COMPOSITIONAL BRAIN DECODING VIA VISUAL CONSTITUENT RECOMBINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in fMRI visual decoding recover increasingly rich semantic and perceptual information, yet it remains unclear whether broadly overlapping object and spatial information can be recovered as independently usable constituents rather than only within their original associations. We propose InsideOut, a framework for compositional brain decoding that operationalizes object category and spatial location as what and where constituents. InsideOut explicitly decodes these constituents and recombines them across responses elicited by different images to test whether each preserves its source-specific information after its original association is broken. In the subject-1 composition evaluation, 58.1% of cross-response compositions preserve both category and location, and read-out-level recombination outperforms voxel- and latent-level alternatives. Together, these findings provide a compositional view of brain decoding in which distinct visual constituents can be separately recovered from brain responses and remain reusable beyond the associations in which they were originally observed.

1 INTRODUCTION

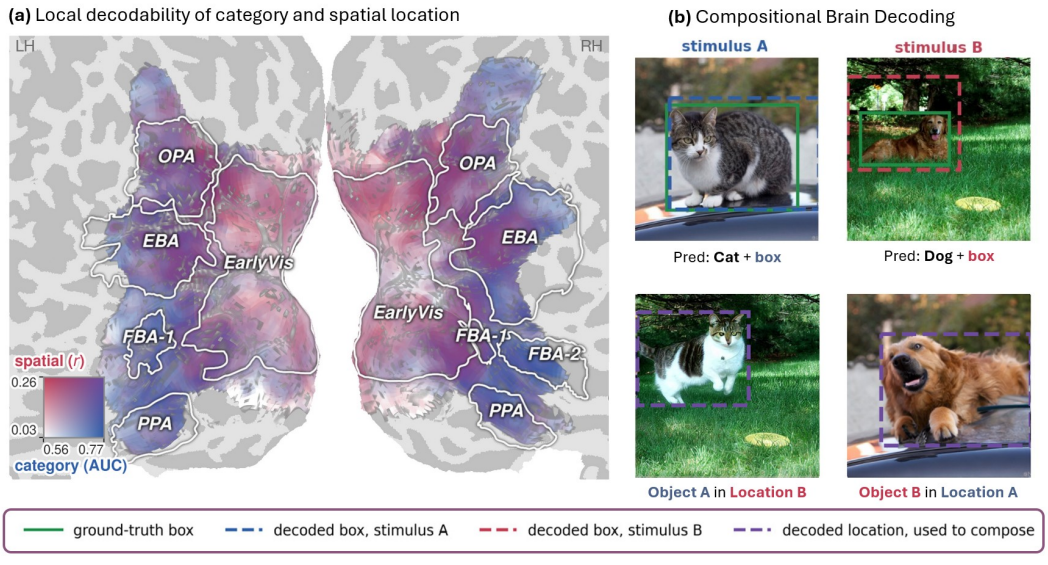


Figure 1: **Object category and spatial location are decodable from overlapping visual cortex and recombined across responses by InsideOut. (a)** Subject-1 searchlight maps (Kriegeskorte et al., 2006) show graded overlap between category and spatial decoding. Blue, red, and purple indicate stronger category decoding, spatial decoding, and both, respectively; white outlines mark visual ROIs. **(b)** InsideOut combines category from one fMRI response with location from another. Green denotes the ground-truth box and dashed lines decoded boxes. Compositions are inpainted into the location-source image for visualization only.

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 major goal of functional magnetic resonance imaging (fMRI) visual decoding is to recover information about a viewed scene from patterns of brain activity. Recent work has progressed from reconstructing viewed images from predicted visual features to recovering increasingly rich semantic and perceptual content with large vision-language and generative models (Shen et al., 2019; Takagi & Nishimoto, 2023; Scotti et al., 2023; 2024; Xia et al., 2024a;b). These approaches decode information represented across multiple stages of the visual system, where different properties of the same stimulus are emphasized to different degrees. Early visual cortex preserves a strongly retinotopic code for stimulus position, whereas representations along the ventral visual hierarchy become increasingly tolerant to changes in position and other visual properties that support robust object recognition (Bosking et al., 2002; Hong et al., 2016). This hierarchy, however, does not imply a clean separation between spatial and object information, because higher ventral areas retain substantial spatial information as well (Connor & Knierim, 2017). Consistent with this view, our searchlight analysis reveals a graded organization of object and spatial information across the visual cortex (Kriegeskorte et al., 2006). Spatial decoding is relatively stronger in and around the early visual cortex, whereas category decoding becomes more prominent in several higher-level visual regions (Fig. 1a). Yet the two signals remain broadly overlapping rather than anatomically segregated, with most local neighborhoods carrying information about both category and location.

Given that object and spatial information are broadly decodable from overlapping regions of visual cortex, an important question is whether a decoder can recover them in a form that remains separately usable. Existing approaches have shown that complementary aspects of a viewed stimulus, such as semantic content and low-level visual or structural information, can be decoded from the same fMRI response and jointly used for reconstruction (Fang et al., 2020; Scotti et al., 2023). However, decoding multiple types of information from the same response does not establish whether each can retain its source-specific content once its original association with the other is broken. We study this question through what we call *compositional brain decoding*: recovering distinct types of visual information from brain responses and testing whether they remain usable when recombined across responses.

We introduce **InsideOut** to instantiate this compositional brain decoding task (Fig. 1b). InsideOut focuses on object category (`what`) and spatial location (`where`) as two visual constituents. InsideOut first recovers a shared visual representation from each fMRI response by aligning brain activity with a pretrained visual embedding. The shared representation is then frozen and passed to separate task-specific read-outs for category and location. Finally, InsideOut recombines responses by pairing the category decoded from one response with the location decoded from another. By breaking the original category-location association, this procedure tests whether each decoded constituent preserves the information from its source response under a novel pairing.

