BRAINSEG: A FRAMEWORK FOR PROBING CATEGORY-SPECIFIC NEURAL REPRESENTATIONS THROUGH DENSE SEMANTIC DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

The human visual system transforms sensory inputs into structured representations of semantic content and spatial organization. Recent deep-learning methods can reconstruct visually compelling images from fMRI responses, but it remains unclear whether the semantic content of these reconstructions is supported by measured brain responses or supplied by generative priors. We introduce BrainSeg, a framework that reformulates fMRI-based visual decoding as dense semantic prediction. Rather than reconstructing image appearance, BrainSeg directly predicts which semantic categories are present and where they occur. BrainSeg maps fMRI responses into complementary semantic queries and multiscale spatial features and progressively couples these two representations to produce a dense semantic map over 133 categories. Training targets are generated offline using a frozen MaskFormer, whereas inference relies on fMRI alone. We evaluate BrainSeg on held-out stimuli from the Natural Scenes Dataset using metrics of category presence, spatial localization, and dense segmentation. We further assess whether its predictions depend on stimulus-specific neural signals using shuffled fMRI–stimulus pairings and category-conditioned layout controls. BrainSeg also enables post-hoc category-conditioned attribution, relating individual decoded components to voxel-level and ROI-level relevance patterns. By making "what was where" an explicit decoding target, BrainSeg provides an interpretable framework for probing the organization of category-specific neural representations in human visual cortex.

1 INTRODUCTION

Visual recognition involves identifying scene components and locating them in the visual field. fMRI decoders can reconstruct images through predicted deep visual features (Shen et al., 2019) or diffusion-based generation (Takagi & Nishimoto, 2023; Scotti et al., 2023). The resulting images are informative, yet a plausible object or layout may reflect the generator's learned prior as well as stimulus-specific neural evidence (Shirakawa et al., 2025). This motivates a more direct question: which categories, and which parts of their spatial arrangement, can be decoded from fMRI alone?

Existing outputs resolve different parts of this question. Semantic retrieval and captions capture scene content, while fMRI-based grounding and detection localize selected objects with boxes (Xia et al., 2024; Deng et al., 2025). Boxes do not describe object shapes or the background regions that make up much of a scene. Segmentation-based evaluation of reconstructed RGB images measures dense structure after generation (Xia & Öztireli, 2026), but its category scores cannot directly identify which voxel signals supported the generated details. A decoder whose output is itself a semantic map allows category presence and dense spatial overlap to be evaluated separately, and its category scores provide explicit targets for voxel attribution.

BrainSeg implements this formulation with a frozen MaskFormer teacher (Cheng et al., 2021). Separate fMRI projections initialize semantic queries and spatial tokens; cross-attention couples them during multiscale decoding to produce class-specific masks, and inference uses fMRI alone. We evaluate the resulting maps on an identity-disjoint NSD validation split viewed by four participants. Our contributions are as follows:

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.

- **A direct semantic-layout decoding target.** We reformulate fMRI-based visual decoding as dense prediction of 133-category semantic maps, supervised by a frozen MaskFormer teacher and requiring only fMRI at inference, so that category presence (*what*) and spatial support (*where*) become separately measurable quantities.
- **A controlled empirical analysis.** On identity-disjoint NSD validation stimuli across four participants, we quantify decoding quality against neural decoders, non-neural priors, and shuffled-fMRI controls, and further check agreement with human COCO annotations and cross-subject consistency.
- **Category-conditioned cortical attribution.** The decoded category scores provide differentiable probes for voxel-level attribution; ROI enrichment (scenes in PPA/OPA, persons in EBA/FFA) and occlusion sensitivity are validated with permutation and size-matched random controls.

2 Related Work

2.1 Visual Reconstruction from fMRI

Early image-space decoders recovered contrast patterns or used encoding models with natural-image priors (Miyawaki et al., 2008; Naselaris et al., 2009). Subsequent approaches decoded hierarchical visual features or mapped fMRI into learned generative spaces (Shen et al., 2019; Seeliger et al., 2018; VanRullen & Reddy, 2019). Diffusion systems now combine decoded image latents, low-level features, and CLIP representations to synthesize recognizable natural scenes (Takagi & Nishimoto, 2023; Ozcelik & VanRullen, 2023; Scotti et al., 2023). Cross-subject training improves data efficiency in related reconstruction frameworks (Scotti et al., 2024; Wang et al., 2024). These systems target image or embedding fidelity; their semantic layout remains implicit in the generated RGB output. BrainSeg addresses a different output space: a dense class map directly decoded from the measured response.

2.2 Semantic and Structured Brain Decoding

Distributed cortical patterns support object-category discrimination (Haxby et al., 2001), and decoded visual features can support category recognition beyond the training classes (Horikawa & Kamitani, 2017). Multimodal decoders extend this direction through image–text alignment, retrieval, captioning, and question answering (Du et al., 2023; Ma et al., 2025; Ferrante et al., 2024; Horikawa, 2025; Huang et al., 2025). These outputs identify or describe content without assigning a category to every image location.

Spatially structured decoders bring the target closer to the stimulus image. UMBRAE and MindDETR predict object boxes from fMRI (Xia et al., 2024; Deng et al., 2025); NEURONA represents semantic relations over brain parcels rather than pixel locations (Wang et al., 2026). Y-Net produces segmentation-like maps using both fMRI and the observed RGB image (Chatterjee et al., 2024). BrainSeg instead predicts masks for both *thing* and *stuff* categories from fMRI alone. Its output supports direct evaluation of full-scene overlap and class-conditioned voxel attribution.

2.3 Dense Brain–Vision Alignment and Structured Evaluation

Dense visual representations have served several roles in brain decoding. Segmentation models provide spatially indexed targets for pixel classification (Cheng et al., 2021; Xie et al., 2021); BrainSAIL relates observed-image features to voxels through an image-to-fMRI encoding model (Luo et al., 2025). Brain-Imager uses panoptic latents as an intermediate cue for RGB reconstruction (Meng et al., 2025). BASIC and SEED evaluate semantic structure after reconstructing images (Xia & Öztireli, 2026; Park et al., 2026). BrainSeg uses the frozen segmenter only to define teacher maps. The decoded output is the semantic map itself, so category presence and spatial support can be evaluated without generating RGB images.

Direct semantic prediction also requires careful controls. Frequent categories and stereotyped layouts may agree with a reference without conveying information about a particular stimulus. We therefore compare with class-frequency and spatial priors, a category-conditioned layout template,

and shuffled fMRI–stimulus pairings. These controls probe stimulus specificity rather than generic plausibility. Because the same frozen MaskFormer defines both training targets and held-out references, segmentation scores quantify agreement with that teacher, not accuracy against independent human annotations.

