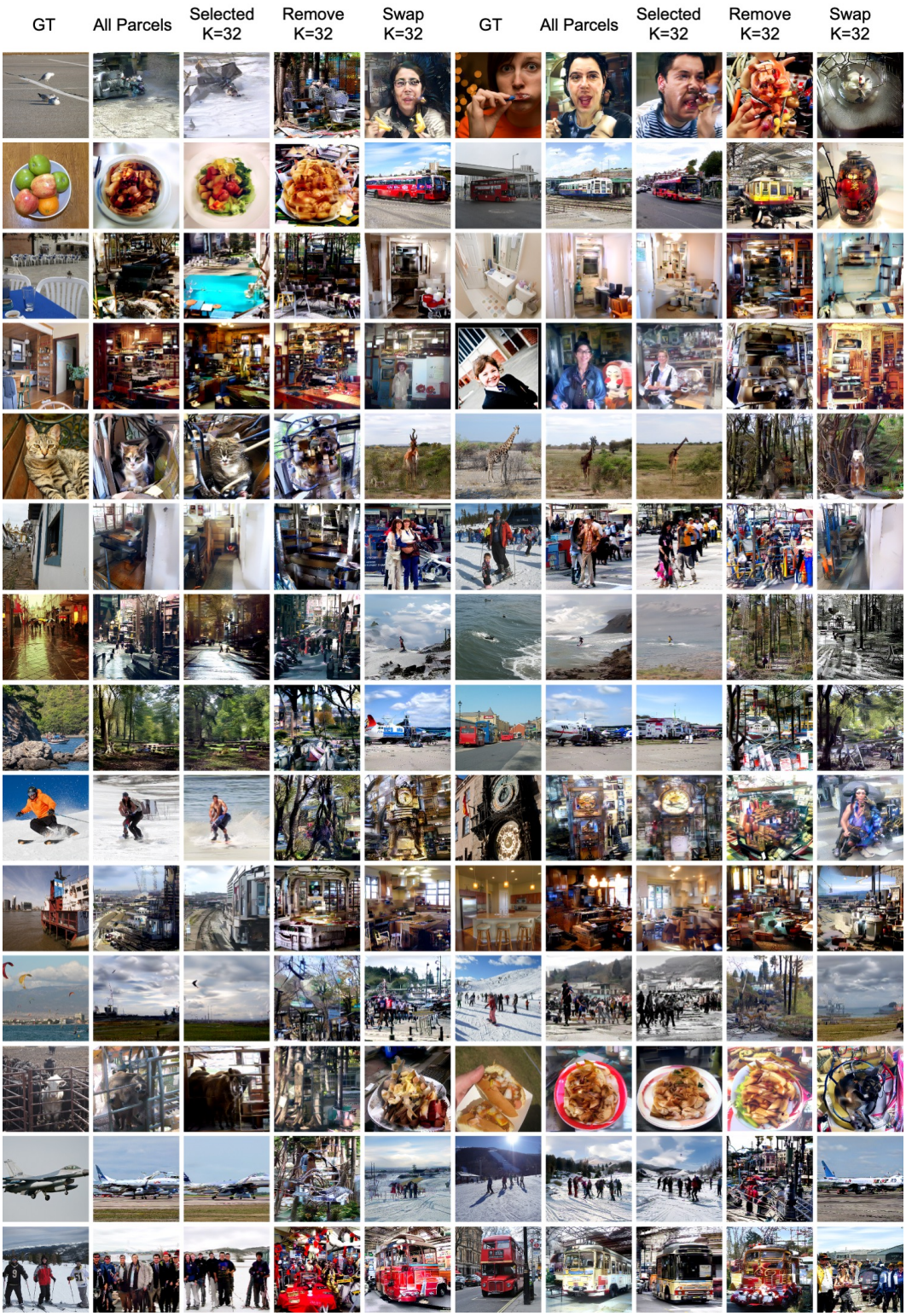


Figure 22: Additional qualitative examples of retention, removal, and cross-stimulus substitution using MindEye2. The layout and intervention protocol follow Fig. 21.



Figure 23: Additional qualitative examples of retention, removal, and cross-stimulus substitution using MindEye2. The layout and intervention protocol follow Fig. 21.

