BRAININTERV: IN SILICO INTERVENTIONS FOR MAPPING SEMANTIC SELECTIVITY IN THE BRAIN

Anonymous authors
Paper under double-blind review

ABSTRACT

Characterizing the selectivity of brain areas is important for understanding their underlying computation. Here we propose an in silico approach that allows us to address this challenge by performing interventions and studying their effect on a downstream decoding task. Brain-tuned visual representations are created by routing and transforming self-supervised visual features into region-of-interest (ROI) tokens, each supervised by functional magnetic resonance imaging (fMRI) responses from a cortical area. The resulting ROI-tuned image tokens support high-accuracy 1000-way semantic retrieval through a lightweight readout to a text-embedding space, reaching $\mathrm{R@10} = 0.942 \pm 0.007$. We further extend the decoding task by predicting the ROI tokens from measured fMRI responses using a voxel-to-token mapper. These decoded ROI tokens achieve $\mathrm{R@10} = 0.541 \pm 0.085$ under the same frozen image-trained semantic readout. We then perform in silico interventions on the model by transplanting ROI tokens across images for the retrieval task and observe that high-level token interventions produce substantially greater donor-semantic transfer than early-visual tokens, whose effects are stronger on lower-level visual probes. This intervention pattern also holds for both ROI-tuned image tokens and decoded tokens, and in different subjects. Overall, the proposed approach provides an intervention-based method to probe the selectivity of different visual brain areas, complementing current encoding- and direct decoding-based methods.

1 INTRODUCTION

Brain encoding models aim to predict neural responses to stimuli (Naselaris et al., 2011). These models have shown that recent deep neural networks provide computational accounts of visually evoked activity in the brain. The prediction accuracy tells us whether a model matches neural activity, but not necessarily what information is represented in a region or how that representation matters for the semantic processing in the brain. Decoding methods ask what information can be recovered from neural activity and usually rely on strong computational priors as well to produce naturalistic texts and images (Tang et al., 2023; Takagi & Nishimoto, 2023). However, a successful decoder does not by itself show which parts of the representation contribute to the decoded information (Vigotsky et al., 2024). This motivates our main question: can region-specific brain representations be made explicitly intervenable, so that their selectivity and contribution to downstream decoding can be tested directly?

We study this question by imposing an ROI-specific bottleneck on frozen self-supervised visual features. Starting from a DINOv3 backbone (Siméoni et al., 2025), we learn 50 ROI-specific tokens associated with visual-cortex regions from the probabilistic visual topography atlas (Wang et al., 2015). The ROI token representation is learned from visual features and tuned using fMRI responses from the corresponding predefined visual-cortex ROIs. We therefore refer to these representations as *ROI-tuned image tokens*.

We first test the semantic retrieval performance of this representation by mapping the frozen token set into an independently trained Qwen3-Embedding-8B caption space (Zhang et al., 2025). We then test whether different token groups have distinct functional roles using controlled token interventions, in which selected ROI tokens from a donor image replace the corresponding tokens of an original image while the rest of the representation and readout remain fixed. Finally, we ask whether

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

this same token space can be recovered from measured fMRI, and whether voxel-to-token inversion preserves the intervention ordering observed on ROI-tuned image tokens.

Our contributions are as follows. First, we introduce an ROI-tuned, brain-supervised visual representation that supports high-accuracy 1000-way semantic retrieval without language supervision in Stage 1. Second, high-level ROI-group interventions produce substantially stronger donor-semantic transfer than early-visual interventions, whereas early tokens preferentially carry lower-level visual structure; this intervention pattern is reproducible across subjects. Third, we show that the same ROI-indexed representation can be partially reconstructed directly from fMRI using a voxel-to-token mapper, under the same frozen image-trained Stage 2 readout, and that the recovered brain representation preserves the same broad intervention ordering, with reduced effect magnitudes.

2 Related Work

Brain encoding and decoding. Hierarchical neural networks have long served as models of the ventral visual stream, predicting responses from intermediate and high-level visual cortex (Yamins et al., 2014; Güçlü & van Gerven, 2015). Large naturalistic datasets such as the Natural Scenes Dataset (NSD) have made it possible to evaluate such models at scale using repeated 7 T fMRI responses to thousands of natural images (Allen et al., 2022). The inverse problem has also progressed rapidly: fMRI has been used to recover visual features, object categories, language-like semantics, and generative latents for image reconstruction (Horikawa & Kamitani, 2017; Takagi & Nishimoto, 2023; Scotti et al., 2024; Tang et al., 2023; Feng et al., 2026). Our approach differs from most encoding and decoding by introducing an intermediate ROI-supervised representation that supports selective intervention before downstream decoding.

In silico experiments to map semantic selectivity in the brain. Different methods have been proposed to study categorical selectivity in the brain. Interpretable encoding models can show which image regions are attended to predict brain activity (Adeli et al., 2025). Encoding models can also be used to generate maximally activating stimuli for different brain regions (Luo et al., 2023; Hwang et al., 2025) or create counterfactual tests of representational selectivity (Golbari et al., 2026). Decoding models can show what information can be read out from activity in different brain areas (Scotti et al., 2024; Feng et al., 2026). We build on these approaches and propose an in silico intervention method that tests how modifying the ROI-conditioned representation associated with a given brain region changes the decoded information, while keeping the downstream readout fixed. We implement this by making the ROI-conditioned components themselves the units of intervention.

Vision, language, and semantic spaces. Recent work shows that higher-level visual cortex contains semantic structure that can be expressed in language-derived embedding spaces (Doerig et al., 2025), and visual and linguistic similarity judgments converge toward related representational structure (Simkova et al., 2026). These results motivate us to use semantic caption decoding as the downstream task to reveal the effect of different interventions: DINOv3 provides frozen visual features, fMRI shapes the intermediate ROI token representation, and Qwen3-Embedding-8B defines the downstream caption-embedding target space.

3 Method

3.1 Overview

Figure 1 summarizes the ROI-tuning of image tokens, decoding ROI tokens, and intervention procedure. Stage 1 produces 50 ROI-tuned image tokens from the frozen visual features; ROI tokens reconstructed from measured fMRI responses using a voxel-to-token mapper are termed *decoded ROI tokens*. Given an image x, a frozen DINOv3 encoder produces a token sequence $H(x)$ containing spatial tokens that represent image patches, a class (CLS) token that provides an image-level representation, and auxiliary register tokens used for internal computation. Stage 1 maps these features to $R = 50$ ROI-conditioned tokens, $Z(x) = \{z_r(x)\}_{r=1}^{R}$, where $z_r(x) \in \mathbb{R}^{512}$. The token ROIs correspond to left- and right-hemisphere instances of visual ROIs from the probabilistic visual-topography atlas (ROI atlas). Each of 50 learned ROI queries independently cross-attends to the frozen DINOv3 sequence and produces a 512-dimensional token. ROI-specific MLP heads

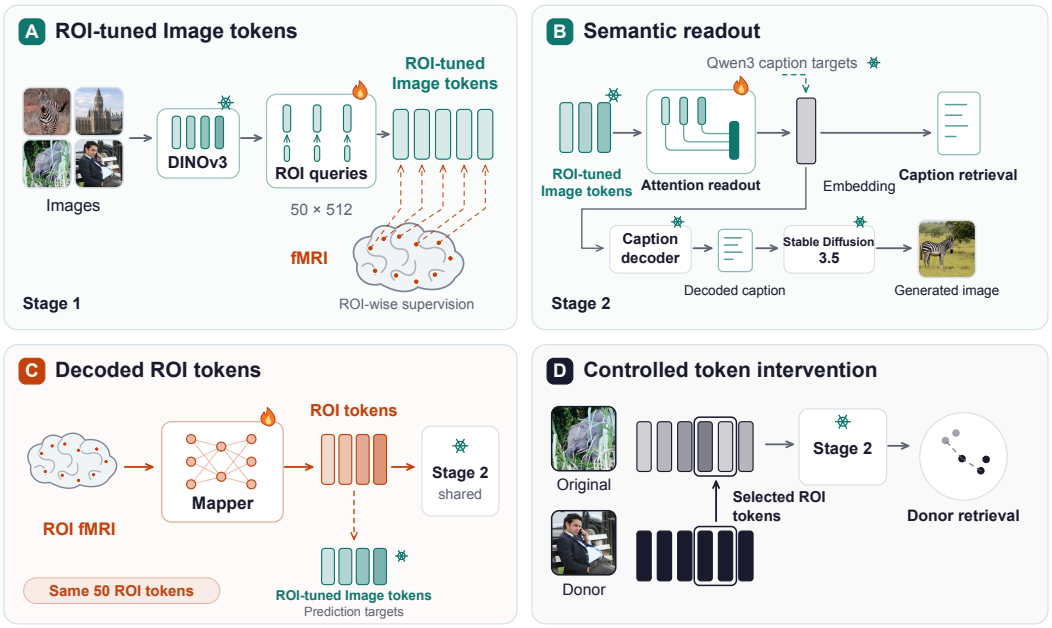

Figure 1: **ROI tokens as an intervenable intermediate representation.** **(A)** Stage 1 maps frozen DINOv3 image features into 50 ROI-tuned image tokens. Regional voxel responses supervise the corresponding tokens through ROI-specific prediction heads, without language supervision. **(B)** Stage 2 learns an attention readout from the frozen token set to the Qwen3 embedding space using fixed caption-derived targets. The predicted embedding supports caption retrieval and a frozen caption decoder, whose output conditions Stable Diffusion 3.5 for image generation. **(C)** Independent ROI-wise voxel-to-token MLPs map measured fMRI responses to 50 decoded ROI tokens, reconstructing the corresponding ROI-tuned image tokens. These decoded ROI tokens are evaluated through the same frozen image-trained Stage 2 readout. **(D)** Controlled interventions replace selected ROI tokens of an original image with the corresponding tokens of a donor image while keeping the readout fixed, allowing changes in donor-semantic retrieval to be measured. Snowflakes denote frozen components; flames denote components trained in the indicated stage. Dashed arrows indicate supervision. Brain locations and token arrays are schematic; only a subset of the ROI correspondences is shown. The complete ROI inventory is reported in Appendix A.2.

predict the corresponding fMRI responses; Stage 1 is trained only with voxel-prediction loss, with no language target contributing. Appendix A.3 gives the architecture and training settings.