3 METHOD

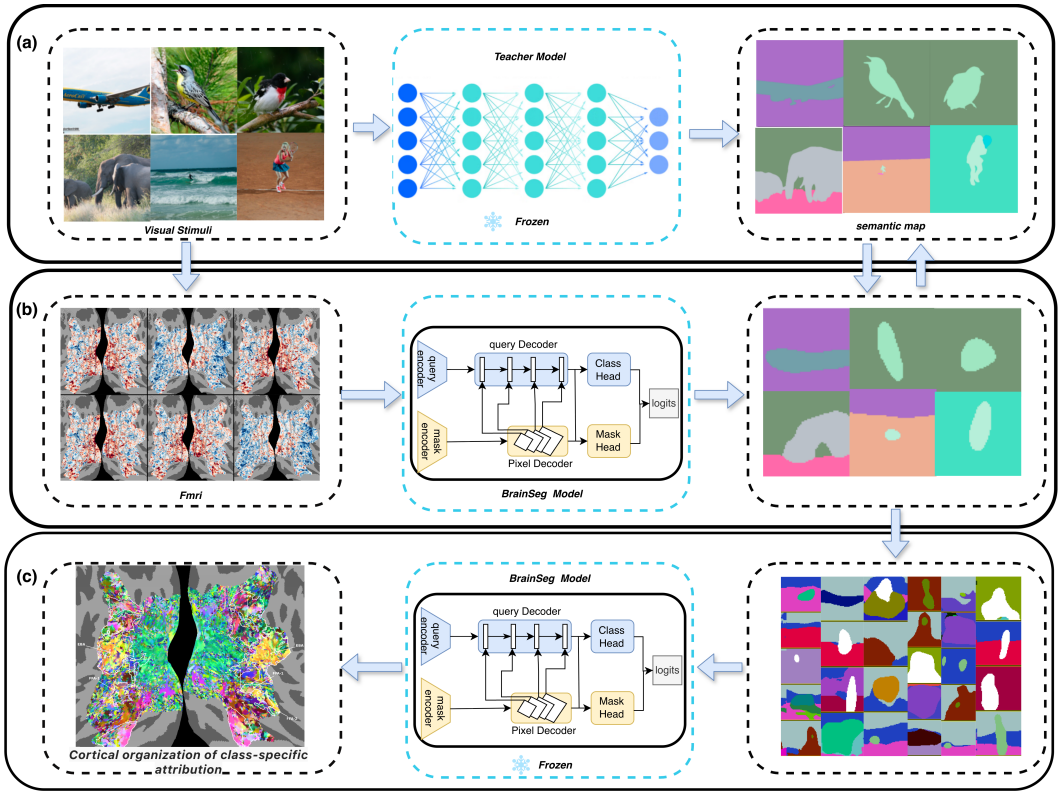

Figure 1: **BrainSeg overview.** (a) A frozen MaskFormer converts paired training stimuli into semantic pseudo-label maps. (b) BrainSeg projects fMRI responses into semantic queries and spatial features; query–feature decoding produces class scores and masks that form a dense semantic map. Training uses the teacher maps, whereas inference uses fMRI alone. (c) Category-conditioned gradients through the frozen trained decoder trace selected output scores back to voxels, yielding cortical attribution maps. Attribution measures the decoder's input dependence, not a causal role in perception.

BrainSeg learns a direct mapping from an fMRI response to a dense semantic layout (Fig. 1). The method has three components. First, a frozen visual teacher model converts each training stimulus into a semantic target in a fixed label space. Second, two fMRI projections produce complementary semantic queries and spatial tokens. Third, a multiscale decoder couples the two representations and predicts a category and spatial mask for every query. After training, the same category scores provide differentiable probes for identifying the voxels associated with each decoded semantic component.

3.1 PROBLEM FORMULATION AND SEMANTIC SUPERVISION

Let $\mathbf{x} \in \mathbb{R}^V$ denote the visually evoked fMRI response and $\mathbf{I}$ its paired stimulus image. Our goal is to learn a decoder F_θ that predicts a dense semantic map

$$F_\theta(\mathbf{x}) = \hat{\mathbf{Y}}, \qquad \hat{\mathbf{Y}} \in \{1, \ldots, C\}^{H \times W}. \tag{1}$$

We use $V = 37{,}984$ visually responsive voxels, $C = 133$ COCO panoptic categories, and $H = W = 64$, consistent with the preprocessing described in Section 4.1. The output represents category identity (*what*) and category-specific spatial support (*where*) in stimulus coordinates.

Dense targets are produced offline by a frozen MaskFormer teacher (Cheng et al., 2021). For each stimulus, the teacher prediction is converted to a semantic map $\mathbf{Y} \in \{1, \dots, C, 255\}^{H \times W}$, where 255 denotes an ignored pixel. The 133-category vocabulary contains 80 thing and 53 stuff classes. We resize the teacher map to 64×64 using nearest-neighbor interpolation to preserve discrete labels. The resulting training examples are pairs $(\mathbf{x}_i, \mathbf{Y}_i)$. The stimulus image and teacher are used only to construct supervision: at inference, BrainSeg receives the fMRI response alone.

3.2 Semantic queries and spatial tokens

A single visually responsive cortex pattern must specify both the semantic components of a scene and their spatial arrangement. BrainSeg therefore projects the same response into two complementary representations:

$$\mathbf{Q}^0 = \mathrm{reshape}(g_q(\mathbf{x})) \in \mathbb{R}^{N_q \times d}, \qquad \mathbf{X}^0 = \mathrm{reshape}(g_s(\mathbf{x})) + \mathbf{E}_{\mathrm{pos}} \in \mathbb{R}^{h_0 w_0 \times d}. \tag{2}$$

Here, $\mathbf{Q}^0$ contains $N_q = 50$ fMRI-derived semantic queries, while $\mathbf{X}^0$ contains a 4×4 grid of spatial tokens ($h_0 = w_0 = 4$). Both branches use two 8192-dimensional hidden layers with GELU activations and produce embeddings of dimension $d = 128$. A learned positional embedding $\mathbf{E}_{\mathrm{pos}}$ supplies the initial grid coordinates. We then contextualize the spatial tokens with an eight-layer Transformer encoder,

$$\mathbf{H}^0 = \mathrm{reshape}_{2\mathrm{D}}\left(\mathrm{TransformerEnc}(\mathbf{X}^0)\right) \in \mathbb{R}^{d \times w_0 \times h_0}. \tag{3}$$

The query branch represents candidate semantic components, whereas the spatial branch provides the substrate on which their locations are decoded. Both are derived from fMRI; no image features enter the decoder.

3.3 Query-guided multiscale decoding

BrainSeg couples semantic and spatial information over four refinement stages. At stage $\ell \in \{0, 1, 2, 3\}$, the queries attend to the current spatial feature map:

$$\mathbf{Q}^{\ell+1} = \mathrm{LN}\left[\mathbf{Q}^{\ell} + \mathrm{MHA}\left(\mathbf{Q}^{\ell}, \mathrm{flat}(\mathbf{H}^{\ell}), \mathrm{flat}(\mathbf{H}^{\ell})\right)\right]. \tag{4}$$

The spatial features are subsequently upsampled by a factor of two and refined by a residual convolutional block,

$$\mathbf{U}^{\ell} = \mathrm{Up}_2(\mathbf{H}^{\ell}), \qquad \mathbf{H}^{\ell+1} = \mathbf{U}^{\ell} + \phi_{\ell}(\mathbf{U}^{\ell}), \tag{5}$$

where ϕ_ℓ consists of a 5×5 convolution, GroupNorm, and SiLU. Cross-attention is thus applied at resolutions 4×4, 8×8, 16×16, and 32×32, after which the final update produces a 64×64 feature map. This coarse-to-fine interaction lets each query gather spatial evidence while preserving a compact semantic representation.

The final query $\mathbf{q}_q$ is mapped to class logits $\mathbf{a}_q \in \mathbb{R}^C$ and a mask embedding $\mathbf{e}_q \in \mathbb{R}^d$. If $\mathbf{h}_p \in \mathbb{R}^d$ denotes the final pixel feature at location p, the query-specific mask logit is

$$m_{q,p} = \mathbf{e}_q^{\top} \mathbf{h}_p. \tag{6}$$

Following query-based segmentation, we combine category and mask predictions into a dense semantic score:

$$S_{c,p} = \sum_{q=1}^{N_q} \mathrm{softmax}(\mathbf{a}_q)_c \cdot \mathrm{sigmoid}(m_{q,p}), \qquad P_{c,p} = \frac{S_{c,p}}{\sum_{c'=1}^{C} S_{c',p} + \epsilon}. \tag{7}$$

The decoded map is $\hat{Y}_p = \arg\max_c P_{c,p}$. This factorization assigns every query both a semantic identity and a spatial support, making category presence and localization separately measurable.