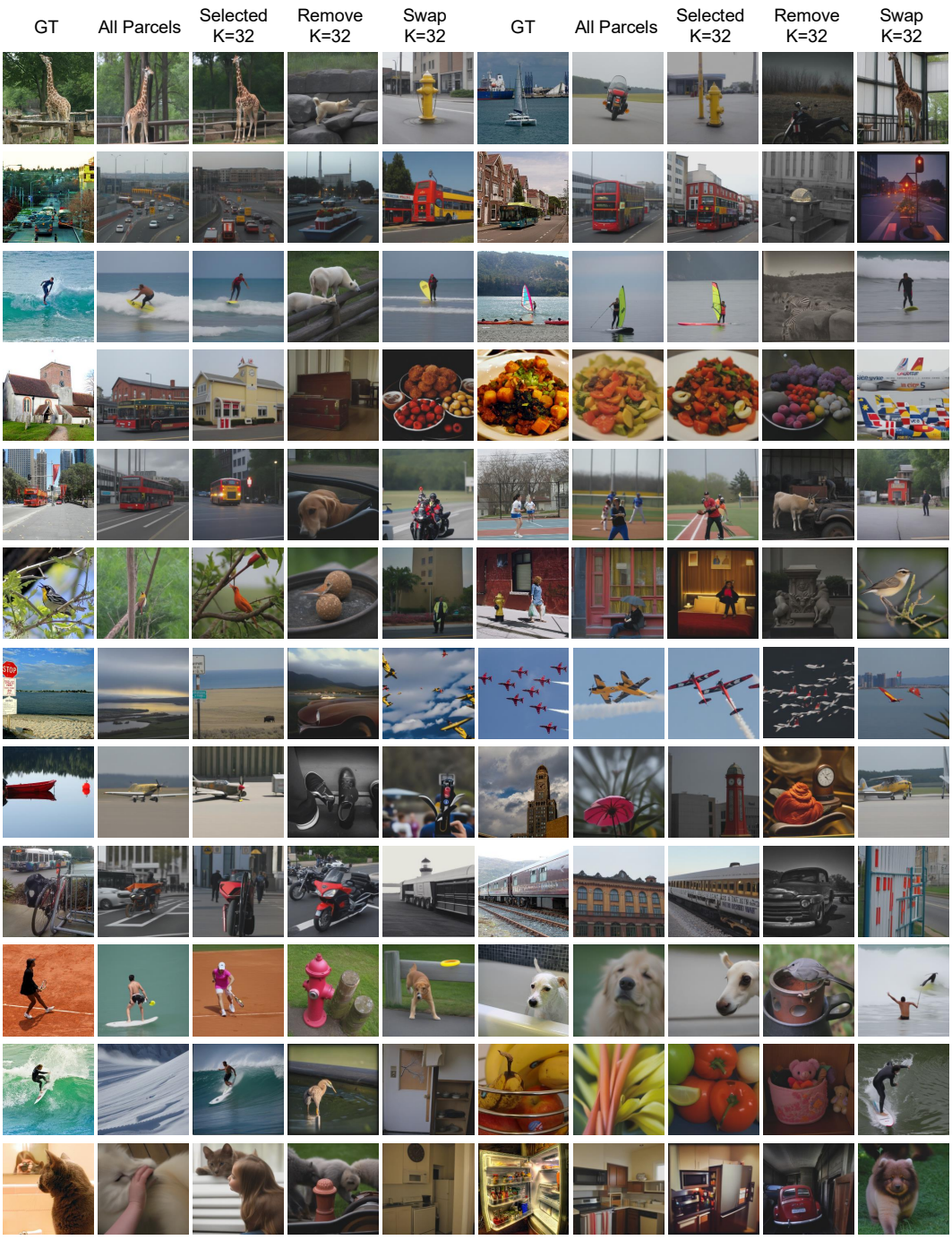


Figure 24: Qualitative retention, removal, and cross-stimulus substitution using Brain-IT. Each row shows a left-right stimulus pair, with GT, all-parcel, SANE $K = 32$, removal, and cross-stimulus swap reconstructions.

evidence representation to the native conditioning space required by the reconstruction system. The downstream generative and post-processing components remain unchanged.