Stage 2 is a lightweight semantic readout from the frozen token set to the Qwen3-Embedding-8B caption space. It uses a single learned query over the 50 tokens and predicts a 4096-dimensional embedding. Caption targets are computed offline by embedding up to five MS-COCO captions per image (Lin et al., 2014) with frozen Qwen3-Embedding-8B as text encoder, normalizing each caption embedding, and averaging them into an image-level target. Stage 2 is trained with a directional image-to-caption InfoNCE objective (van den Oord et al., 2018) with temperature $\tau_{\text{image}} = 0.07$ and then frozen for all primary intervention and brain-recovery analyses. Appendix A.4 gives the forward equations and training objective.

3.2 Controlled Token Interventions

For an *original* image $x^{(o)}$, a *donor* image $x^{(d)}$, and an ROI set G, we form an intervened representation

$$\tilde{z}_r(\alpha) = \begin{cases} (1-\alpha)z_r^{(o)} + \alpha z_r^{(d)}, & r \in G, \\ z_r^{(o)}, & r \notin G, \end{cases}$$

where hard intervention uses $\alpha = 1$. Appendix B provides the complete formulation, evaluation metrics, and size-matched null definition. The intervened tokens are passed through the same frozen

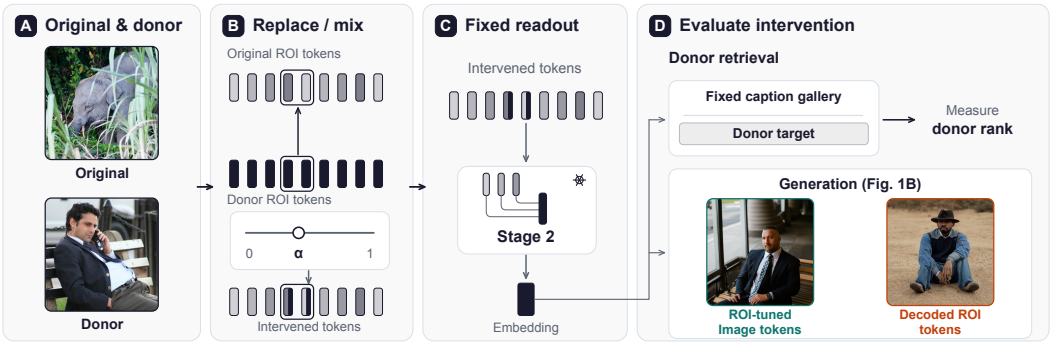

Figure 2: **Controlled original–donor token intervention.** **(A)** An original–donor stimulus pair supplies the two ROI-token sets. **(B)** Selected original tokens are replaced by, or mixed with, the corresponding donor tokens; all other tokens remain unchanged. **(C)** The intervened tokens pass through the same frozen image-trained Stage 2 readout to produce a semantic embedding. **(D)** Donor retrieval rank measures how the intervention changes access to the donor target in a fixed caption gallery. A separate frozen caption decoder produces text that conditions Stable Diffusion 3.5 for qualitative inspection; retrieved captions are not used for generation. The procedure is applied separately to ROI-tuned image tokens and decoded ROI tokens. The generated images illustrate a High-level intervention on both token representations for the same original–donor pair (S01, pair 6).

Stage 2 readout and evaluated by retrieval against the fixed 1000-way caption gallery. Intervention analyses using ROI-tuned image tokens and decoded ROI tokens use the same 1000 original–donor pairs drawn from *shared1000*, the held-out set of 1000 images shared across subjects in the Natural Scenes Dataset (NSD) (Allen et al., 2022). Pair ordering, ROI groups, retrieval gallery, and semantic readout are identical across the two token types. Figure2 illustrates the procedure.

3.3 Decoding ROI tokens from fMRI

To recover the token representation from fMRI, we train subject-specific voxel-to-token mappers with 50 independent ROI-wise MLPs. For ROI r, the measured voxel response v_r is mapped as

$$v_r \to 1024 \to \mathrm{GELU} \to 1024 \to \mathrm{GELU} \to 512,$$

producing $\widehat{z}_r$. We initialize the voxel-to-token mapping with token-space reconstruction,

$$\mathcal{L}_{\mathrm{tok}} = \mathcal{L}_{\mathrm{MSE}} + \mathcal{L}_{1-\cos},$$

and subsequently optimize the same architecture with a multi-term objective through the frozen semantic readout. Only mapper parameters are trained. Stage 2 remains frozen, the embedded targets remain offline and frozen, and gradients propagate through the frozen Stage 2 only to update the mapper. The objective is

$$\lambda_{\mathrm{tok}}\mathcal{L}_{\mathrm{tok}} + \lambda_{\mathrm{sem}}\mathcal{L}_{\mathrm{sem}} + \lambda_{\mathrm{ret}}\mathcal{L}^{\mathrm{sym}}_{\mathrm{InfoNCE}}.$$

The symmetric retrieval loss used for mapper fine-tuning has temperature $\tau_{\mathrm{mapper}} = 0.05$. The semantic term aligns the Stage 2 representation produced from decoded ROI tokens with that produced from true image-derived ROI tokens; the retrieval term is symmetric InfoNCE against the text encoder caption targets. Loss weights and optimization settings are reported in the appendix.

4 Experimental Setup

We evaluate on NSD subjects 01, 02, 05, and 07. Single-trial fMRI beta estimates on the fsaverage cortical surface are z-scored within session, averaged across available stimulus repetitions, and restricted to the atlas-defined ROIs.

Stage 1, Stage 2, and the voxel-to-token mapper are trained and selected within each subject's train9k set, using held-out validation subsets. The mapper is initialized by token reconstruction, with its

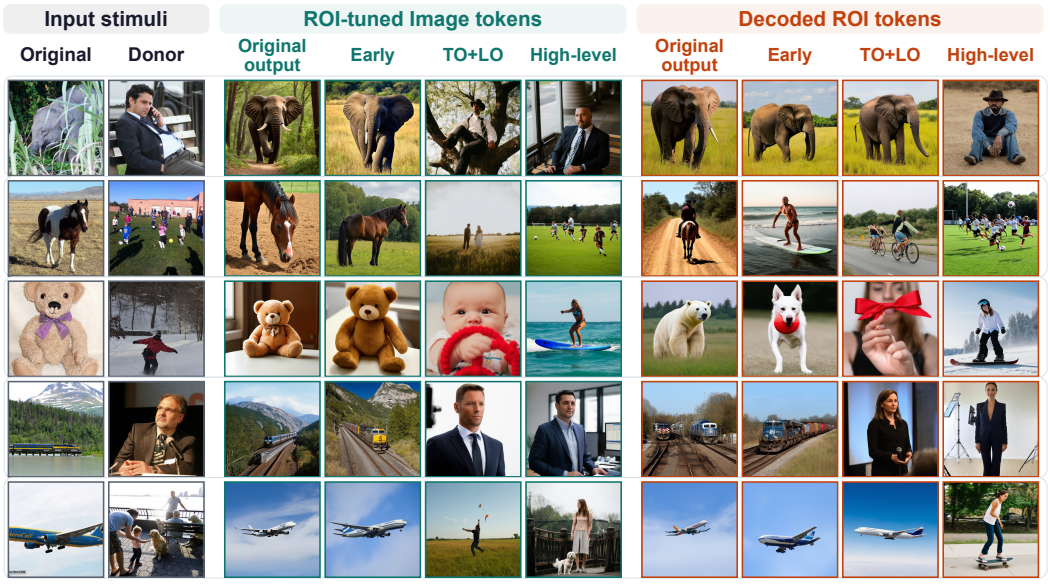

Figure 3: **Generated-image effects of ROI token intervention.** Each row shows an original–donor pair and caption-conditioned generations from ROI-tuned Image tokens (teal) and decoded ROI tokens (orange), before intervention and after replacing Early, TO+LO, or High-level tokens with corresponding donor tokens.

checkpoint selected by validation token cosine, then fine-tuned through the frozen image-trained Stage 2 readout. Shared1000 is reserved exclusively for final evaluation. Component-specific splits and selection criteria are detailed in Appendix A.5.

We measure semantic retrieval by ranking the 1000 shared1000 image-level caption targets by cosine similarity to the predicted embedding. Recall at rank k (R@k) is the fraction of queries whose correct image target appears among the top k, for $k \in \{1, 5, 10\}$; chance R@10 is 0.01. For intervention analyses, the correct target is the donor image. Primary semantic-intervention analyses use each subject's frozen image-trained Stage 2 attention readout and the same 1000 original–donor pairs across subjects and token types. Results are reported as mean $\pm$ sample standard deviation across subjects; pair-level bootstrap confidence intervals and empirical permutation p-values are identified separately.

5 Results

5.1 Semantic information in ROI-tuned image tokens

ROI-tuned image tokens support accurate caption retrieval, reaching $\mathrm{R@10} = 0.942 \pm 0.007$ in the 1000-image gallery, compared with a chance level of 0.01. An attention readout trained directly on the 201 frozen DINOv3 tokens reaches 0.965 ± 0.005. The 50-token ROI representation therefore retains much of the backbone's caption-retrieval performance. Stage 1 also predicts held-out fMRI responses, with a mean voxelwise Pearson correlation of $r = 0.411 \pm 0.061$ across subjects. Replacing the atlas ROIs with 50 random spatial regions and retraining both stages yielded chance-level caption retrieval ($\mathrm{R@10} = 0.010$ in every subject). Control details and limitations are reported in Appendix A.21.

5.2 ROI-tuned image token donor-semantic interventions

We next ask whether different groups of ROI-tuned image tokens have different effects on semantic retrieval. We replace the corresponding original ROI tokens with donor ROI tokens while keeping the semantic readout fixed. We consider six predefined groups: early visual cortex (Early), V1, LO, TO, TO+LO, and High-level (Appendix A.2).