3.4 TRAINING OBJECTIVE

Pixelwise cross-entropy treats the semantic scores $\mathbf{S}$ as logits:

$$\widetilde{P}_{c,p} = \frac{\exp(S_{c,p})}{\sum_{c'=1}^{C} \exp(S_{c',p})}, \qquad \mathcal{L}_{\mathrm{CE}} = -\frac{1}{|\Omega|} \sum_{p \in \Omega} \log \widetilde{P}_{Y_p,p}, \qquad \Omega = \{p : Y_p \neq 255\}, \quad (8)$$

Class weights are uniform. The Dice term and inference instead use the sum-normalized $P_{c,p}$ defined above. Two auxiliary objectives act on a *foreground subset* of 109 categories: the 80 *thing* classes and the 29 unmerged *stuff* classes of the contiguous COCO-panoptic vocabulary. The 24 remaining *stuff* classes (the LVIS-"merged" labels such as sky, tree, and grass, which occupy large but low-precision regions) are excluded from the auxiliary objectives and retain supervision only through $\mathcal{L}_{\mathrm{CE}}$.

The auxiliary Dice term operates on the per-pixel renormalized probabilities $P_{c,p}$, computes a soft Dice coefficient per foreground class that is present in the target, and averages $1 - \mathrm{Dice}$ over those classes. The prototype objective encourages pixels sharing a foreground label to have similar query-mask activation profiles. Writing $\mathbf{u}_p \in \mathbb{R}^{N_q}$ for the ℓ_2-normalized mask-logit profile at pixel p, each foreground region of the target image with at least 8 pixels contributes a prototype

$$\boldsymbol{\mu}_c = \mathrm{norm}\left(\frac{1}{|\Omega_c|} \sum_{p \in \Omega_c} \mathbf{u}_p \right), \qquad \Omega_c = \{p : Y_p = c\}, \quad (9)$$

and every pixel of a qualifying region is then classified among the image's prototypes by a cross-entropy over cosine logits scaled by a temperature of 0.1. The full objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + 2.0\, \mathcal{L}_{\mathrm{Dice}} + 0.1\, \mathcal{L}_{\mathrm{proto}}. \quad (10)$$

Section 4.5 reports that the prototype term has negligible effect on final scores; it is retained in all main models for consistency with the trained checkpoints.

3.5 CATEGORY-CONDITIONED CORTICAL ATTRIBUTION

After training, each category's decoded score provides a differentiable target for voxel attribution. We compute gradient-times-input relevance with respect to the participant's fMRI response and aggregate absolute relevance over held-out images containing the category. For each independently defined functional ROI, enrichment is the mean relevance in that ROI divided by mean relevance over the visually responsive cortex. Values above one indicate that the trained decoder draws relatively more on the ROI for that category. We also set high-relevance input voxels to zero and measure the change in category IoU, comparing the effect with size-matched random-voxel controls. These analyses characterize dependencies of the trained decoder; input occlusion does not establish a causal role for the corresponding cortex in human perception.

4 RESULTS

4.1 EXPERIMENTAL SETUP

Data. We use the Natural Scenes Dataset (NSD) (Allen et al., 2022) and its four 40-session participants (S1, S2, S5, S7). Preprocessed single-run betas (`betas_fithrf_GLMdenoise_RR`) are mapped to fsaverage and restricted to the NSD general visual mask, giving V=37,984 voxels per participant in left-then-right hemisphere order. We average repeated responses and z-score with training-set statistics. The 1,000 shared COCO images (Lin et al., 2014) form the validation split, disjoint by image identity from each participant's remaining $\sim$9,000 training images. Validation pairs never enter gradient updates but drive learning-rate scheduling and checkpoint selection. Reported shared-image scores are validation results, not independent test results.

Teacher and training. A frozen COCO-panoptic MaskFormer (Swin-Base) (Cheng et al., 2021) processes paired stimuli at shortest-edge resolution 896. Its 133-class maps are resized to 64×64 by nearest-neighbor interpolation; unlabeled pixels use ignore index 255. Training images supply

Table 1: **Dense semantic decoding of NSD validation images.** Presence metrics under per-image and macro conventions, and per-image mIoU against the frozen MaskFormer teacher.

	Per-image presence				
Subj.	P	R	F1	Macro F1	mIoU
S1	.7165	.4542	.5559	.2792	.1657
S2	.7013	.4227	.5275	.2533	.1550
S5	.7163	.4575	.5584	.2950	.1545
S7	.7011	.4187	.5243	.2448	.1460
Mean	.7088	.4383	.5415	.2681	.1553

gradient targets; validation teacher maps supply selection metrics. Inference uses fMRI alone. Per-participant BrainSeg models train for 150 epochs with Adam (AMSGrad), learning rate 10^{-4}, batch size 64, and ReduceLROnPlateau on validation loss (factor 0.5, patience 10), using DistributedDataParallel; the best checkpoint uses validation mIoU. Linear and MLP baselines train for 25 epochs (AdamW, learning rate 10^{-3} with cosine decay, batch size 64) with best-validation checkpoint selection. Both plateau by epoch 21 (Appendix D), so the BrainSeg–baseline gap does not reflect under-trained baselines. The shuffled control follows the BrainSeg schedule.

Metrics. Category presence is scored by per-image precision, recall, and F1 (averaged over stimuli) and by their macro counterparts (averaged over the 133 classes). A category counts as present in the reference if it occupies at least one valid (non-ignored) pixel, and as predicted-present if it occupies at least one predicted pixel; per-image statistics average these image-level counts over stimuli, and macro statistics accumulate per-class counts over the held-out set before averaging over classes. Per-image mIoU averages IoU over categories present in either the prediction or reference, then over images. All methods ignore index-255 pixels and share the same post-processing.

Baselines and controls. Non-neural controls use training-set class frequencies, per-pixel majority labels, or class-conditioned spatial templates supplied with the true validation-image category set (oracle). Neural baselines are fMRI kNN retrieval (cosine similarity over training responses; per-pixel one-hot voting over the k retrieved teacher maps, with $k \in \{1, 5, 25\}$), a linear probe ($133\times16\times16$ logits, bilinearly upsampled), and a two-hidden-layer MLP (4,096 units; $133\times32\times32$ logits, bilinearly upsampled). The shuffled-fMRI control trains BrainSeg with within-batch fMRI–stimulus pairing permutations. All controls use the same held-out references and post-processing.

4.2 Dense semantic decoding on validation stimuli

Table 1 reports decoding quality on the shared validation images. BrainSeg identifies which categories are present with per-image presence F1 of .5243–.5584 (mean .5415) and macro F1 of .2681, and produces dense maps with per-image mIoU of .1460–.1657 (mean .1553) against the teacher. The two conventions bracket complementary failure modes: per-image scores are dominated by the frequent categories that populate typical scenes, whereas the macro scores weight all 133 categories equally and are correspondingly lower—decoding quality is strongly modulated by category frequency and area, and rare, small categories are recovered least reliably. Performance is consistent across the four participants (spread of 5.0 macro-F1 and 3.4 per-image-F1 points), indicating that dense semantic decoding does not depend on idiosyncrasies of a single cortical anatomy.

Agreement with human annotations. Because the teacher defines both training targets and the references above, we additionally evaluate against the human COCO panoptic annotations of the same held-out stimuli (center-cropped and resized identically to the NSD stimuli). The teacher reaches mIoU .5538 against human labels; BrainSeg reaches .1506–.1688 (mean .1593) —matching its teacher-agreement within a point—and the shuffled control remains at the prior level (S7: .0421). The pseudo-label supervision therefore does not inflate the reported decoding: agreement transfers unchanged from teacher-defined to human-defined layouts, at 28.8% of the teacher's own ceiling.