Consequently, differences between the native and SANE-driven variants primarily reflect the information supplied by the neural front-end and the difficulty of aligning compact SANE evidence to the corresponding native representation. They should not be interpreted as arising from the use of a stronger image generator.

Why NeuroAdapter is particularly compatible with SANE. Among the five main reconstruction backbones, NeuroAdapter provides the clearest example in which SANE replacement improves several aggregate reconstruction metrics. Its neural interface is comparatively direct: the native brain front-end produces a sequence of conditioning tokens that is subsequently consumed by the preserved image-projection, attention, and Stable Diffusion modules. In the SANE-driven variant, a lightweight adapter maps the frozen SANE evidence directly into this conditioning-token space.

This interface is favorable to SANE for two reasons. First, SANE already provides a nonlinear and visually supervised neural representation through parcel-specific encoding, stimulus-dependent evidence selection, and post-selection cross-parcel contextualization. Second, the downstream alignment problem remains relatively constrained because the adapter only needs to reproduce the conditioning representation expected by the preserved NeuroAdapter generation pipeline rather than reconstruct several heterogeneous semantic and low-level intermediate spaces.

The SANE-NeuroAdapter interface is additionally trained using explicit token-level alignment to the native conditioning representation. This provides a direct supervision signal at the interface boundary and may act as a regularized projection of the native neural representation. Together with the visual structure already learned by SANE, this can yield a stable conditioning signal for the frozen generator.

We do not interpret this result as evidence that a compact SANE representation is intrinsically superior to the native NeuroAdapter representation. Rather, it demonstrates that, for a relatively direct conditioning interface, the information retained by the $K = 32$ SANE representation is sufficient not only to drive the reconstruction system but also to improve several aggregate reconstruction metrics.

Why richer native neural interfaces are harder to replace. MindEye2, Brain-IT, NeuroPictor, and NeuroFlow present a different regime. Their native neural encoders and intermediate representations are more specialized and preserve richer semantic, spatial, and low-level information before image generation.

MindEye2 and Brain-IT contain distinct semantic and low-level pathways. NeuroPictor relies on a high-dimensional latent representation that contributes to both semantic and low-level generation. NeuroFlow uses a strong pretrained neural representation coupled to its downstream generative pathway. In these systems, the neural encoder is itself a substantial component of the reconstruction architecture rather than a relatively simple mapping from fMRI to a single conditioning sequence.

SANE, by design, imposes a compact evidence bottleneck. At the default operating point, the contextualized neural representation contains only 32 selected tokens of dimension 512. Several downstream interfaces are substantially richer; for example, the MindEye2 and NeuroFlow integrations require representations of approximately 256×1664 dimensions, while the NeuroPictor interface operates on an approximately 257×1024 latent representation.

The dimensional difference itself should not be interpreted as the cause of the performance gap. Instead, it illustrates the amount of structured information that the lightweight adapter must recover from a deliberately compact neural representation. When the native front-end already explicitly represents detailed spatial or low-level information, information discarded by the $K = 32$ bottleneck cannot necessarily be reconstructed by increasing the dimensionality of the adapter output.

This yields a natural trade-off. SANE contributes a compact, visually supervised representation in which stimulus-relevant evidence has already been identified and integrated, but replacing a rich native neural encoder also removes information that the original reconstruction architecture may exploit. The relative importance of these two effects varies across reconstruction systems.

Table 12: Supplementary reconstruction results on the Takagi-Nishimoto pipeline. The native pipeline uses linear ridge regression to map fMRI to Stable Diffusion conditioning representations, whereas the SANE-driven variant predicts the same downstream representations from frozen $K = 32$ SANE evidence. The Stable Diffusion reconstruction stage is otherwise preserved. Results are reported for NSD Subjects 1, 2, 5, and 7 and the four-subject macro average. † denotes our reproduced results. Higher is better for PixCorr, SSIM, Alex(2), Alex(5), Inception, and CLIP, while lower is better for EfficientNet and SwAV.