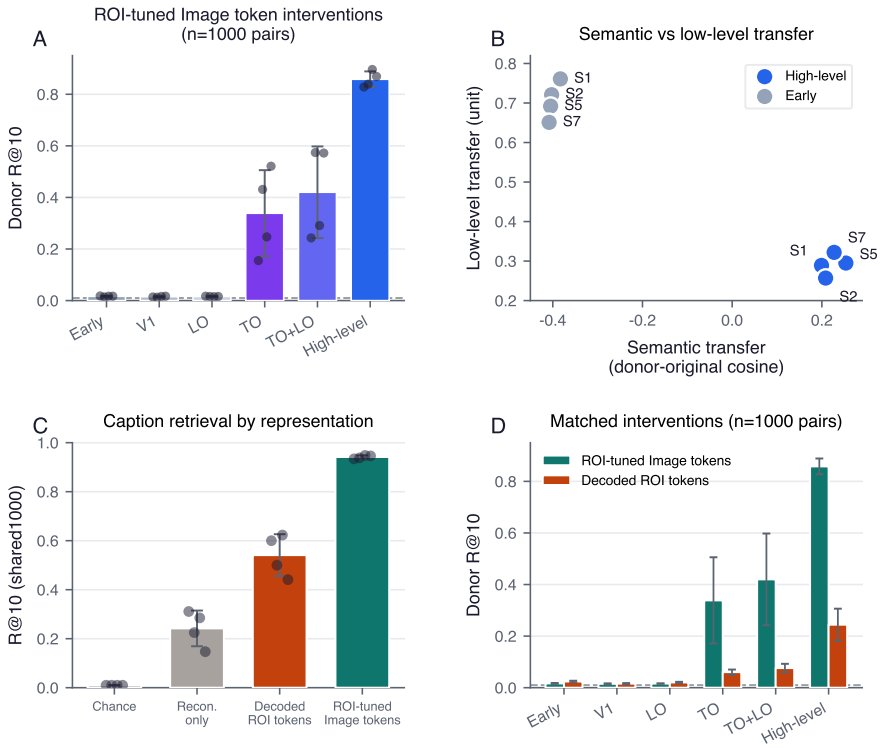

Figure 4: **Functional organization and brain recovery.** **(A)** Donor R@10 after interventions on ROI-tuned Image tokens for 1000 fixed original–donor pairs. **(B)** Semantic transfer (donor-minus-original cosine) versus normalized low-level transfer for High-level (blue) and Early (gray) interventions. Points represent subjects; low-level probes are independent of Stage 2. **(C)** Caption retrieval on shared1000 for chance, reconstruction-only decoded ROI tokens, task-aligned decoded ROI tokens, and ROI-tuned Image tokens. **(D)** Matched interventions on ROI-tuned Image tokens and decoded ROI tokens using the same 1000 original–donor pairs. Bars in A, C, and D show means with sample SD across four subjects; dots in A and C show individual subjects. Dashed lines indicate chance R@10 $= 0.01$.

Donor retrieval depends strongly on which token group is replaced (Figure 4A). Replacing High-level tokens shifts retrieval toward the donor, reaching R@10 $= 0.858 \pm 0.031$, compared with 0.017 ± 0.002 for Early tokens. TO also produces greater donor retrieval than LO (0.339 ± 0.167 versus 0.016 ± 0.001). The High-level–Early difference is consistent across subjects, with pair-level bootstrap confidence intervals excluding zero in all four (Appendix A.16).

This pattern is not explained by the number of replaced ROI tokens: High-level, TO, and TO+LO interventions exceed size-matched random-subset replacements in every subject ($p_{\text{gt}} \leq 0.004$). Gradually mixing donor tokens into the original representation produces the same overall pattern, with donor retrieval increasing for High-level, TO, and TO+LO tokens but changing little for Early tokens (Figure 7).

5.3 Visual transfer, readout dependence, and cross-subject consistency

The weak effect of Early ROI token interventions on semantic retrieval does not mean that Early tokens carry little visual information. Early interventions produce greater normalized transfer on low-level probes than High-level interventions in all four subjects (Figure 4B). These probes measure color, spatial layout, edge structure, and early DINO features independently of Stage 2. The semantic readout shows the opposite pattern: donor-minus-original cosine is $+0.222 \pm 0.024$ for High-level interventions and -0.400 ± 0.011 for Early interventions.

Table 1: **fMRI-to-token recovery.** Values are mean ± sample SD across $n = 4$ subjects on shared1000. Token cosine compares the flattened decoded and ROI-tuned image token representations for corresponding images. Stage 2 embedding cosine compares the corresponding outputs of the frozen semantic readout. The direct fMRI baseline uses the same response coordinates, with model selection performed on held-out data within train9k.

Path	R@10	Token cosine	Stage 2 embedding cosine
ROI-tuned image tokens	0.942 ± 0.007	–	–
Direct fMRI → Qwen embedding	0.525 ± 0.083	–	–
Decoded ROI tokens	0.541 ± 0.085	0.896 ± 0.010	0.391 ± 0.059

These effects depend on how the tokens are read out. In a matched comparison, mean pooling achieves similar overall retrieval performance to attention ($\mathrm{R@10} = 0.909 \pm 0.012$ versus 0.930 ± 0.009), but does not preserve the intervention ordering. Attention allocation across ROI groups is reported in Appendix A.19. Separately trained ridge decoders retrieve caption targets above chance from individual tokens, including Early tokens. Thus, information can be recoverable from a token even when replacing it has little effect under the learned attention readout.

The intervention patterns are also similar across subjects. Fingerprints combining the six ROI groups and five measurements have a mean pairwise Spearman correlation of $\rho = 0.963$ across the four subjects (Figure 5). The mean correlation remains at least 0.934 when any one measurement is removed (Appendix A.9), so no single measurement accounts for the agreement.

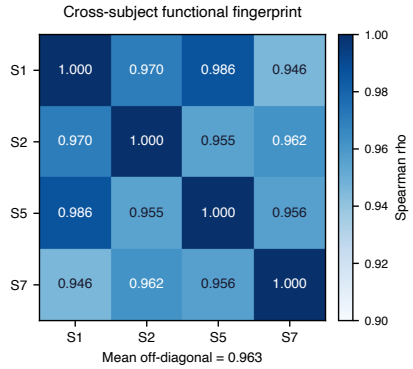

Figure 5: **Cross-subject reproducibility of the functional fingerprint.** Each subject is represented by the same six ROI groups and five measurements. Definitions and sensitivity analyses appear in Appendix A.22.

5.4 fMRI-to-token recovery

The voxel-to-token mapper predicts ROI tokens from fMRI responses, which are evaluated using the frozen image-trained Stage 2 readout. Decoded ROI tokens achieve $\mathrm{R@10} = 0.541 \pm 0.085$, with $\mathrm{R@1} = 0.138 \pm 0.044$ and $\mathrm{R@5} = 0.389 \pm 0.083$.

As a direct decoding baseline, we train an fMRI-to-Qwen embedding decoder using the same 26,400 response coordinates, with model selection restricted to held-out data within train9k. This baseline reaches $\mathrm{R@10} = 0.525 \pm 0.083$ on shared1000, giving similar retrieval performance for the direct and token-mediated decoding paths.

Decoded ROI tokens have a mean cosine similarity of 0.896 ± 0.010 to the corresponding ROI-tuned image tokens. Because raw token cosine contains a substantial component shared across images, we additionally report a centered similarity analysis in Appendix A.20. The cosine between the Stage 2 embeddings produced by decoded ROI tokens and their ROI-tuned image token counterparts is 0.391 ± 0.059. Retrieval remains below the ROI-tuned image token reference of $\mathrm{R@10} = 0.942 \pm 0.007$, indicating substantial but incomplete recovery from fMRI.

5.5 Intervention ordering in decoded ROI tokens

We repeat the matched interventions on decoded ROI tokens. The same broad ordering is preserved, although with smaller effects than for ROI-tuned image tokens (Figure 4D; Table 2). High-level donor $\mathrm{R@10}$ decreases from 0.858 ± 0.031 for ROI-tuned image tokens to 0.244 ± 0.062 for decoded ROI tokens. For decoded ROI tokens, the High-level–Early difference is 0.220 ± 0.059. This difference is positive in every subject, with pair-level bootstrap confidence intervals excluding zero; High-level interventions also produce a better donor rank than Early interventions for 88.4% of pairs on average (Appendix A.16).

Table 2: **Matched intervention comparison.** Donor R@10 for 1000 fixed original–donor pairs from shared1000. Values are mean ± sample SD across $n = 4$ subjects. Reconstruction-only ROI tokens are evaluated before task-aligned mapper fine-tuning. For both ROI-tuned image tokens and decoded ROI tokens, subject-level pair-bootstrap confidence intervals for the High-level–Early difference exclude zero.

Representation	Early	V1	LO	TO	TO+LO	High-level
ROI-tuned image tokens	0.017 ± 0.002	0.015 ± 0.002	0.016 ± 0.001	0.339 ± 0.167	0.420 ± 0.178	0.858 ± 0.031
Reconstruction-only ROI tokens	0.015 ± 0.004	0.014 ± 0.003	0.015 ± 0.003	0.055 ± 0.014	0.065 ± 0.017	0.146 ± 0.051
Decoded ROI tokens	0.024 ± 0.003	0.016 ± 0.002	0.020 ± 0.003	0.060 ± 0.010	0.076 ± 0.017	0.244 ± 0.062

Table 3: **Semantic decoding and generation.** Values are mean ± sample SD across $n = 4$ subjects. Caption and generated-image similarities provide supporting evidence for semantic recovery rather than detailed stimulus reconstruction.

Representation	Retrieval R@10	Caption semantic similarity	Generated-image DINO similarity
ROI-tuned image tokens	0.942 ± 0.007	0.777 ± 0.009	0.355 ± 0.008
Decoded ROI tokens	0.541 ± 0.085	0.648 ± 0.016	0.219 ± 0.031

The ordering is already present before task-aligned mapper fine-tuning. Reconstruction-only ROI tokens reach High-level donor $\mathrm{R@10} = 0.146 \pm 0.051$, compared with 0.015 ± 0.004 for Early interventions, arguing against the ordering arising solely from the task-aligned semantic objective. Continuous donor mixing likewise produces graded transfer for TO, TO+LO, and High-level groups, with much weaker effects for Early groups (Appendix A.13). These three groups also exceed size-matched random token subsets in every subject (Appendix A.14).

5.6 Semantic recovery in captions and generated images

We finally examine whether the recovered semantic information is visible in decoded captions and generated images. Stage 2 embeddings are decoded using a frozen ClipCap-style caption pathway (Mokady et al., 2021) with Qwen3-0.6B (Yang et al., 2025). The first decoded sentence is used for caption-semantic evaluation and to condition Stable Diffusion 3.5 Medium (Esser et al., 2024); image generation is therefore conditioned on decoded text rather than directly on the Stage 2 embedding.

Decoded ROI tokens preserve substantial coarse semantic content, although both caption similarity and generated-image similarity remain below the ROI-tuned image token reference (Table 3). Figure 6 shows examples stratified by decoded ROI-token retrieval performance, including top-1, top-10, and outside-top-10 retrieval cases.