Qualitative results. Figure 2 shows held-out stimuli, teacher references, and the predictions of the four participants' independently trained models. Predictions recover the dominant thing and

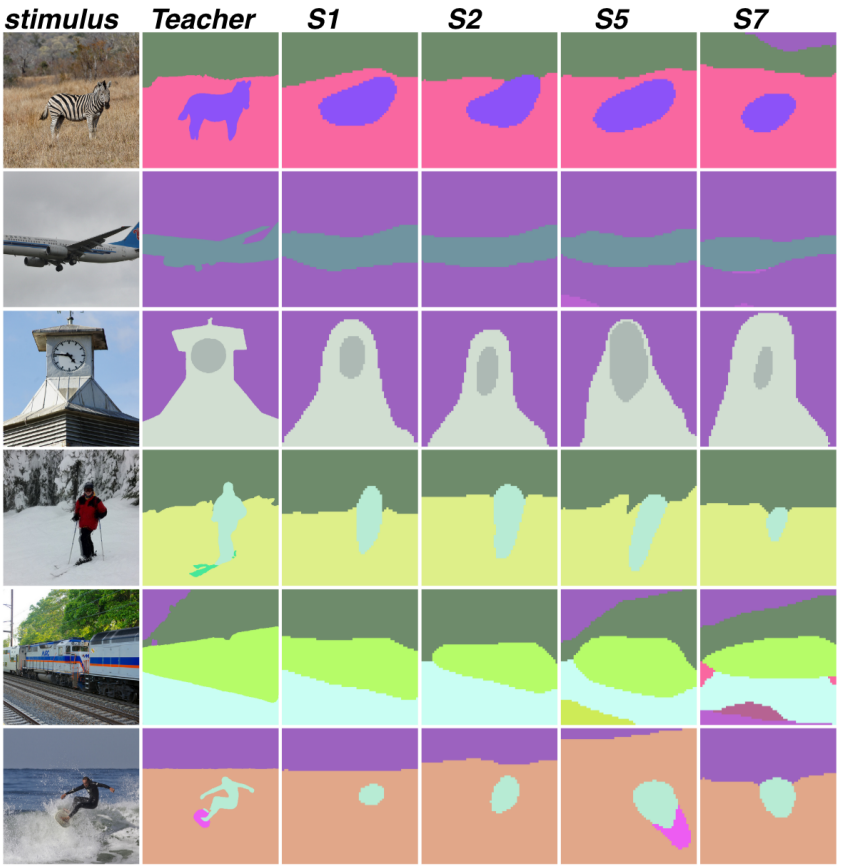

Figure 2: **Qualitative decoding across subjects.** Each row shows the stimulus, the MaskFormer teacher segmentation, and predictions from BrainSeg trained independently for each subject (consistent class colors).

stuff regions and their coarse spatial arrangement—grass, sky, water, snow, and building facades are laid out correctly, and persons are localized to their coarse position—while boundaries are visibly limited by the 64×64 output resolution. The four columns agree substantially despite fully separate models, and typical failures involve small objects, thin structures, and semantically close categories.

4.3 Controls and Baselines

Table 2 isolates the contribution of the measured responses. On S7, the strongest non-neural controls reach macro F1 .0582 (class-frequency prior) and mIoU .0410 (pixelwise spatial prior). The class-conditioned layout template, although given the true validation-image category set, reaches mIoU .0055: averaged masks overlap the layout of any individual image very little, quantifying how little fixed dataset-level statistics explain stimulus-specific arrangement. The standard decoding baselines show where the difficulty lies: kNN retrieval and the linear probe remain close to the prior level on mIoU (.0575 and .0449), and the MLP recovers much of the category-presence signal (per-image F1 .4970) but only .1049 mIoU—category existence is comparatively easy, whereas spatially resolved segmentation is where the architecture matters. BrainSeg improves over the MLP by 4.7 macro-F1 and 5.0 mIoU points on average (mIoU $+48\%$ relative). This comparison measures the gain of the full architecture over the MLP baseline.

Retraining *all four participants* with permuted fMRI–stimulus pairings collapses performance to mIoU .0389–.0413 and macro F1 .0170–.0210 (four-participant mean .0403 and .0182)—indistinguishable from the prior level and nearly $4\times$ below the real models on mIoU. Notably, the four shuffled models converge to nearly the same solution (per-participant mIoU within .0025 of each other, range .0389–.0413): without genuine fMRI–stimulus correspondence, the best this ar-

Table 2: **Baselines and controls.** Standard decoding baselines and the shuffled-fMRI control use identical data, teacher references, and evaluation as BrainSeg and report the four-participant mean; non-neural priors are evaluated on the same held-out images (S7). Template† uses the true validation-image category set as an oracle input to the tested template procedure.

Method	Macro F1	Per-image F1	mIoU
Class-frequency prior (S7)	.0582	–	.0076
Pixelwise spatial prior (S7)	.0211	–	.0410
Layout template† (S7)	.0135	–	.0055
kNN retrieval (k=25)	.1646	.4231	.0575
Linear probe	.1815	.3573	.0449
MLP	.2214	.4970	.1049
Shuffled-fMRI BrainSeg	.0182	.3158	.0403
BrainSeg (ours)	**.2681**	**.5415**	**.1553**

chitecture can learn is the dataset-level layout prior, reached independently from four different sets of shuffled inputs.

Cross-subject consistency. Across all six participant pairs, predictions agree more for the same stimulus than for different stimuli (mean pairwise IoU .2354, 95% CI [.2293, .2414] versus .0292 [.0285, .0298]). The shuffled controls provide the prior-agreement floor for this comparison: two shuffled models also agree strongly (.6923) but *identically* for same- and different-stimulus pairs (.6923 versus .6919, gap $<$.001), and real–shuffled pairs behave the same way (.0699 versus .0692). The stimulus-specific component of cross-subject agreement is therefore the same-minus-different gap (.2062), not the absolute overlap, which includes a stimulus-blind prior-driven part. This supports shared stimulus information in independently trained decoders without identifying the source of their residual disagreement.

4.4 Category-specific cortical support

We aggregate gradient-times-input attribution over six category families (person, animal, vehicle, food, indoor/furniture, and scene/stuff) and six independently defined fLOC ROIs (PPA, OPA, FFA, OFA, EBA, FBA) (Fig. 3a).

Relevance concentrates in category-selective cortex. Scene-category relevance is enriched in PPA ($1.6818\pm0.0985\times$ the cortex mean, permutation p=.001 against 1,000 size-matched random masks) and OPA (1.1965 ± 0.1422, p=.001); person-category relevance is enriched in EBA (1.4423 ± 0.1385) and FFA (1.3790 ± 0.3482). These associations appear in all four participants and persist in direction under an alternative group-probability ROI definition (Appendix B; 4/4 participants for PPA, EBA, and FFA, 3/4 for OPA). Non-matched pairs are close to unity (PPA$\times$person 1.0822, FFA$\times$scene 0.9982, OPA$\times$vehicle 0.9012), ruling out a global account in which every region enriches for every family. OFA does not show person-family enrichment (0.8628), whereas FFA does (1.3790). Enrichment alone cannot separate semantics from correlated visual features, which motivates the occlusion test below. The enrichment analysis is repeated with Integrated Gradients and SmoothGrad in Appendix B; the associations are robust to the attribution method.

Decoder sensitivity to high-attribution voxels. We compare occlusion of attribution-ranked voxels with 100 size-matched random controls (Fig. 3b,c). On the full held-out set, removing the top 2% reduces *playingfield* IoU by .0635–.2524 and *person* IoU by .0162–.0482 (all p=.0099), whereas random-voxel effects remain negligible through 50% occlusion. The top 10% of scene-attribution voxels accounts for 43.9–77.4% of the total damage from full occlusion; person damage rises through 75%, suggesting more distributed support. Within EBA, removing the top person-attribution voxels impairs person decoding in all four participants, unlike equally ranked voxels outside EBA. The top three attribution-ranked fLOC ROIs account for 51.1–75.3% (person) and 91.5–100% (playingfield) of the damage from occluding all six. Split-half attribution correlations are .6457–.6829 within categories versus .2870–.3158 across categories. The S1 cross-category exception and full controls appear in Appendix C.