		Low-Level				High-Level			
Method	**Subject**	**PixCorr ↑**	**SSIM ↑**	**Alex(2) ↑**	**Alex(5) ↑**	**Incep ↑**	**CLIP ↑**	**Eff ↓**	**SwAV ↓**
Takagi & Nishimoto†	Subject 1	0.280	0.229	75.9%	77.6%	65.3%	63.3%	0.942	0.669
	Subject 2	0.238	0.229	73.0%	75.8%	62.8%	61.9%	0.948	0.674
	Subject 5	0.197	0.222	70.1%	74.3%	64.4%	63.6%	0.944	0.673
	Subject 7	0.183	0.216	68.5%	72.2%	61.9%	62.0%	0.950	0.680
	4-subject macro	**0.225**	**0.224**	**71.9%**	**75.0%**	**63.6%**	**62.7%**	**0.946**	**0.674**
SANE + Takagi & Nishimoto	Subject 1	0.224	0.328	75.0%	83.3%	84.8%	82.9%	0.796	0.489
	Subject 2	0.207	0.325	74.4%	81.9%	83.2%	81.0%	0.811	0.497
	Subject 5	0.222	0.330	75.0%	83.0%	86.0%	84.1%	0.784	0.484
	Subject 7	0.197	0.323	72.3%	80.5%	82.4%	80.9%	0.817	0.505
	4-subject macro	**0.212**	**0.326**	**74.2%**	**82.2%**	**84.1%**	**82.2%**	**0.802**	**0.494**

E.3.1 Supplementary Evaluation on a Basic fMRI-to-Diffusion Pipeline

To complement the five modern reconstruction backbones, we additionally evaluate SANE in the comparatively simple Takagi-Nishimoto reconstruction pipeline. This experiment is not included as a sixth backbone in the main portability comparison. Instead, it serves as a controlled supplementary case in which the native neural-to-image interface is substantially simpler.

The original pipeline follows a direct fMRI-to-Stable-Diffusion design. Standardized fMRI responses are mapped by linear ridge regression to two visual representations: semantic conditioning for Stable Diffusion and an initial VAE latent used for low-level reconstruction. The subsequent Stable Diffusion reconstruction procedure is then applied without modifying the image generator.

In the SANE-driven variant, the same downstream Stable Diffusion reconstruction stage is retained. The linear brain-to-visual mappings are replaced by lightweight nonlinear readouts from the frozen contextualized SANE evidence, which predict the corresponding semantic conditioning and low-level latent representations. Thus, the comparison isolates the effect of replacing a relatively simple linear neural mapping with the compact SANE representation while preserving the downstream generator.

SANE is effective in the simple fMRI-to-diffusion setting. The supplementary experiment shows that SANE can be integrated successfully even into a basic fMRI-to-Stable-Diffusion reconstruction pipeline. At the four-subject macro level, replacing the native ridge-based neural mapping with SANE increases SSIM from 0.224 to 0.326, Alex(5) from 75.0% to 82.2%, Inception similarity from 63.6% to 84.1%, and CLIP similarity from 62.7% to 82.2%. EfficientNet and SwAV distances decrease from 0.946 to 0.802 and from 0.674 to 0.494, respectively. PixCorr decreases slightly from 0.225 to 0.212.

The metric pattern is informative. The strongest gains occur in semantic and perceptual measures, whereas exact pixel-level correspondence does not improve uniformly. This is consistent with the role of SANE as a compact visually supervised representation: it preserves stimulus-relevant structure and semantic content particularly well, while the $K = 32$ bottleneck may discard some fine-grained information required for precise pixel alignment.

A backbone-dependent view of interface replacement. Taken together, the five main portability experiments and the supplementary Takagi-Nishimoto experiment reveal a consistent pattern. The effectiveness of SANE replacement depends on the native neural interface that it replaces.

When the original neural mapping is comparatively simple and the alignment target is direct, as in the Takagi-Nishimoto pipeline and, to a lesser extent, NeuroAdapter, the pretrained structure provided by SANE can compensate for or exceed the information supplied by the native neural front-end. In these settings, the compact evidence representation can preserve or improve a substantial fraction of reconstruction quality.