6 Discussion

BrainInterv introduces a way to study regionally supervised visual representations through controlled intervention. Encoding models tell us how well a representation predicts neural responses, while decoding models test what information can be recovered from them. Here, we add a complementary operation: selected ROI tokens can be replaced while the rest of the representation and downstream readout are kept fixed. This makes it possible to test how changes to different ROI-associated components affect downstream semantic retrieval. Importantly, this intervenable representation retains much of the semantic retrieval performance of the original visual backbone while also predicting held-out fMRI responses.

Early ROI token interventions have stronger effects on lower-level visual measurements, including color, spatial layout, edge structure, and early-DINO features, whereas TO and High-level interventions produce substantially greater donor-semantic transfer. The same overall pattern is observed across subjects. We interpret this as a group-level difference in how the ROI-supervised representation contributes to visual and semantic decoding of natural scenes (Figure 3).

The fMRI pathway extends the same analysis to representations recovered from measured brain responses. Decoded ROI tokens reach semantic retrieval comparable to a direct fMRI-to-Qwen de-

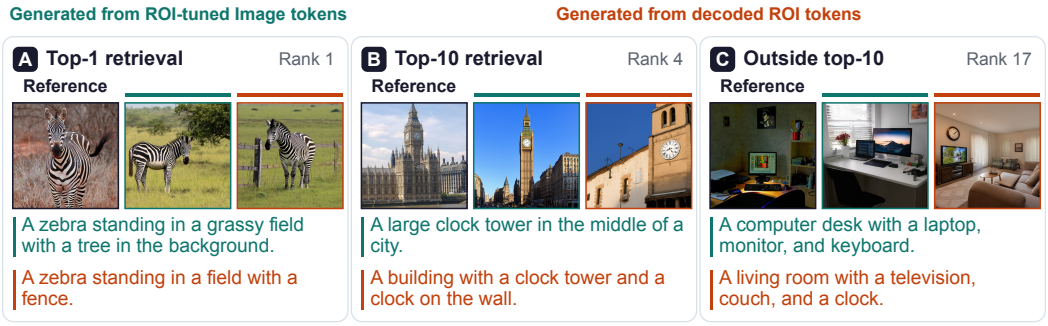

Figure 6: **Semantic recovery in captions and generated images from ROI-tuned image tokens and decoded ROI tokens.** Panels A–C show from top-1, top-10, and outside-top-10 decoded-ROI-token retrieval groups. For each reference image, the ROI-tuned image-token and decoded ROI-token representations are decoded to captions and then to Stable Diffusion 3.5 images. Teal denotes ROI-tuned image-token outputs and orange denotes decoded ROI-token outputs. The number in the upper-right corner of each panel gives the rank of the correct image-level caption target among the 1000 SHARED1000 targets, based on cosine similarity to the decoded-ROI-token Stage 2 embedding. Captions show the first decoded sentence.

coder, but the purpose of the intermediate token space is not to improve retrieval over direct decoding. Rather, it provides a structured representation in which the same ROI groups can be selectively manipulated and compared across the image and fMRI pathways. The same broad group ordering observed for ROI-tuned image tokens is also present in decoded ROI tokens and is already visible in reconstruction-only ROI tokens before task-aligned mapper fine-tuning. This makes it possible to carry the intervention analysis from a brain-supervised image representation to a representation reconstructed from measured fMRI.

More generally, the framework moves beyond describing regional selectivity by testing how selectively changing ROI-supervised representations affects downstream computation. The same approach can be extended to other ROI definitions, visual representations, and downstream tasks.

Several aspects of the interpretation remain specific to the present model and experimental setting. The intervention ordering depends on the downstream readout: mean pooling preserves strong overall retrieval but does not preserve the same ordering. The results therefore describe the behavior of the ROI-token representation together with the tested semantic readout, rather than a readout-independent organization of the token space. The current experiments are also limited to four NSD subjects, one visual-topography atlas, one visual backbone, and one semantic readout. Testing other ROI definitions, architectures, tasks, and subject-generalization settings will be important for determining which aspects of the observed organization generalize.

Finally, the interventions are performed on model representations, not on cortical activity. Their role is therefore to characterize how a brain-supervised computational representation behaves when selected regional components are changed. They can generate and refine hypotheses about regional contributions to natural-scene processing, but they should not be interpreted as evidence for the effects of biological intervention.

7 CONCLUSION

BrainInterv combines neural encoding, semantic decoding, and controlled intervention within a shared ROI-token representation. Selective token replacement reveals stronger low-level visual effects for Early ROI tokens and greater donor-semantic transfer for TO and High-level groups, with a broad group pattern that is consistent across subjects and also present in ROI tokens decoded from fMRI. By making regionally supervised representations selectively intervenable, BrainInterv provides a framework for testing how different regional representations affect downstream computation.

Reproducibility Statement

We describe the architecture, objectives, data splits, controls, and evaluation procedures in the main text and appendix. Code to reproduce all experiments will be released in an open-source repository. All reported decoded ROI token results use a held-out validation split of train9k for model selection, with shared1000 reserved for final evaluation.

AI Use Statement

Generative AI tools were used to support research brainstorming, refinement of analyses and controls, literature exploration, manuscript organization, and editing. All experimental results, numerical values, methodological descriptions, citations, and AI-assisted text included in the final manuscript were reviewed and verified by the authors against the underlying implementations, experimental outputs, and original sources. The authors made all final decisions regarding methodology, interpretation, claims, and manuscript content.

References

Hossein Adeli, Minni Sun, and Nikolaus Kriegeskorte. Transformer brain encoders explain human high-level visual responses. In *Advances in Neural Information Processing Systems*, volume 38, pp. 63288–63319, 2025. doi: 10.52202/085713-1902.

Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, 2022. doi: 10.1038/s41593-021-00962-x.

Adrien Doerig, Tim C. Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. High-level visual representations in the human brain are aligned with large language models. *Nature Machine Intelligence*, 7(8):1220–1234, 2025. doi: 10.1038/s42256-025-01072-0.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 12606–12633, 2024.

Pinyuan Feng, Hossein Adeli, Wenxuan Guo, Fan L. Cheng, Ethan Hwang, and Nikolaus Kriegeskorte. Towards interpretable visual decoding with attention to brain representations. *International Conference on Learning Representations*, 2026.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021. doi: 10.48550/arXiv.2106.02997.

Yuval Golbari, Navve Wasserman, Matias Cosarinsky, Roman Beliy, Aude Oliva, Antonio Torralba, Michal Irani, and Tamar Rott Shaham. From activation to specificity: Automating counterfactual testing of visual representations in the human brain. *arXiv preprint arXiv:2605.23895*, 2026. doi: 10.48550/arXiv.2605.23895.

Umut Güçlü and Marcel A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27): 10005–10014, 2015. doi: 10.1523/JNEUROSCI.5023-14.2015.

Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8:15037, 2017. doi: 10.1038/ncomms15037.

Ethan Hwang, Hossein Adeli, Wenxuan Guo, Andrew Luo, and Nikolaus Kriegeskorte. In silico mapping of visual categorical selectivity across the whole brain. *Advances in Neural Information Processing Systems*, 2025.

Michael A. Lepori, Alexa R. Tartaglini, Wai Keen Vong, Thomas Serre, Brenden M. Lake, and Ellie Pavlick. Beyond the doors of perception: Vision transformers represent relations between objects. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ed99503856b2ba85092e5413add65d86-Paper-Conference.pdf.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

Andrew Luo, Maggie Henderson, Leila Wehbe, and Michael Tarr. Brain diffusion for visual exploration: Cortical discovery using large scale generative models. *Advances in Neural Information Processing Systems*, 36:75740–75781, 2023.

Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

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

Paul Steven Scotti, Mihir Tripathy, Cesar Torrico, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq Mathew Abraham. MindEye2: Shared-subject models enable fMRI-to-image with 1 hour of data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44038–44059, 2024.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. doi: 10.48550/arXiv.2508.10104.

Katerina M. Simkova, Adrien Doerig, Clayton Hickey, and Ian Charest. Representations in vision and language converge in a shared, multidimensional space of perceived similarities. *Journal of Vision*, 26(5):7, 2026. doi: 10.1167/jov.26.5.7.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14453–14463, 2023. doi: 10.1109/CVPR52729.2023.01389.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023. doi: 10.1038/s41593-023-01304-9.

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

Andrew D. Vigotsky, Gian Domenico Iannetti, and A. Vania Apkarian. Mental state decoders: Game-changers or wishful thinking? *Trends in Cognitive Sciences*, 28(10):884–895, 2024. doi: 10.1016/j.tics.2024.06.004.

Liang Wang, Ryan E. B. Mruczek, Michael J. Arcaro, and Sabine Kastner. Probabilistic maps of visual topography in human cortex. *Cerebral Cortex*, 25(10):3911–3931, 2015. doi: 10.1093/cercor/bhu277.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. doi: 10.1073/pnas.1403112111.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. doi: 10.48550/arXiv.2506.05176.

A Appendix

A.1 Data and Preprocessing

We use the NSD subjects 01, 02, 05, and 07 with GLMdenoise surface betas in fsaverage space. Responses are session-z-scored, averaged over repeated presentations, aligned to train9k and shared1000 splits, and masked with the 50 hemisphere-specific visual ROIs from the probabilistic visual-topography atlas.

Table 4: **Dataset and preprocessing summary.** Subjects share the same final evaluation image identities.

Item	Setting
Subjects	NSD 01, 02, 05, 07
Beta representation	`betas_fithrf_GLMdenoise_RR`, fsaverage surface
Response preparation	Session-wise z-scoring before averaging repeated stimulus responses
ROI masking	50 hemisphere-specific probabilistic visual-topography atlas slots
Image sets	Subject-specific train9k; shared1000 for final evaluation only
Readout/mapper split	8000 training and 1000 validation images within train9k; seed 42
Pair manifests	Seeds 0–4; seed 0 defines primary results
Subset null	1000 random subsets per subject/group; seed 0

A.2 Exact ROI Token Definitions

Slots retain the same ordering across subjects, with separate left- and right-hemisphere tokens. Table 5 gives all 50 ROI slots, zero-based indices, and effective per-slot mapper input dimensions. These dimensions are identical across the four subject-specific mappers. The PHC label denotes parahippocampal cortex.

Table 5: **ROI slots and effective mapper input dimensions.** Indices are zero-based. Each dimension is identical across the four subject-specific mapper checkpoints.