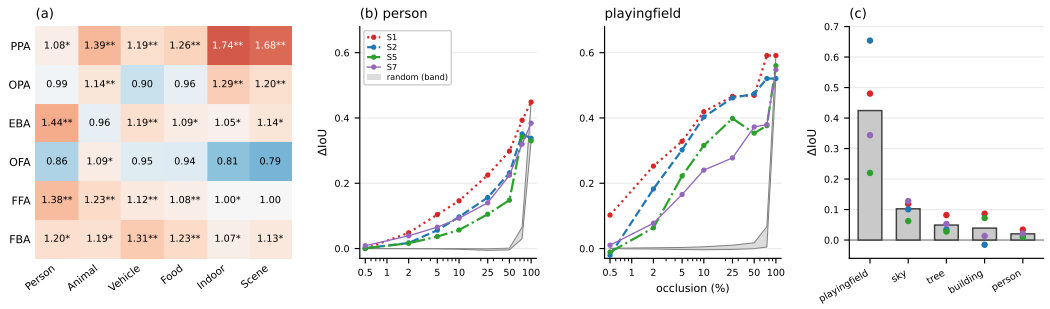

Figure 3: **Neural evidence for the decoded semantic code.** (a) Attribution enrichment (ROI mean / cortex mean) on official per-subject fLOC parcellations (mean over S1/S2/S5/S7; **: replicated 4/4 subjects, permutation $p<.05$; *: $p<.05$ in at least one participant). (b) Occlusion dose–response on the full held-out set: category IoU drop (ΔIoU, y-axis) versus occluded-voxel percentage (log x-axis) for *person* and *playingfield* (S1 dotted, S2 dashed, S5 dash-dot, S7 thin solid; shaded band: min–max of the random-voxel control, which coincides with the target arms at 100% by construction). (c) Change in decoder IoU (ΔIoU at 2% occlusion) per category; bars are four-subject means, dots are individual subjects.

4.5 Component ablations

On S7, replacing the query–mask head with per-pixel classification over the same spatial tokens reduces mIoU from .1460 to .1120 and macro F1 from .2448 to .1621; the same ablation on S1 replicates the ordering (.1657→.1256, .2792→.1921), an order of magnitude above seed variability ($\leq$.0023 across three S7 seeds). The prototype auxiliary loss has negligible effect in both participants. Full component results appear in Appendix A.

5 Discussion

BrainSeg decodes semantic layout directly from fMRI, rather than a by-product of RGB reconstruction, which may also reflect learned image priors (Shirakawa et al., 2025). Its four-participant mIoU (.1553 versus .1049 for the MLP and .0403 for the shuffled control) and the same-stimulus agreement gap support stimulus-dependent recovery beyond these baselines, and decoded scores provide targets for tracing the model's reliance to measured voxels. Attribution is enriched for person categories in EBA and FFA and scene categories in PPA and OPA, with occlusion and split-half analyses supporting stable, category-dependent reliance: dependence on recorded voxels, not causal necessity for perception, possibly reflecting correlated visual features.

These are validation scores, not independent-test estimates, bounded by MaskFormer pseudo-labels despite similar agreement with human COCO annotations. Averaged-repetition predictions at 64×64, four separately trained models, rare or small categories, one layout-template prior, and the fixed COCO vocabulary bound the claims, leaving single-trial responses, fine boundaries, and cross-subject transfer untested. Within these limits, BrainSeg links decoded category presence and coarse spatial support to its cortical inputs.

AI Use Statement

Generative AI tools assisted with refining the research framing, discussing evaluation controls, interpreting preliminary analyses, drafting and editing manuscript text, and preparing schematic figures. All AI-assisted text was reviewed and revised by the authors, and every quantitative claim in the paper was verified by the authors against the experiment outputs generated by the authors' own analysis code. The authors are responsible for the reported results, citations, and final claims, including any content produced with the aid of generative AI.

Reproducibility Statement

All experiments use the publicly released Natural Scenes Dataset (Allen et al., 2022) and its official preprocessing stream; no new human data were collected. Section 4.1 specifies the voxel selection, trial averaging, splits, and training protocol; Section 3.4 gives the complete loss formulation and weights; Sections 3.2 and 3.3 specify the architecture. Appendix B documents ROI construction and the permutation test, and Appendix C documents the occlusion protocol. Training, evaluation, and analysis code together with the trained per-participant checkpoints will be released upon publication.

Ethics Statement

This study uses the Natural Scenes Dataset, a publicly released fMRI resource collected by its own investigators under the consent and ethics approval of that project; the present authors had no contact with participants and collected no new human-subject data. Decoding analyses operate on group-level, de-identified, publicly shared measurements and do not attempt to identify individuals. The decoded outputs are coarse semantic maps of laboratory stimuli and carry no sensitive personal information; we note, however, that neural decoding research is broadly relevant to mental-privacy questions, and we support ongoing discussion of appropriate consent and safeguards for future applications.

References

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

Subhrasankar Chatterjee, Shreyansh Vansh Verma, Debasis Samanta, and Monalisa Sarma. Y-Net: Fusing fMRI and natural scenes to generate neuro-interpretable heat maps. In *Proceedings of the British Machine Vision Conference Workshops*, 2024. DIFA: Deep Learning-based Image Fusion and Its Applications.

Bowen Cheng, Alexander G Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17864–17875, 2021.

Han Deng, Xinzhu Ma, Yuan Meng, Chen Tang, Chunfeng Song, Zhi Wang, Wenwu Zhu, Xiangyu Yue, and Wanli Ouyang. MindDETR: Beyond semantics, exploring positional cues from brain activity, 2025. OpenReview preprint, ICLR 2025 submission #6071.

Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10760–10777, 2023.

Matteo Ferrante, Tommaso Boccato, Furkan Ozcelik, Rufin VanRullen, and Nicola Toschi. Multimodal decoding of human brain activity into images and text. In *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, volume 243 of *Proceedings of Machine Learning Research*, pp. 87–101. PMLR, 2024.

James V. Haxby, M. Ida Gobbini, Maura L. Furey, Alumit Ishai, Jennifer L. Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

Tomoyasu Horikawa. Mind captioning: Evolving descriptive text of mental content from human brain activity. *Science Advances*, 11(45):eadw1464, 2025.

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

Wanqiu Huang, Ke Ma, Tingyu Xie, and Hongwei Wang. BrainChat: Interactive semantic information decoding from fMRI using large-scale vision-language pretrained models. In *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5. IEEE, 2025.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Andrew F. Luo, Jacob Yeung, Rushikesh Zawar, Shaurya Dewan, Margaret M. Henderson, Leila Wehbe, and Michael J. Tarr. Brain mapping with dense features: Grounding cortical semantic selectivity in natural images with vision transformers. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

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.

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.

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

Juhyeon Park, Peter Yongho Kim, Jiook Cha, Shinjae Yoo, and Taesup Moon. SEED: Towards more accurate semantic evaluation for visual brain decoding. In *The Fourteenth International Conference on Learning Representations*, 2026.

Paul S Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J Dempster, Nathalie White Verlinde, Elad Yundler, David Weisberg, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. In *Advances in Neural Information Processing Systems*, volume 36, pp. 16112–16124, 2023.

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. PMLR, 2024.

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

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

Ken Shirakawa, Yoshihiro Nagano, Misato Tanaka, Shuntaro C Aoki, Yusuke Muraki, Kei Majima, and Yukiyasu Kamitani. Spurious reconstruction from brain activity. *Neural Networks*, 190: 107515, 2025.

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.

Rufin VanRullen and Leila Reddy. Reconstructing faces from fMRI patterns using deep generative neural networks. *Communications Biology*, 2:193, 2019.

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

Yanchen Wang, Joy Hsu, Ehsan Adeli, and Jiajun Wu. Neuro-symbolic decoding of neural activity. In *The Fourteenth International Conference on Learning Representations*, 2026.

Weihao Xia and Cengiz Öztireli. Multigranular evaluation for brain visual decoding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(21):17868–17876, 2026.

Weihao Xia, Raoul de Charette, Cengiz Öztireli, and Jing-Hao Xue. Umbrae: Unified multimodal brain decoding. In *European Conference on Computer Vision*, pp. 237–254. Springer, 2024.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12077–12090, 2021.

A COMPONENT ABLATION DETAILS

Table 3: **Component ablations** (S7 and S1, identical 150-epoch protocol and evaluation) and S7 seed variability.