In contrast, when the native neural encoder is already highly specialized and produces rich semantic, spatial, or low-level representations, as in MindEye2, Brain-IT, NeuroPictor, and NeuroFlow, replacing it with a $K = 32$ compact evidence representation creates a more demanding information bottleneck. The lightweight adapter must reconstruct information that the original architecture encoded explicitly, and some of this information may no longer be available after compact evidence selection.

The results therefore support a backbone-dependent interpretation of SANE interface replacement: *the more direct the native interface and the simpler the original neural mapping, the more effectively compact SANE evidence can replace it; as the native encoder becomes richer and increasingly dependent on detailed spatial or low-level representations, the information bottleneck imposed by* $K = 32$ *becomes more consequential.*

This interpretation also clarifies the purpose of the main portability experiment. SANE is not intended to be a universally superior replacement for every task-specific neural encoder. Rather, the central result is that a single compact evidence representation, learned independently of any image generator, remains sufficiently informative to condition heterogeneous frozen reconstruction systems after lightweight alignment. The fact that it can additionally improve reconstruction in simpler neural-interface settings provides complementary evidence that the learned representation is not merely a compressed substitute, but can also provide a useful neural-to-visual representation when the native mapping is comparatively limited.

Scope of the interpretation. The backbone-dependent explanation above is supported by the observed reconstruction patterns and the architectural differences among the evaluated systems, but it should not be interpreted as a controlled causal decomposition of encoder strength, interface dimensionality, or adapter complexity. Establishing the independent contribution of each factor would require matched experiments controlling neural-encoder capacity, alignment objective, intermediate representation dimensionality, numerical precision, and downstream generation settings. We therefore present this analysis as a mechanistic interpretation of when compact SANE evidence transfers effectively rather than as evidence that any single architectural property uniquely determines reconstruction performance.

E.4 Neural-Side Efficiency Across Reconstruction Backbones

Beyond reconstruction quality, we examine the computational footprint of replacing each backbone's native neural front-end with compact SANE evidence. The analysis covers the five reconstruction systems used in the main portability experiments: MindEye2, Brain-IT, NeuroAdapter, NeuroFlow, and NeuroPictor.

Evaluation scope. To isolate the cost of neural representation construction, we measure only the pathway from fMRI input to the conditioning representation consumed by the downstream generative model. Diffusion sampling, unCLIP, VAE image decoding, UNet or ControlNet denoising, and final image generation are excluded. The comparison therefore characterizes the computational cost of the neural interface rather than the end-to-end image-generation pipeline.

For each backbone, we compare its native neural front-end with the complete SANE-driven interface. The latter consists of the SANE evidence pathway together with the backbone-specific adapter required to produce the corresponding native conditioning representation.

Parameter-count convention. The reported parameter counts measure the learned capacity required to obtain the complete neural-side representation. Accordingly, the SANE total includes both the parameters trained to obtain the SANE representation and those of the corresponding backbone-specific adapter. We do not count the SANE backbone as zero simply because it is frozen during downstream integration.

This quantity should be distinguished from the number of parameters updated when SANE is connected to a particular reconstruction backbone. After SANE pretraining, the SANE representation is frozen and only the corresponding lightweight adapter is optimized. Thus, the total parameter count measures the capacity required to obtain the complete SANE-driven neural pathway, whereas the adapter-stage trainable parameter count measures the additional optimization burden of integrating an already trained SANE representation with a new reconstruction system.

Table 13: Neural-side efficiency of native and SANE-driven reconstruction interfaces. Statistics cover only the pathway from fMRI input to the conditioning representation supplied to the downstream generator; image generation is excluded. "SANE" denotes the complete SANE neural pathway together with its backbone-specific adapter. Positive reduction values indicate lower cost for SANE, whereas negative values indicate additional cost. CLIP retention is the SANE-driven reconstruction CLIP score divided by the corresponding native score. Latency and peak-allocated-memory measurements use a batch size of 1 under an NVIDIA GeForce GTX TITAN X in FP32, with 20 warm-up iterations and 100 timed iterations. Brain-IT contains simulation-based efficiency estimates for components without complete direct measurements.