ROI	LH slot	LH input dim.	RH slot	RH input dim.
V1v	0	1106	25	1035
V1d	1	1047	26	1088
V2v	2	920	27	889
V2d	3	723	28	859
V3v	4	600	29	626
V3d	5	714	30	678
hV4	6	410	31	504
VO1	7	303	32	231
VO2	8	348	33	419
PHC1	9	385	34	162
PHC2	10	280	35	424
TO2	11	10	36	152
TO1	12	420	37	385
LO2	13	348	38	316
LO1	14	448	39	568
V3B	15	441	40	453
V3A	16	690	41	800
IPS0	17	1060	42	835
IPS1	18	888	43	682
IPS2	19	898	44	612
IPS3	20	586	45	804
IPS4	21	127	46	134
IPS5	22	52	47	15
SPL1	23	232	48	285
FEF	24	308	49	100

Table 6: **Exact ROI-group membership.** Indices refer to Table 5; left/right slots remain separate. PHC is a component of the high-level group, not a separate primary fingerprint row.

Group	ROI stems	k	Left slots	Right slots
Early	V1d, V1v, V2d, V2v, V3d, V3v	12	0, 1, 2, 3, 4, 5	25, 26, 27, 28, 29, 30
V1	V1d, V1v	4	0, 1	25, 26
LO	LO1, LO2	4	13, 14	38, 39
TO	TO1, TO2	4	11, 12	36, 37
TO+LO	LO1, LO2, TO1, TO2	8	11, 12, 13, 14	36, 37, 38, 39
High-level	LO1, LO2, PHC1, PHC2, TO1, TO2, VO1, VO2, hV4	18	6, 7, 8, 9, 10, 11, 12, 13, 14	31, 32, 33, 34, 35, 36, 37, 38, 39
PHC	PHC1, PHC2	4	9, 10	34, 35

A.3 Stage-1 Architecture and Optimization

Stage 1 uses a per-ROI cross-attention encoder with a frozen DINOv3 ViT-B/16 backbone. The backbone's final hidden-state tokens are projected from 768 to 512 dimensions; the key/value sequence contains one CLS token, four register tokens, and 196 spatial patch tokens. The learned ROI tokens and positional parameters each have shape $1 \times 50 \times 512$. Each ROI query independently attends to this 201-token sequence through its own cross-attention block; there is no ROI–ROI self-attention or other cross-ROI mixing. ROI-specific heads predict the corresponding fMRI responses, with 26,400 output coordinates in total. For a batch of B images and n_r response coordinates in ROI r, the objective is

$$\mathcal{L}_{\text{Stage1}} = \frac{1}{50} \sum_{r=1}^{50} \frac{1}{B n_r} \sum_{i=1}^{B} \sum_{v=1}^{n_r} (\widehat{y}_{irv} - y_{irv})^2.$$

Thus each ROI contributes equally, irrespective of its number of response coordinates. No language target contributes to this objective. On held-out fMRI responses, Stage 1 achieves a mean voxelwise

Table 7: **Stage-1 architecture and optimization.** All four subjects use the same architecture and hyperparameters.

Component	Setting
Backbone	`facebook/dinov3-vitb16-pretrain-lvd1689m`, frozen
Input preprocessing	224×224; `DINOv3ViTImageProcessor` resize and normalization; mean (0.485, 0.456, 0.406), SD (0.229, 0.224, 0.225)
Features	`last_hidden_state`; 201 tokens, 768 coordinates; affine $768 \to 512$; patch size 16, spatial grid 14×14
ROI attention	50 independent learned queries; one block per ROI; 8 heads, 64 coordinates/head; query/key/value width 512
Block	Pre-LayerNorm on query and key/value; attention residual; feed-forward $512 \to 2048 \to 512$, GELU and dropout, with residual; dropout 0.1
ROI prediction head	LayerNorm, affine $512 \to 1024$, GELU, dropout, affine $1024 \to n_r$
Output	50 ROI tokens of width 512; 26,400 response coordinates
Optimization	AdamW; learning rate 3×10^{-4}; weight decay 0.01; batch 32; 15 epochs; cosine annealing; gradient norm clipped at 1.0
Split and selection	8100/900 within train9k; random split seed 42; minimum validation voxel loss
Selected epoch	S1: 5; S2: 4; S5: 5; S7: 4

Pearson correlation of $r = 0.411 \pm 0.061$ across subjects. The subject-level values are 0.420, 0.421, 0.476, and 0.329 for S1, S2, S5, and S7, respectively. Mean parcel-level correlations are highest for TO (0.564) and PHC (0.557), followed by VO (0.520); the corresponding means are 0.411 for Early visual cortex and 0.289 for the parietal group.

A.4 Stage-2 Architecture and Optimization

The readout uses a learned query over ROI tokens and predicts a 4096-dimensional caption embedding. The learned parameter dimensions are summarized in Table 8. For each token z_r,

$$k_r = W_k z_r + b_k, \qquad v_r = W_2\,\mathrm{GELU}(W_1 z_r + b_1) + b_2, \tag{1}$$

$$a_r = \frac{\exp(q^\top k_r/\sqrt{128})}{\sum_{j=1}^{50}\exp(q^\top k_j/\sqrt{128})}, \qquad F(Z) = \sum_{r=1}^{50} a_r v_r. \tag{2}$$

Key and value projections are shared across slots; attention softmax is over the 50 ROI slots. The readout has no LayerNorm, dropout, residual block, or additional projection after attention pooling. Its output is unnormalized; loss and retrieval computations apply L2 normalization.

Let $p_i = F(Z_i)$, let t_i be the corresponding image-level caption target, and define $s_{ij} = \cos(p_i, t_j)/\tau$ in a batch of B images. The image readout minimizes the directional objective

$$\mathcal{L}_{\rightarrow} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(s_{ii})}{\sum_{j=1}^{B}\exp(s_{ij})}, \qquad \tau = 0.07.$$

For mapper fine-tuning, the symmetric form instead averages the two directions:

$$\mathcal{L}^{\mathrm{sym}}_{\mathrm{InfoNCE}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(s_{ii})}{\sum_{j=1}^{B}\exp(s_{ij})} + \log\frac{\exp(s_{ii})}{\sum_{j=1}^{B}\exp(s_{ji})}\right], \qquad \tau = 0.05.$$

When a cached target has a caption axis, the loader takes its arithmetic mean; an already image-level target is used directly. This averaging operation does not itself renormalize the mean; the resulting target is L2-normalized when computing either objective or retrieval similarities.

Table 8: **Semantic readout dimensions and training schedule.** Dimensions and schedule are shared across all four subjects.

Component	Setting
Input	50×512 ROI tokens
Learned query	One 128-dimensional vector
Key projection	Affine $512 \to 128$
Value projections	Affine $512 \to 1024$, GELU, affine $1024 \to 4096$
Output space	Qwen3-Embedding-8B, 4096 dimensions
Attention	Scaled dot product; softmax over ROI slots
Normalization/dropout	None in head; L2 normalization for objective and retrieval
Objective	Image-to-caption InfoNCE, $\tau = 0.07$
Optimizer	AdamW; weight decay 10^{-4}; constant learning rate
Training schedule	400 epochs; learning rate 10^{-3}; batch size 512
Checkpoint selection	Maximum held-out train9k R@10, checked every 10 epochs and at the final epoch; first maximum retained
Validation split	8000/1000 inside train9k; seed 42
Primary evaluations	Readout parameters frozen

As a direct-backbone reference, we train the same attention-readout architecture on the 201 frozen DINOv3 tokens, without the 50-slot ROI bottleneck. Checkpoints are selected using held-out validation data within train9k and evaluated once on shared1000. This baseline reaches $\mathrm{R@10} = 0.965 \pm 0.005$, compared with 0.942 ± 0.007 for image-derived ROI tokens; the subject-wise ratio of ROI-token to direct-DINOv3 R@10 is 0.975 ± 0.004.

Attention allocation and single-slot decoding. We average the Stage 2 attention weights over shared1000 and sum them within each predefined ROI group. Across subjects, the 18 High-level slots receive 0.592 ± 0.043 of the total attention mass, compared with 0.36 under uniform allocation. The four TO slots receive 0.254 ± 0.084, compared with 0.08 uniformly, whereas the 12 Early slots receive 0.116 ± 0.019, compared with 0.24 uniformly.

To measure linearly decodable information independently of the Stage 2 attention readout, we fit a separate ridge decoder from each 512-dimensional token to the caption-embedding target using train9k and evaluate it on shared1000. Every slot supports above-chance retrieval, with individual-slot R@10 values spanning approximately 0.60 to 0.76. Group averages are 0.666 ± 0.048 for Early, 0.663 ± 0.049 for V1, 0.709 ± 0.023 for LO, 0.740 ± 0.011 for TO, 0.726 ± 0.021 for PHC, and 0.719 ± 0.025 for High-level slots. Thus, linear decodability from an individual slot and influence on the learned attention readout are empirically distinct.

A.5 Training and Validation Splits

Model selection is restricted to train9k; Table 9 distinguishes training and selection from final evaluation for each component.

Table 9: **Training splits and model selection.** For the reported decoded ROI token analyses, model selection is contained within train9k and shared1000 is final-test-only.

Component	Training / selection	shared1000 role
Stage 1	8100/900 within train9k; minimum validation voxel loss	final evaluation
Image Stage 2	train9k with train9k-contained validation	final evaluation
Reconstruction initialization	train9k token-space reconstruction; held-out train9k token-cosine selection	final evaluation only
Task-aligned voxel-to-token mapper	8000/1000 split within train9k	final evaluation only
Matched transplant evaluation	no training; fixed 1000-pair evaluation	final evaluation only
Caption/generation evaluation	no model selection on shared1000	final evaluation only

A.6 ROI-Tuned Image Token Intervention Controls

The ROI-tuned image token intervention analyses, size-matched subset nulls, and dose–response experiments use the frozen image-trained Stage 2 attention readout and the same fixed 1000 original–donor pairs. High-level, TO, and TO+LO donor R@10 are 0.858 ± 0.031, 0.339 ± 0.167, and 0.420 ± 0.178, whereas Early, V1, and LO reach 0.017 ± 0.002, 0.015 ± 0.002, and 0.016 ± 0.001, respectively. These group-level results are shown in Figure 4A.

For the size-matched null, High-level per-subject p_{gt} ranges from 0.001 to 0.003, and TO and TO+LO satisfy $p_{\mathrm{gt}} \leq 0.004$. Early transfer falls below the corresponding null ($p_{\mathrm{gt}} \geq 0.97$), while LO and V1 do not differ significantly from it. The group-wise results are reported in Table 10.