Variant	Macro F1	Per-image F1	mIoU
S7			
BrainSeg (full, seed 0)	.2448	.5243	.1460
seed 1	.2436	.5245	.1444
seed 2	.2373	.5189	.1437
− query–mask head (per-pixel)	.1621	.4777	.1120
MLP+conv decoder	.1815	.4961	.1186
S1			
BrainSeg (full)	.2792	.5559	.1657
− query–mask head (per-pixel)	.1921	.5066	.1256
MLP+conv decoder	.2146	.5294	.1359

The query–mask ablation replaces query-conditioned mask prediction with per-pixel classification over the same spatial tokens. The MLP-with-convolutions variant provides a second, lighter decoder comparison.

B ROI ENRICHMENT ANALYSIS DETAILS

ROI definitions. The primary masks use the official NSD fLOC parcellations of each participant (faces, places, and bodies localizer suites), defined in the participant's own FreeSurfer surface. Because decoding operates in fsaverage space, we map each parcellation to fsaverage by nearest-neighbor matching on the spherical registration coordinates (`sphere.reg`). The three localizer suites are thresholded independently and can overlap; a voxel claimed by several ROIs is assigned to the suite whose localizer contrast has the higher t-value at that voxel, yielding six mutually exclusive masks (PPA, OPA, FFA, OFA, EBA, FBA). FFA-1 and FFA-2 are merged (FFA-2 contains as few as 56 voxels in S7), and FBA-1/FBA-2 are merged into FBA. As a robustness check, all analyses are repeated with an alternative mask definition based on the fsaverage group probability maps (top-800 voxels per ROI).

Robustness to mask definition. Table 5 reports the four key associations under both definitions. Direction and significance are preserved under the probability-map definition: person-family effect sizes are essentially unchanged (EBA $1.4423 \rightarrow 1.4854$, FFA $1.3790 \rightarrow 1.2937$), while the scene-family cells attenuate (PPA $1.6818 \rightarrow 1.5010$, OPA $1.1965 \rightarrow 1.0594$; part of this shift reflects the family partition—the Scene cells in the probability-map column use the semantic partition of Table 4, under which the official-mask values are PPA 1.4692 and OPA 1.0905). PPA, EBA, and FFA replicate in 4/4 participants under both definitions; OPA replicates in 4/4 under the official masks and in 3/4 under the probability maps (S1: 0.9867).

Robustness to the attribution method. Table 4 repeats the enrichment analysis with two alternative attribution methods—Integrated Gradients (32 steps) and SmoothGrad (20 samples, noise 0.1)—computed from the same checkpoints on the same held-out images under one fixed semantic family partition. (Person, Animal, Vehicle, and Food cells coincide with the official partition; the Scene and Indoor cells differ slightly in class composition, so their absolute values are not comparable to Table 5—the comparison of interest is across columns.) All four headline associations persist under Integrated Gradients, which if anything sharpens the person-family enrichment (EBA $1.4423 \rightarrow 1.7007$, FFA $1.3790 \rightarrow 1.9217$), and remain at or above unity under SmoothGrad, which attenuates magnitudes toward one. Across the full ROI$\times$family matrix, Integrated Gradients agrees with the gradient-times-input direction in 35 of 36 cells, and the per-category voxelwise relevance profiles correlate with the gradient-times-input profiles at r=.7559–.7879 (SmoothGrad: r=.5275–.5353).

Table 4: **Attribution-method robustness of the enrichment analysis** (mean±sd over participants; all cells replicate in 4/4 participants). Methods: gradient-times-input (g×i), Integrated Gradients (IG), SmoothGrad (SG).

Pair	g×i	IG	SG
PPA×scene	1.4692±.0550	1.5123±.0617	1.2819±.0540
OPA×scene	1.0905±.0625	1.1252±.0602	1.0308±.0508
EBA×person	1.4423±.1385	1.7007±.2208	1.0330±.0333
FFA×person	1.3790±.3482	1.9217±.6486	1.1658±.0843
PPA×person	1.0822±.1168	1.1183±.1401	1.0588±.0436
FFA×scene	0.9875±.0525	0.9645±.0782	1.0099±.0308
OPA×vehicle	0.9012±.0550	0.8885±.0639	1.0086±.0416

Table 5: **Attribution enrichment under two independent ROI definitions** (mean±sd over participants; official masks carry permutation p from 1,000 size-matched random masks). Probability-map scene cells use the semantic family partition of Table 4 (person-family cells coincide with the official partition). Replication: 4/4 participants for all official-mask cells and for PPA/EBA/FFA under probability maps; OPA 3/4 under probability maps (S1: 0.99).

Pair	Official fLOC masks	Probability maps
PPA×scene	1.6818±.0985 (p=.001)	1.5010±.1017
EBA×person	1.4423±.1385 (p=.001)	1.4854±.1439
FFA×person	1.3790±.3482 ($p\leq$.002)	1.2937±.1369
OPA×scene	1.1965±.1422 (p=.001)	1.0594±.0489

Additional patterns in the enrichment matrix. The complete ROI×family matrix appears in Fig. 3a. PPA enrichment for indoor/furniture categories (1.7397±.1421, replicated in 4/4 participants) is at least as strong as for scene/stuff. The two remaining face- and body-selective ROIs dissociate from the headline pairs: OFA shows no person-family enrichment (0.8628, four-participant mean), and FBA is weaker and less consistent than EBA (mean 1.1990±.1601; 0.945 in S5, the only non-replicating cell among the person-family ROIs), whereas FFA replicates in 4/4 participants (1.106–1.964, all permutation $p\leq$.002).

C Occlusion analysis details

C.1 Voxel-wise occlusion over the full held-out set

Table 6 lists the per-participant IoU changes at the 2% occlusion ratio for the five tested categories (100 held-out images per category). The *tree* effect is significant against random controls but not specific: occluding another category's top voxels impairs tree to the same degree, indicating shared rather than exclusive support. Because the 100-image subsample and the full-set evaluation below use different image draws, their effect sizes differ in both directions (playingfield effects are larger in the subsample, person effects smaller); the random-voxel null and the specificity contrasts agree in the two evaluations, and we treat the full-set values as definitive.

Table 6: **Change in category IoU after zeroing the top 2% of attribution voxels** (760 voxels; 100 held-out images per category). Positive values denote an IoU drop. Random size-matched controls change IoU by less than .0030 in every cell. *Significant against random controls but not category-specific: occluding another category's top voxels impairs tree to the same degree (see text).

ΔIoU@2%	S1	S2	S5	S7
playingfield	+.4802	+.6541	+.2200	+.3438
sky	+.1187	+.1009	+.0624	+.1279
tree*	+.0820	+.0342	+.0279	+.0527
building	+.0864	−.0154	+.0722	+.0133
person	+.0343	+.0163	+.0092	+.0213

Evaluated on the *full* held-out set (all images containing the category) for the two headline categories (Table 7), the person effect is substantially larger than in the 100-image sample, confirming that the small person effect is a property of distributed—not weak—coding. Extending the occlusion ratio to 100% of voxels shows that scene damage saturates by 10–25% occlusion (top-10% attribution voxels account for 44–77% of the total achievable damage), whereas person damage rises monotonically through 75%. At 100% the target and random arms coincide by construction (identical operation), a degeneracy check the data satisfy.

Table 7: **Full held-out set, extended ratios.** IoU drop when zeroing the top-k attribution voxels of each category. Random controls are ≈0 through 50%.

ΔIoU	person			playingfield		
ratio	2%	10%	50%	2%	10%	50%
S1	+.0482	+.1462	+.2985	+.2524	+.4192	+.4692
S2	+.0171	+.0968	+.2313	+.1826	+.4030	+.4741
S5	+.0162	+.0568	+.1482	+.0635	+.3156	+.3535
S7	+.0390	+.0929	+.2244	+.0776	+.2402	+.3723

C.2 ROI-RESTRICTED OCCLUSION

Table 8 restricts the top-voxel selection to inside or outside a target ROI. For person, occluding top voxels *inside* EBA impairs decoding more than occluding equally ranked voxels *outside* EBA in all four participants (e.g., +.0294 versus −.0006 in S1), and the outside effects remain at or below .0085. Playingfield shows the opposite pattern in S1/S2 (larger damage from outside PPA), indicating that scene support, though strongest in PPA by attribution, extends across a broader cortical extent—consistent with the distributed scene coding observed at high occlusion ratios.