	Total Params. (M)			GFLOPs			B1 Latency (ms)			Peak Alloc. (MiB)			CLIP
Method	**Native**	**SANE**	**Red.**	**Native**	**SANE**	**Red.**	**Native**	**SANE**	**Red.**	**Native**	**SANE**	**Red.**	**Ret.**
MindEye2	1967.430	33.916	98.28%	8.619	5.473	36.50%	32.800	15.260	53.48%	7538.5	162.0	97.85%	94.38%
Brain-IT	155.824	23.396	84.99%	21.700	3.684	83.02%	141.345	82.980	41.29%	2325.5	1844.7	20.68%	97.10%
NeuroAdapter	88.451	19.527	77.92%	0.411	1.762	−328.20%	1.452	17.272	−1089.28%	348.7	93.3	73.25%	102.62%
NeuroFlow	56.082	24.445	56.41%	14.287	3.259	77.19%	6.872	16.406	−138.74%	143.8	110.7	22.99%	90.75%
NeuroPictor	311.949	49.544	84.12%	168.193	49.345	70.66%	66.298	40.402	39.06%	1212.4	260.2	78.54%	93.60%

Main findings. Across the five backbones, SANE reduces total learned parameter count and peak allocated memory in every case, and reduces GFLOPs for MindEye2, Brain-IT, NeuroFlow, and NeuroPictor. MindEye2, Brain-IT, and NeuroPictor retain 94.38%, 97.10%, and 93.60% of their native CLIP scores, respectively, while also reducing latency. NeuroAdapter is the main exception to the computational trend: its native interface is already very lightweight, so SANE increases GFLOPs and latency despite reducing parameters and memory and reaching 102.62% CLIP retention. NeuroFlow likewise reduces GFLOPs but not latency. These results show that the quality-efficiency trade-off depends on the native interface and that lower arithmetic cost does not necessarily translate into lower wall-clock latency.

Measurement sources and limitations. Latency is measured on an NVIDIA GeForce GTX TITAN X in FP32 with batch size 1, but measurement provenance is not identical across all components. MindEye2 uses complete two-branch same-GPU measurements; Brain-IT includes fixed simulation-based estimates and a same-GPU weight-based proxy; NeuroPictor includes a same-GPU partial-weight proxy; and NeuroAdapter and NeuroFlow use their existing same-GPU benchmark outputs. Absolute cross-backbone latency comparisons should therefore be interpreted cautiously, and the strongest comparison is the native-versus-SANE change within each backbone. The available NeuroFlow native-quality benchmark metadata flag the quality path used for its retained-CLIP calculation as containing target leakage; the 90.75% value is included for completeness but is not used as evidence for a controlled quality-efficiency trade-off.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and improving the clarity and readability of the manuscript, as well as for limited assistance with code development and debugging. Generative AI tools were not used to autonomously generate research hypotheses, experimental results, or scientific conclusions. All AI-assisted text was reviewed and revised by the authors, and all AI-assisted code was manually inspected and tested for correctness. The reported experimental results were obtained from our implemented experimental pipeline and independently verified by the authors. We take full responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly available neuroimaging datasets collected and released by prior studies under their respective ethical approvals and data-use agreements. We did not recruit new human participants or collect new human-subject data. Our experiments use neuroimaging data only for the research purposes permitted by the corresponding datasets. We acknowledge that visual reconstruction from brain activity may raise potential privacy and misuse concerns as the technology develops. The methods presented in this work are intended for scientific research, particularly for understanding and modeling neural representations, and should not be used to infer or reconstruct

an individual's private neural information without appropriate authorization and informed consent. We follow the applicable licenses, data-use requirements, and research ethics guidelines associated with the datasets used in this work.

REPRODUCIBILITY STATEMENT

We have taken several steps to facilitate reproducibility. The main paper provides the overall architecture, methodological formulation, training objectives, and experimental setup. Additional implementation details, hyperparameter settings, dataset preprocessing procedures, model configurations, and supplementary analyses are provided in the Appendix. We report results across multiple subjects and compare our method with established baselines using standard evaluation metrics and consistent evaluation protocols. Code and the necessary configurations for reproducing the experiments will be released to facilitate independent verification of our results.