The dose–response protocol uses $\alpha \in \{0, 0.1, 0.25, 0.5, 0.75, 0.9, 1.0\}$. Donor retrieval increases with donor mixing for TO, TO+LO, and High-level groups, while remaining weak for Early ROI tokens (Figure 7).

Table 10: **Image-token transplants versus size-matched subsets.** Donor R@10 is mean $\pm$ sample SD; null ranges span subject-specific null means. Probabilities are the maximum subject-wise upper-tail values from 1000 random subsets, rounded to three decimals. Full values remain in supplementary data.

Group	k	Observed	Null mean range	max p_{gt}
Early	12	0.017 ± 0.002	0.071–0.120	1.000
V1	4	0.015 ± 0.002	0.018–0.031	0.895
LO	4	0.016 ± 0.001	0.018–0.031	0.947
TO	4	0.339 ± 0.167	0.018–0.029	0.001
TO+LO	8	0.420 ± 0.178	0.031–0.062	0.004
High-level	18	0.858 ± 0.031	0.205–0.257	0.003

We additionally evaluate architecture-matched regional-assignment controls. With a fixed permutation of the ROI-to-response assignment and retraining of both stages, ROI-tuned image-token retrieval is at chance ($\mathrm{R@10} = 0.010$) in every subject. A random spatial 50-region parcellation likewise yields $\mathrm{R@10} = 0.010$ in every subject (Appendix A.21). These controls show sensitivity to the tested regional assignments under the training protocol, rather than establishing that the visual-topography atlas is uniquely necessary.

A.7 Single-Stem and Category-Stratified Transplants

We first replace the left- and right-hemisphere tokens belonging to each atlas stem while holding all other tokens fixed. TO1 has the largest mean donor-retrieval effect, but it also shows substantial variation across subjects. The remaining stems produce donor R@10 values close to chance.

Table 11: **Largest single-stem intervention effects.** Values are ROI-tuned image token donor R@10. The final row gives the range of cross-subject means across the remaining 21 stems.

Stem	S1	S2	S5	S7	Mean ± SD
TO1	0.032	0.231	0.076	0.238	0.144 ± 0.106
TO2	0.023	0.019	0.021	0.036	0.025 ± 0.008
PHC1	0.023	0.017	0.017	0.024	0.020 ± 0.004
PHC2	0.017	0.014	0.024	0.020	0.019 ± 0.004
Remaining stems	–	–	–	–	0.013–0.016

We next stratify the fixed transplant evaluation by donor-image category. The descriptive results do not show a selective PHC advantage for scenes or a TO advantage restricted to animate categories. Instead, PHC remains close to chance, while TO and the broader High-level group transfer donor identity across all four categories.

Table 12: **Category-stratified donor retrieval.** Entries are donor R@10, reported as mean ± sample SD across $n = 4$ subjects. Each column restricts evaluation to original–donor pairs whose donor belongs to the indicated category.

Group	Person	Animal	Vehicle	Scene
PHC	0.015 ± 0.004	0.017 ± 0.006	0.018 ± 0.009	0.023 ± 0.012
TO	0.345 ± 0.195	0.469 ± 0.209	0.289 ± 0.152	0.336 ± 0.182
High-level	0.846 ± 0.034	0.888 ± 0.022	0.882 ± 0.024	0.862 ± 0.018

A.8 Semantic and Low-Level Analyses

The semantic-versus-low-level analysis uses the frozen image-trained Stage 2 semantic donor $\mathrm{R@10}$ and donor-minus-original cosine. For early tokens these are 0.017 ± 0.002 and -0.400 ± 0.011, and for high-level tokens they are 0.858 ± 0.031 and $+0.222 \pm 0.024$.

Low-level transfer is measured using four fixed ridge probes targeting color, spatial layout, edge structure, and early-DINO features. The color, layout, and edge probes are fit on a fixed subset of 2,000 train9k images, while the early-DINO probe is fit on all 9,000 train9k images. All probes use fixed ridge regularization $\lambda = 25$, with no hyperparameter selection.

For each probe, donor transfer is measured by cosine similarity between the prediction from the transplanted tokens and the donor-image target. Scores are normalized so that untouched original tokens define 0 and complete 50-token donor replacement defines 1, and the four normalized probe scores are averaged. These low-level measurements are independent of Stage 2 and the Qwen semantic targets.

A.9 Cross-Subject Reproducibility

Each subject contributes a complete 6×5 matrix for Early, V1, LO, TO, TO+LO, and high-level groups. The five measurements are hard-transplant donor R@10, mean donor-minus-original cosine

margin, the size-matched-null score $z = (\mathrm{R@10}_{\mathrm{observed}} - \mu_{\mathrm{null}})/\sigma_{\mathrm{null}}$, normalized low-level visual transfer, and dose–response area. The low-level component is the Stage 2-independent normalized score averaged across the color, layout, edge, and early-DINO probes. Dose area is the trapezoidal integral of donor R@10 over $\alpha \in \{0, 0.1, 0.25, 0.5, 0.75, 0.9, 1\}$. The semantic components use the same frozen image-trained readout and fixed seed-0 pairs as the primary interventions. We flatten each matrix in the same group/feature order and compute Spearman correlations between the resulting 30-dimensional subject vectors, without imputation. The mean of the six unique off-diagonal correlations is 0.963 (Figure 5).

No single feature accounts for the high cross-subject agreement: the lowest leave-one-feature-out mean correlation is 0.934. Removing high-level, Early, or both groups also preserves high agreement (Table 13). As a simpler complementary check, correlations between the six-dimensional donor-R@10 group vectors average 0.902. All subjects follow high-level > TO+LO > TO, with these groups above Early/V1/LO; only the internal ordering among the latter groups varies. This analysis concerns intervention effects, not geometric similarity of individual image-token representations or localization of individual concepts.

Table 13: **Fingerprint sensitivity analyses.** Each entry is the mean of the six pairwise cross-subject Spearman correlations after the indicated component is omitted.

Analysis	Mean ρ
Omit donor R@10	0.975
Omit semantic cosine shift	0.968
Omit size-matched null z-score	0.934
Omit low-level transfer	0.954
Omit dose–response area	0.970
Omit High-level group	0.944
Omit Early group	0.963
Omit High-level and Early groups	0.944
Donor R@10 group vector only	0.902

A.10 Decoded ROI Token Results

A direct decoder from the same 26,400 fMRI response coordinates to the Qwen embedding space reaches $\mathrm{R@10} = 0.525 \pm 0.083$, using held-out train9k validation for model selection and shared1000 only for final evaluation. The final decoded ROI token results use the task-aligned voxel-to-token mapper selected entirely within train9k validation. On shared1000, decoded ROI tokens reach $\mathrm{R@1} = 0.138 \pm 0.044$, $\mathrm{R@5} = 0.389 \pm 0.083$, and $\mathrm{R@10} = 0.541 \pm 0.085$. The mean cosine similarity between corresponding decoded ROI tokens and ROI-tuned image tokens is 0.896 ± 0.010, compared with 0.779 ± 0.036 for shuffled image–brain pairs. The cosine between the decoded ROI token and ROI-tuned image token Stage 2 readout embeddings is 0.391 ± 0.059.

A.11 Voxel-to-Token Objective and Training

We first train the voxel-to-token mapper for token reconstruction and then perform task-aligned fine-tuning using the frozen Stage 2 semantic and retrieval objectives.

The reconstruction initialization is selected using held-out train9k token cosine and reaches $\mathrm{R@10} = 0.242 \pm 0.073$ on shared1000 when evaluated by the frozen image-trained readout. The final task-aligned mapper is then optimized with token reconstruction, Stage 2 semantic alignment, and symmetric InfoNCE terms. The loss weights are $(\lambda_{\mathrm{tok}}, \lambda_{\mathrm{sem}}, \lambda_{\mathrm{ret}}) = (0.03, 1.0, 1.0)$ for all four subjects. The complete mapper configuration is given in Table 14. Writing $\widehat{z}_{ir}$ and z_{ir} for predicted and image-derived tokens, the reconstruction objective averages squared error over all batch, slot, and feature coordinates and averages cosine error over batch/slot pairs:

$$\mathcal{L}_{\mathrm{tok}} = \frac{1}{BRD} \sum_{i,r,d} (\widehat{z}_{ird} - z_{ird})^2 + \frac{1}{BR} \sum_{i,r} [1 - \cos(\widehat{z}_{ir}, z_{ir})], \quad R = 50, \quad D = 512.$$

The task-aligned term is $\mathcal{L}_{\text{sem}} = B^{-1}\sum_i[1 - \cos(F(\widehat{Z}_i), F(Z_i))]$, and the caption retrieval term is symmetric InfoNCE as defined in Appendix A.4. The image-token semantic teacher is detached; Stage 2 parameters remain frozen while gradients propagate through the reconstructed-token branch to the mapper. Each ROI MLP has GELU after its first two affine layers, with no dropout, normalization, or cross-ROI mixing.

Table 14: **Subject-specific voxel-to-token mapper.** All readout parameters remain frozen; gradients pass through Stage 2 only to update mapper parameters.

Component	Setting
ROI mapping	50 independent subject-specific MLPs
Linear dimensions	ROI voxel count $\rightarrow 1024 \rightarrow 1024 \rightarrow 512$
Input dimensions	Per-slot counts in Table 5
Reconstruction initialization	80 epochs; learning rate 10^{-3}; held-out train9k token-cosine selection every 5 epochs and at the final epoch
Task-aligned fine-tuning	80 epochs; learning rate 10^{-4}
Optimizer/batch	AdamW, weight decay 10^{-4}, batch size 256 in both phases; no learning-rate scheduler
Task checkpoint selection	Held-out train9k R@10 every 10 epochs and at the final epoch; first maximum retained
Training/validation	8000/1000 images within train9k, seed 42
Loss weights, all subjects	$(\lambda_{\text{tok}}, \lambda_{\text{sem}}, \lambda_{\text{ret}}) = (0.03, 1, 1)$
Retrieval temperature	$\tau = 0.05$

A.12 Brain-Specific Semantic Readout Control

This control tests whether an image-trained semantic readout is mismatched to decoded ROI tokens. We freeze the final train9k-selected voxel-to-token mapper and initialize a brain-specific Stage 2 from the image-trained Stage 2. Only the readout is trained, using 8000 train9k images and 1000 held-out train9k images for validation and checkpoint selection by R@10. The tested schedule uses 100 epochs, learning rate 10^{-4}, and batch size 512; epoch 0 is included among the candidates. Shared1000 is reserved for final evaluation after selection.