Table 8: **ROI-restricted occlusion** (ΔIoU at 2%, 760 voxels): top attribution voxels selected inside vs. outside the ROI.

	inside ROI				outside ROI			
	S1	S2	S5	S7	S1	S2	S5	S7
person / EBA	+.0294	+.0134	+.0081	+.0162	−.0006	+.0085	+.0036	+.0023
playf. / PPA	+.1977	+.0820	+.1394	+.2057	+.3828	+.3437	+.0214	+.0640

C.3 CUMULATIVE ROI OCCLUSION

We order the six fLOC ROIs by their mean attribution for the target category and occlude them cumulatively from the highest-ranked (Table 9). Three findings: (i) the attribution-ranked top three regions, selected without any anatomological input, contain EBA for person in 4/4 participants (EBA ranked first in 3/4) and both PPA and EBA for scene in 4/4 (PPA first in 3/4), with the remaining member varying across participants (person: FBA/FFA/PPA; scene: OPA/FBA/FFA); (ii) they cause 51.1–75.3% (person) and 91.5–100% (scene) of the damage of all six regions, well above random region ordering at the same rank (e.g., person at rank 3: .1325–.1901 vs. .0657–.1099); (iii) the six ROIs ($\sim$40% of cortex) impair decoding 2.92–7.95× more than occluding the complementary 60% of cortex (person 2.92–3.36×, scene 2.92–7.95×).

C.4 SPLIT-HALF ATTRIBUTION STABILITY

Splitting the held-out images by even/odd stimulus identity and recomputing per-category attribution (|grad×input|) in each half, within-category correlation across halves averages r=.6457–.6829 (70 categories per participant; person r=.9212–.9493, scene categories r=.7311–.9826), versus r=.2870–.3158 between different categories. The attribution maps are therefore stable and category-specific rather than noise.

Table 9: **Cumulative ROI occlusion** (ΔIoU): attribution-ranked regions added one at a time, vs. the mean of 10 random region orderings, and vs. occluding the complement (non-ROI cortex, $\sim$22k voxels).

	k=1	k=2	k=3	k=6	rand. k=3	complement
person S1	+.0511	+.0920	+.1592	+.2694	+.0685	+.0923
person S2	+.0549	+.1212	+.1325	+.2591	+.0657	+.0770
person S5	+.0368	+.1116	+.1901	+.2874	+.1099	+.0910
person S7	+.0030	+.1323	+.1575	+.2091	+.0779	+.0637
scene S1	+.4890	+.5994	+.6325	+.6325	+.3560	+.0796
scene S2	+.2035	+.5869	+.5997	+.5997	+.2997	+.1499
scene S5	+.3766	+.3018	+.7862	+.7862	+.4248	+.2691
scene S7	+.2291	+.4696	+.4529	+.4948	+.3197	+.1490

C.5 The S1 specificity exception

In S1, occluding the control category's top voxels also impairs playingfield (control-arm drop .4341 vs. target .4802). Two observations localize the cause. First, the person and playingfield top-760 voxel sets overlap in only 97 voxels, so the two arms lesion nearly disjoint tissue. Second, random voxels remain harmless throughout. The pattern therefore reflects a globally heightened sensitivity of the S1 model to removing *any* high-attribution voxels—the strongest-decoder participant relies on a shared high-gain input substrate—rather than mislocalized category attribution.

Dose–response curves. Figure 3b shows the full dose–response for both fully-tracked categories on the held-out set, from 0.5% to 100% occlusion (log axis): playingfield saturates between 10% and 25% in three of four participants, person rises monotonically through 75%, and the random-voxel band stays below .02 through 50% occlusion, coinciding with the target arms at 100% by construction. Figure 4 shows the per-participant panels for all five tested categories at the low occlusion ratios $\{0.5\%, 2\%, 5\%\}$.

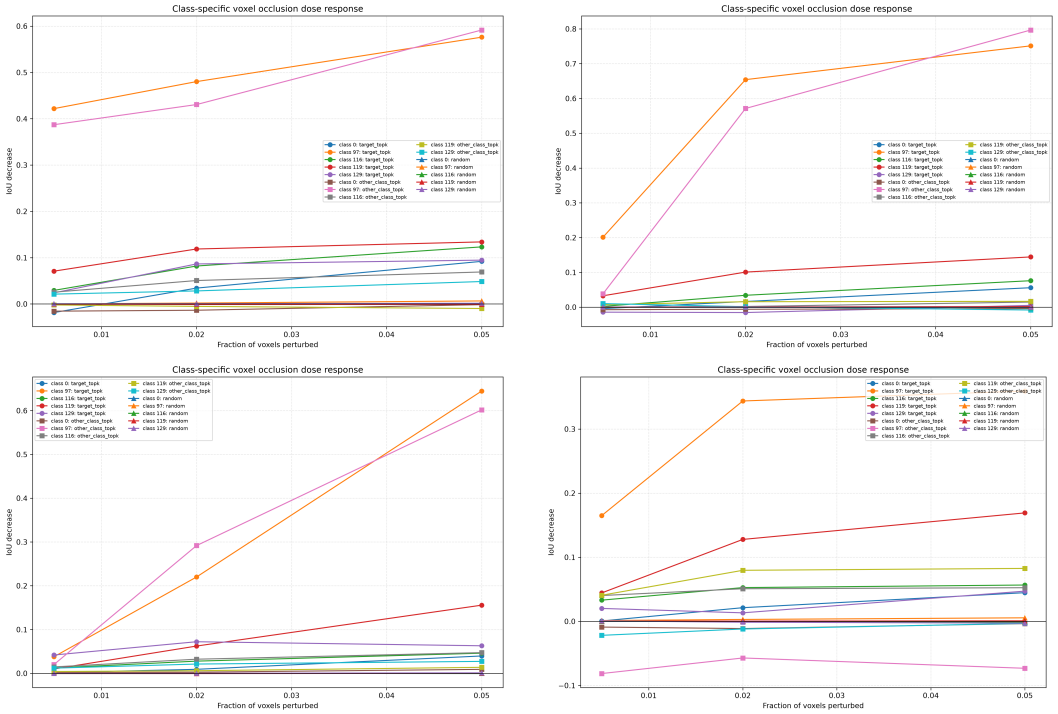

Figure 4: **Occlusion dose–response per participant** (S1, S2, S5, S7). Each panel shows category IoU as a function of the fraction of occluded voxels for the target-attribution arm and the random-voxel control arm.

D BASELINE CONVERGENCE

Figure 5 shows per-epoch validation metrics for the linear and MLP baselines of all four participants. Validation mIoU and macro F1 plateau well before the 25-epoch budget: the best validation epoch falls between 16 and 21 for every subject–model pair, and the mean of the final five epochs differs from the final value by at most .0003 mIoU and .0014 macro F1. The best-versus-final gap is at most .0017 mIoU. The 4.7–5.0-point BrainSeg advantage over the MLP is therefore not an artifact of stopping the baselines early.

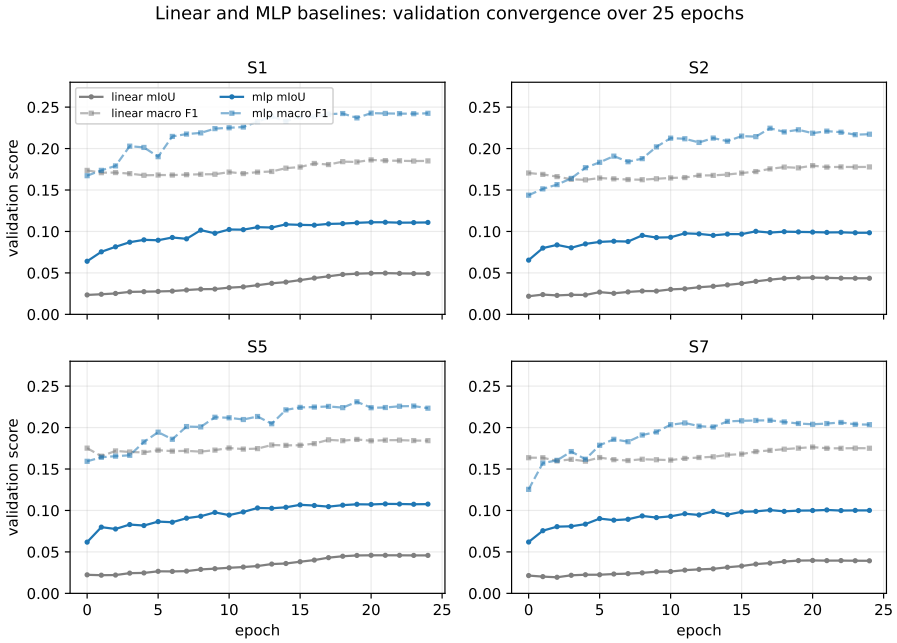

Figure 5: **Linear and MLP baseline convergence.** Per-epoch validation mIoU (solid) and macro F1 (dashed) for each participant over the full 25-epoch budget; both models plateau before the budget is exhausted.

E SHUFFLED-FMRI CONTROL, PER PARTICIPANT

Table 10 reports the shuffled control for each participant separately. All four runs use the identical architecture, budget, and pipeline as the corresponding real model; only the fMRI–stimulus pairing is permuted within each batch.

Table 10: **Shuffled-fMRI control per participant** (150 epochs, identical protocol). The four shuffled models converge to nearly the same prior-level solution; the S2 and S5 rows differ only in the fifth decimal (e.g., mIoU .04133 vs. .04134) and coincide after rounding.

Subj.	Macro F1	Per-image F1	mIoU
S1	.0180	.3121	.0389
S2	.0170	.3102	.0413
S5	.0170	.3102	.0413
S7	.0210	.3308	.0397
Mean	.0182	.3158	.0403
Real (mean)	.2681	.5415	.1553

Pairwise agreement of shuffled models. Table 11 extends the cross-subject consistency analysis to the shuffled controls. Shuffled–shuffled pairs agree strongly (.5954–.7858 mean pairwise IoU)

but show no stimulus specificity, and real–shuffled pairs agree at the low prior-overlap level with no stimulus specificity either. Only real–real pairs separate same- from different-stimulus agreement.

Table 11: **Pairwise agreement by model type** (mean per-image IoU between the two predicted maps, same stimulus versus different stimuli; ranges over participant pairs).

Pair type	Same stim.	Diff. stim.	Gap
Real–Real (6 pairs)	.2354 [.2267, .2492]	.0292	+.2062
Shuffled–Shuffled (6)	.6923 [.5954, .7858]	.6919	+.0004
Real–Shuffled (16)	.0699 [.0644, .0765]	.0692	+.0007

F CATEGORY-CONDITIONED ATTRIBUTION TOPOGRAPHY

Figure 6 displays the three-dimensional UMAP projection of voxelwise attribution profiles for the four participants. The embedding is fitted separately for each participant, so colors encode within-participant similarity only.

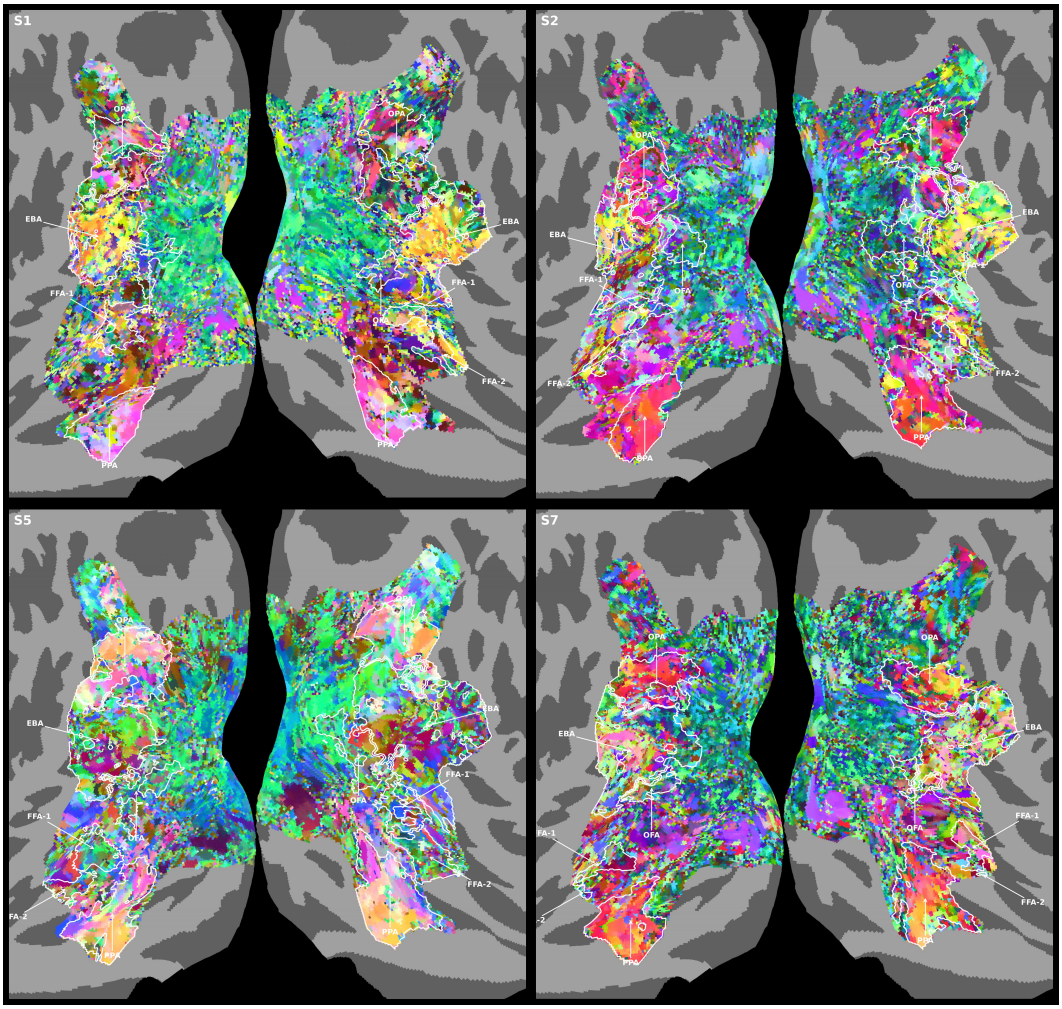

Figure 6: **Exploratory attribution topography.** Voxels are colored by a three-dimensional UMAP embedding (cosine distance) of their per-class attribution profiles. White outlines mark official NSD fLOC parcellations; panels show the posterior cortex covered by the NSD general visual mask.

G PER-SUBJECT TOP-200 SEGMENTATION GALLERIES

Figure 2 shows validation images selected by four-subject mean IoU. To characterize each participant's decoder individually, we rank all 1,000 shared validation images by each subject's own per-image IoU against the teacher and display the top 200 for each participant as triplet walls—stimulus, teacher segmentation, and prediction—in Figs. 7–10. The per-subject top-200 ranges are closely matched across participants (best image .8407–.8723; 200th image .2324–.2553; top-200 mean .3708–.3979), indicating that the qualitative behavior of the paper's examples is not an artifact of cross-subject selection.



Figure 7: **S1: top-200 segmentation triplets.** The 200 highest-IoU validation images for S1, ordered by per-image IoU (descending, left to right and top to bottom). Each triplet shows the stimulus, its teacher segmentation, and the model's prediction (IoU range .8563–.2553, mean .3979). Class colors follow the training-time palette (person white; ignored pixels black).



Figure 8: **S2: top-200 segmentation triplets** (layout as in Fig. 7; IoU range .8723–.2324, mean .3851).



Figure 9: **S5: top-200 segmentation triplets** (layout as in Fig. 7; IoU range .8407–.2401, mean .3708).



Figure 10: **S7: top-200 segmentation triplets** (layout as in Fig. 7; IoU range .8694–.2343, mean .3723).