Validation selected epoch 0 in all four subjects (Table 15). Both readouts therefore yield $\text{R@10} = 0.541 \pm 0.085$, with paired change 0.000 ± 0.000. The unchanged baseline also has $\text{R@1} = 0.138 \pm 0.044$ and $\text{R@5} = 0.389 \pm 0.083$. A secondary readout trained from scratch under train9k-contained selection reaches $\text{R@10} = 0.499 \pm 0.071$, a paired change of -0.042 ± 0.027, and is worse in every subject. Under this adaptation protocol, specializing the readout does not explain the remaining retrieval gap; this result does not rule out improvements from other adaptation procedures.

Table 15: **Brain-specific semantic readout control.** Shared1000 R@10 with the same frozen voxel-to-token mapper. The adapted readout starts from the image-trained readout; the scratch readout starts from random initialization. Delta is adapted minus image-trained R@10. Epoch denotes the selected adapted-readout epoch. The final row reports mean $\pm$ sample SD ($n = 4$).

Subject	Image-trained	Adapted	Delta	Epoch	Scratch
01	0.600	0.600	0.000	0	0.537
02	0.500	0.500	0.000	0	0.442
05	0.623	0.623	0.000	0	0.579
07	0.441	0.441	0.000	0	0.437
Mean $\pm$ SD	0.541 ± 0.085	0.541 ± 0.085	0.000 ± 0.000	–	0.499 ± 0.071

A.13 Continuous interventions on decoded ROI tokens

We mix corresponding original and donor decoded ROI token slots in native token space at $\alpha \in \{0, 0.1, 0.25, 0.5, 0.75, 0.9, 1.0\}$, keeping all other slots fixed. The final train9k-selected mapper and frozen image-trained Stage 2 remain fixed. The intervention uses the same fixed 1000 original–donor pairs as the hard-transplant analysis and the ROI-tuned image token dose curves, with shared1000

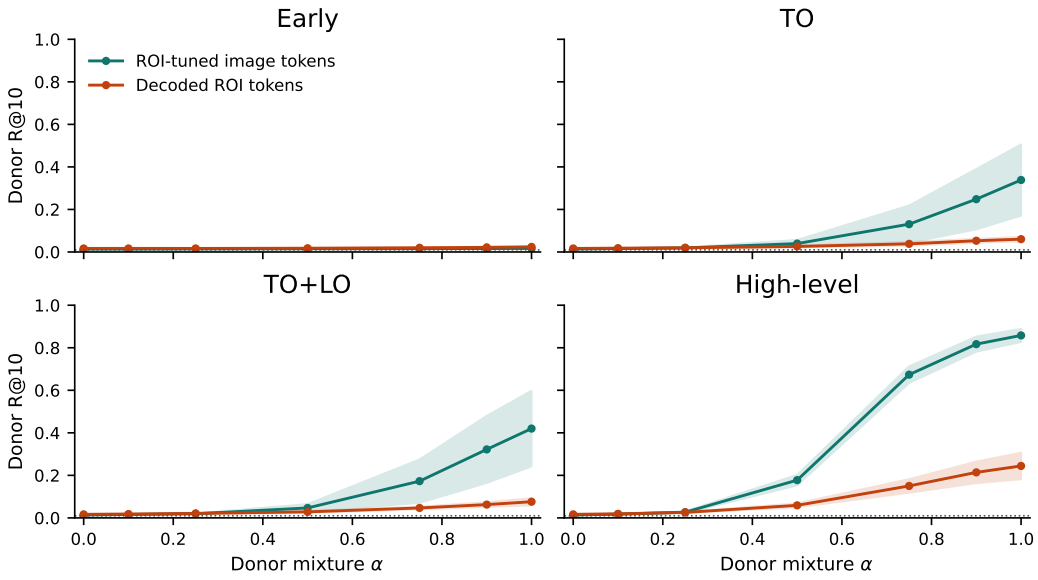

Figure 7: **Continuous donor mixing in ROI tokens.** Donor R@10 for ROI-tuned image tokens (teal) and decoded ROI tokens (orange) as Early, TO, TO+LO, or High-level tokens are mixed with corresponding donor tokens. The mixing coefficient α ranges from the original representation ($\alpha = 0$) to full donor replacement within the selected group ($\alpha = 1$); all other tokens remain unchanged. Curves show means and shaded bands show sample SD across four subjects at seven mixing levels. All panels use common axes; dotted lines indicate chance R@10 $= 0.01$.

used only for final evaluation. Figure 7 compares ROI-tuned image and decoded ROI tokens at all seven doses for Early, TO, TO+LO, and High-level groups. Mean donor transfer increases with α for TO, TO+LO, and high-level groups, reaching 0.060 ± 0.010, 0.076 ± 0.017, and 0.244 ± 0.062, respectively, at $\alpha = 1$. These effects remain substantially attenuated relative to image-derived tokens. Early, V1, and LO effects are weak; the V1 curve is flat and non-monotonic, and LO also shows a small final decrease.

A.14 Size-matched null for decoded ROI tokens

We compare each predefined reconstructed-token group with 1000 random subsets of the same cardinality, separately in each subject, using the fixed 1000 source–donor pairs. The mapper, gallery, and image-trained Stage 2 readout remain fixed. For observed donor retrieval T_G and random-subset values T_b, the upper-tail empirical probability is

$$p_{\mathrm{gt}} = \frac{1 + \sum_{b=1}^{1000} \mathbf{1}[T_b \geq T_G]}{1001}.$$

TO, TO+LO, and high-level reconstructed-token groups exceed size-matched random replacement in every subject (Table 16); Early, V1, and LO do not. These subject-wise comparisons do not imply that every predefined ROI group has a distinct donor-transfer effect.

A.15 Pair-Set Sensitivity

We evaluate five fixed sets of 1000 source–donor pairs (seeds 0–4) on the same shared1000 gallery and both representation paths. Every pair has distinct source and donor identities; each manifest is reused across all four subjects and both paths. Seed 0 remains the primary pair set throughout the main results. These manifests are sensitivity perturbations of an evaluation on the same images, not five independent datasets. For every manifest, subject-mean donor retrieval follows high-level $>$ TO+LO $>$ TO $>$ Early/V1/LO on both paths (Table 17). The image-to-brain attenuation of the TO, TO+LO, and high-level effects persists across all five manifests. Across groups and paths, the mean

Table 16: **Reconstructed-token transplants versus size-matched subsets.** Donor R@10 is mean $\pm$ sample SD; null ranges span subject-specific null means. Probabilities are the maximum subject-wise upper-tail values from 1000 random subsets, rounded to three decimals. Full values remain in supplementary data.

Group	k	Observed	Null mean range	max p_{gt}
Early	12	0.024 ± 0.003	0.039–0.053	0.968
V1	4	0.016 ± 0.002	0.019–0.023	0.998
LO	4	0.020 ± 0.003	0.019–0.023	0.668
TO	4	0.060 ± 0.010	0.020–0.023	0.005
TO+LO	8	0.076 ± 0.017	0.029–0.035	0.039
High-level	18	0.244 ± 0.062	0.070–0.106	0.002

Table 17: **Pair-manifest sensitivity of donor R@10.** Entries are means across four subjects for each fixed manifest; seed 0 remains the primary analysis.

Path	Seed	Early	V1	LO	TO	TO+LO	High-level
Image	0	0.017	0.015	0.016	0.339	0.420	0.858
Image	1	0.018	0.017	0.016	0.338	0.425	0.885
Image	2	0.019	0.019	0.019	0.324	0.406	0.869
Image	3	0.015	0.015	0.016	0.342	0.414	0.889
Image	4	0.025	0.022	0.025	0.345	0.419	0.876
Brain	0	0.024	0.016	0.020	0.060	0.076	0.244
Brain	1	0.024	0.019	0.019	0.053	0.068	0.254
Brain	2	0.026	0.017	0.017	0.052	0.070	0.243
Brain	3	0.020	0.017	0.015	0.052	0.065	0.247
Brain	4	0.031	0.025	0.025	0.059	0.082	0.246

subject-wise range (maximum minus minimum across the five manifests) is 0.008–0.032 absolute R@10. This range measures pair-set sensitivity, not uncertainty from independent subjects.

Pair construction uses caption-string heuristics: concatenated, lowercased captions are matched to substring lexicons for person, animal, vehicle, food, sport, and scene. The sampler requests 250 different-category pairs, 250 same-category pairs, 125 scene-source/scene-and-object-donor pairs, and 125 object-source/object-donor pairs; object masks combine person, animal, and vehicle. After removing duplicates and self-pairs, random distinct pairs fill the manifest to 1000. These masks do not verify scene or object identity. Caption-cosine tertiles annotate the pairs rather than balance sampling (286/318/396 low/medium/high pairs for seed 0). The exact lexicon and all five manifests remain in supplementary code and data.

A.16 Pair-Level Bootstrap Analysis

For each subject and representation path, we resample the fixed set of 1000 original–donor pairs with replacement $10{,}000$ times and recompute the High-level-minus-Early donor-R@10 difference. These confidence intervals quantify variation across the evaluated image pairs within each subject; cross-subject variation is reported separately in the main text.

A.17 Replacing decoded ROI tokens with ROI-tuned image tokens

This is a same-stimulus repair diagnostic: for each shared1000 image, we replace a predefined group of decoded ROI token slots with the corresponding ROI-tuned image tokens for that same image. All other slots and the frozen image-trained Stage 2 remain fixed; no original–donor transplantation is performed. The final train9k-selected mapper is frozen, and shared1000 is used only for final evaluation. For each subject and group, 1000 random subsets of the same token count provide a size-matched repair null. The upper-tail empirical probability is $p_{\mathrm{gt}} = (1 + \#\{\Delta_{\mathrm{random}} \geq \Delta_{\mathrm{group}}\})/1001$, where each delta is the paired R@10 change from the all-brain baseline.

Table 18: **Pair-level bootstrap of the High-level–Early intervention difference.** Delta is High-level minus Early donor R@10. Confidence intervals are computed separately within each subject. $P(\text{HL rank} < \text{Early rank})$ is the fraction of pairs for which the High-level transplant gives the better donor rank.

	ROI-tuned image tokens		decoded ROI tokens	
Subject	Δ [95% CI]	Rank fraction	Δ [95% CI]	Rank fraction
S1	0.810 [0.785, 0.834]	0.997	0.219 [0.193, 0.246]	0.866
S2	0.825 [0.801, 0.848]	0.996	0.225 [0.198, 0.253]	0.875
S5	0.879 [0.859, 0.899]	0.997	0.291 [0.262, 0.320]	0.926
S7	0.852 [0.830, 0.874]	0.994	0.146 [0.123, 0.170]	0.868

Replacing TO, TO+LO, and high-level groups improves retrieval more than size-matched random repair in all four subjects, whereas early, V1, and LO repair does not (Table 19). High-level repair increases R@10 from 0.541 ± 0.085 to 0.918 ± 0.011, compared with 0.942 ± 0.007 for all image-derived tokens. The high-level random-repair null itself gives a substantial mean gain (0.287 ± 0.055), so token count explains part of the improvement. These results identify task-relevant repair effects within the model; they do not localize the entire remaining error to high-level slots or imply effects of biological intervention.

Table 19: **Same-image ROI repair.** Repaired R@10, paired gain, and random gain are mean ± sample SD across subjects. k is the number of repaired slots. The final column is the maximum subject-wise upper-tail probability from 1000 random repairs.

Group	k	Repaired R@10	Paired gain	Random gain	max p_{gt}
Early	12	0.557 ± 0.098	$+0.0160 \pm 0.0197$	0.220 ± 0.036	1.000
V1	4	0.522 ± 0.094	-0.0193 ± 0.0095	0.094 ± 0.013	0.988
LO	4	0.590 ± 0.103	$+0.0495 \pm 0.0359$	0.091 ± 0.010	0.826
TO	4	0.864 ± 0.028	$+0.3232 \pm 0.1030$	0.093 ± 0.012	0.007
TO+LO	8	0.872 ± 0.019	$+0.3307 \pm 0.0998$	0.168 ± 0.025	0.023
High-level	18	0.918 ± 0.011	$+0.3772 \pm 0.0839$	0.287 ± 0.055	0.008

A.18 Generated Results

For caption visualization and caption-semantic similarity, we use the first decoded sentence of each greedy decode, matching the text used to condition Stable Diffusion 3.5; the auxiliary caption decoder was trained almost entirely on single-sentence caption targets. Figure 8 illustrates transplant-conditioned generations across subjects and pairs. Caption semantic similarity is the cosine similarity between the Qwen3-Embedding-8B embedding of the decoded caption's first sentence and the corresponding image-level caption target used by Stage 2. Generated-image similarity is the cosine similarity between DINOv3 embeddings of the reference NSD image and the Stable Diffusion 3.5 image generated from that sentence. We extract the first sentence of the captions. The caption reference is the mean-pooled Qwen3-Embedding-8B target embedding, rather than an individual reference caption. For generated-image evaluation, we use visual encoder of the DinoV3 ViT-B/16, taking the pooled output when available and otherwise the CLS token, followed by L2 normalization.

A.19 Attention allocation across ROI token groups

We examined the attention weights of the learned Stage 2 readout to characterize how strongly different ROI token groups contribute to the semantic readout. For ROI group G, we define its attention mass for image x as

$$A_G(x) = \sum_{r \in G} a_r(x),$$

where $a_r(x)$ is the Stage 2 attention weight assigned to ROI token r. We report A_G averaged across evaluation images and subjects.



Figure 8: **Example generations under ROI-token interventions.** Each example compares ROI-tuned image tokens and decoded ROI tokens for the same original–donor pair, showing the original output and the outputs after Early, TO+LO, and High-level interventions. Examples are median selections from intermediate and lower decoded-ROI-token donor-retrieval groups, restricted to cases where High-level transfer exceeds Early transfer in both representation paths. The generations illustrate intervention effects and decoding limitations rather than providing quantitative evidence for the intervention ordering. These examples are separate from the fixed 1000-pair evaluation.

The High-level group contains 18 of the 50 ROI tokens and receives approximately 59% of the total attention mass, compared with 36% under uniform allocation. In contrast, the 12 Early tokens receive approximately 12% of the attention mass, compared with 24% under uniform allocation. TO, which contributes four tokens and is included in the High-level group, receives approximately 25% of the total attention mass.

These weights show that the learned semantic readout places more emphasis on higher-level ROI tokens than would be expected from their token counts alone. However, attention weight should not be interpreted as a direct measure of intervention importance. The separate mean-pooling control shows that high retrieval performance can be retained even when the intervention ordering changes. We therefore use the attention analysis only to describe the allocation learned by the Stage 2 readout, rather than as an explanation for the ROI intervention effects.

A.20 Centered token-similarity control

The raw token cosine reported in Table 1 compares the flattened 50×512 decoded ROI-token representation with the corresponding ROI-tuned image token representation and yields 0.896 ± 0.010. This value includes representational components that are shared across images and therefore should not be interpreted directly as image-specific correspondence.

We additionally evaluate token similarity after removing this shared component. For each ROI slot, we subtract the mean token vector computed across the 1000 evaluation images from both the decoded and ROI-tuned image tokens. The 50 centered tokens are then flattened and cosine similarity is computed between the two representations.

After centering, corresponding decoded and ROI-tuned image token representations have a cosine similarity of 0.583 ± 0.045 across subjects $(0.591, 0.618, 0.606, 0.517)$. Shuffling the correspondence between fMRI and images reduces the centered cosine to 0.000 ± 0.000. Thus, the high raw cosine partly reflects structure shared across images, while substantial image-specific correspondence remains after this shared component is removed.

A.21 Random Spatial ROI Control

To test sensitivity to the atlas-defined ROI assignments, we replace the 50 visual ROIs with 50 random spatial ROIs over the same 26,400 fsaverage vertices, with 25 ROIs per hemisphere. ROIs are obtained by k-means on fsaverage spherical coordinates (seed 0), and the same assignment is used for train9k and shared1000.

Both Stage 1 and Stage 2 are retrained from scratch using the same training protocol as the atlas-based model. On shared1000, the resulting ROI-tuned image tokens reach $\mathrm{R@10} = 0.010$ in every subject, equal to chance.

The control matches the cortical vertex set and number of ROIs, but not their sizes; for subject 01, random ROIs contain a median of 536 voxels (range 100–651), compared with 450 (range 10–1106) for the atlas ROIs. Because only one random ROI assignment is tested, this result shows sensitivity to the tested regional assignment under the training protocol, rather than establishing that the visual-topography atlas is uniquely necessary.

A.22 Cross-subject functional fingerprint

For each subject, we summarize intervention behavior across six ROI groups (Early, V1, LO, TO, TO+LO, and High-level) using five measurements: donor R@10, donor-minus-original semantic cosine shift, a size-matched random-subset z-score, low-level visual transfer, and dose-response area. All measurements use the same fixed 1000 original–donor pairs and ROI-tuned image token interventions.

Low-level transfer averages four independently trained ridge probes for color, spatial layout, edge structure, and early-DINO features, normalized so that untouched original tokens define 0 and complete donor replacement defines 1. Dose-response area summarizes donor R@10 across $\alpha \in \{0, 0.1, 0.25, 0.5, 0.75, 0.9, 1\}$.

The resulting 6×5 matrix is flattened into a 30-dimensional fingerprint for each subject, and cross-subject similarity is measured using Spearman correlation.

B Controlled original–Donor Token Interventions

Interventions measure how replacing corresponding ROI tokens changes a fixed semantic readout (Figure 2). Mathematical indices are one-based, whereas implementation indices are zero-based.

B.1 Original and Donor Representations

For original and donor inputs x_o, x_d, let $Z_o = Z(x_o)$ and $Z_d = Z(x_d)$, with $Z(x) \in \mathbb{R}^{50\times512}$. The index set $G \subseteq \{1, \dots, 50\}$ specifies an ROI group (Table 6).

B.2 Hard Token Transplantation

Only slots in G take donor values:

$$\widetilde{z}_r^{(G)} = \begin{cases} z_r(x_d), & r \in G, \\ z_r(x_o), & r \notin G. \end{cases} \tag{3}$$

B.3 Continuous Token Transplantation

Interpolation takes place in native token space:

$$\widetilde{z}_r^{(G,\alpha)} = \begin{cases} (1-\alpha)z_r(x_o) + \alpha z_r(x_d), & r \in G, \\ z_r(x_o), & r \notin G. \end{cases} \tag{4}$$

We use $\alpha \in \{0, 0.1, 0.25, 0.5, 0.75, 0.9, 1\}$; the endpoints are the intact source and hard transplant.

B.4 Frozen Readout

The image-trained Stage 2 readout $F : \mathbb{R}^{50\times512} \to \mathbb{R}^{4096}$ produces original output $q_o = F(Z_o)$ and transplanted output $\widetilde{q} = F(\widetilde{Z}^{(G,\alpha)})$. Its parameters remain fixed throughout each intervention.

B.5 Measuring Donor Transfer

For original and donor semantic targets y_o, y_d, we measure $C_d = \cos(\widetilde{q}, y_d)$, $C_o = \cos(\widetilde{q}, y_o)$, and their margin $C_d - C_o$. Ranking the 1000 gallery targets by cosine similarity gives donor and original ranks r_d, r_o, with rank 1 best. Across $N = 1000$ matched pairs,

$$\mathrm{DonorR@}k = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[r_{d,i} \le k], \tag{5}$$

$$P_{\mathrm{donor}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[r_{d,i} < r_{o,i}]. \tag{6}$$

Both representation paths use identical pairs and gallery targets.

B.6 Matched-Subset Null Model

For each group, 1000 random subsets G_j^{rand} satisfy $|G_j^{\mathrm{rand}}| = |G|$. For donor retrieval T, the upper-tail probability is

$$p_{\mathrm{gt}} = \frac{1 + \sum_{j=1}^{1000} \mathbf{1}[T(G_j^{\mathrm{rand}}) \ge T(G)]}{1001}. \tag{7}$$

This tests the selected group against replacement of the same number of ROIs.

B.7 Relation to Existing Intervention Methods

Our procedure is related to interchange interventions and activation patching, which replace internal activations from one input with those from another and measure the resulting output change (Geiger et al., 2021; Lepori et al., 2024; Zhang & Nanda, 2024). We apply this principle to components indexed by brain-supervised visual ROIs, using matched natural-image original–donor pairs, a fixed semantic readout, continuous mixtures, and size-matched random subsets controls.

B.8 General Formulation

The intervention requires corresponding original and donor components and a fixed readout. Its interpretation is limited to the tested representation, component groups, and output metric; it does not establish biological effects or generalization to other representations.