Experiments show that InsideOut reliably recovers both object category and spatial location from individual fMRI responses and preserves both under cross-response recombination. Among the recombination levels we evaluate, read-out-level recombination yields the strongest joint preservation. Qualitative results further show that recombined images retain information from both source responses. Together, these results indicate that object and spatial information can remain separately usable after being detached from their original stimulus pairing.

Our main contributions are:

- We view cross-response recombination as a way to test whether distinct visual information decoded from fMRI remains usable beyond its original stimulus pairing.
- We introduce **InsideOut**, which, to our knowledge, is the first framework to break the original association between decoded `what` and `where` through cross-response recombination, testing whether each constituent is preserved beyond its original context.
- We demonstrate that both constituents remain significantly preserved under cross-response recombination, and that read-out-level recombination provides the strongest joint preservation among the recombination levels evaluated.

2 Related Work

Visual Decoding from fMRI. Prior fMRI decoding research has progressed from reconstructing viewed images from predicted visual features to recovering richer perceptual and semantic content

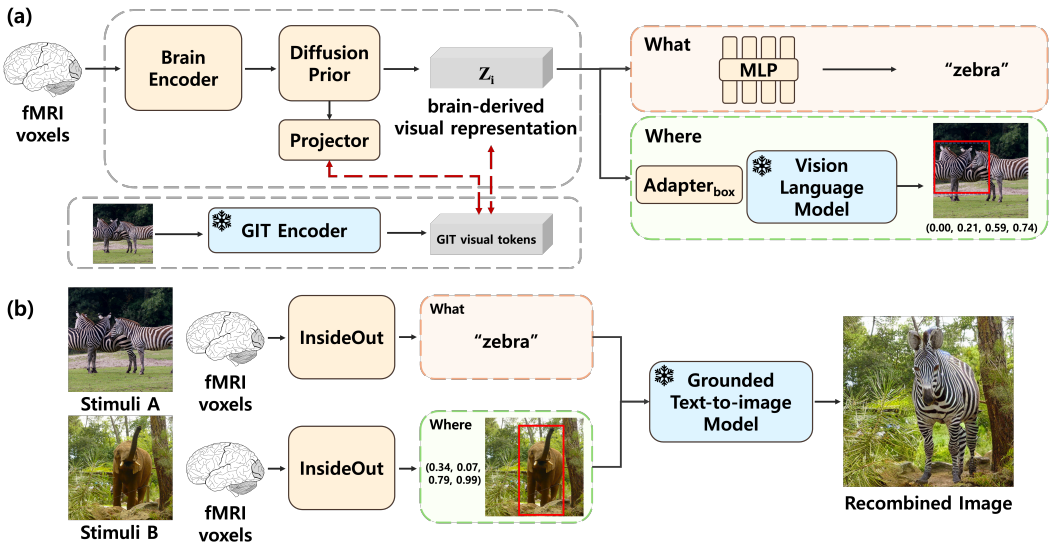


Figure 2: **Overview of InsideOut. (a)** InsideOut maps fMRI responses to a shared visual representation and decodes object-category in what branch and spatial constituents in where branch, together with a scene-level caption read-out. **(b)** Constituents decoded from different fMRI responses are recombined and used to generate a new image with a pretrained grounded text-to-image model, enabling evaluation of source-specific content preservation.

with pretrained vision-language and generative models (Shen et al., 2019; Takagi & Nishimoto, 2023; Ozcelik & VanRullen, 2023; Scotti et al., 2023; 2024). Beyond image reconstruction, some approaches explicitly decode semantic and spatial information, including scene descriptions, object identities, and bounding boxes (Ferrante et al., 2023; Han et al., 2024; Xia et al., 2024b). In parallel, several methods factorize visual information to improve reconstruction, either by separating semantic content from low-level structure or by organizing representations according to functionally motivated components of visual processing (Ozcelik & VanRullen, 2023; Huo et al., 2024; He et al., 2026; Xia et al., 2024a; Joo et al., 2024; Li et al., 2025; Meng et al., 2025). Existing work therefore establishes that semantic and spatial information can be decoded and in some cases factorized, from fMRI. What remains unresolved is whether these decoded constituents remain independently meaningful once detached from the visual associations in which they were originally observed. We address this by explicitly recovering object category and bounding box from individual fMRI responses and evaluating whether each preserves its content under cross-response recombination.

Compositionality in Neural Representations. Prior work has explored compositionality in fMRI through classification-based constituent decoding (Barbu et al., 2015), symbolic compositional generalization (Wang et al., 2026), and algebraic combination of concept-related neural patterns (Ferrante et al., 2025). These approaches examine compositionality under predefined classes, structured queries, or concept-level manipulations. Taken together, however, they leave direct constituent recovery and preservation of recovered content beyond its original context jointly unexplored, particularly for natural scenes without predefined visual factors. In contrast, we study whether visual constituents can be directly recovered from individual fMRI responses to natural scenes and whether the recovered content is preserved when separated from its original visual context.

3 Compositional Recombination in Brain Decoding

This section formalizes compositional recombination of task-specific read-outs obtained from fMRI responses. As shown in Fig. 2, the `what` and `where` read-outs in InsideOut recombine. Let $\mathbf{v}_A$ and $\mathbf{v}_B$ denote fMRI responses in the voxel space $\mathcal{V}$, elicited by visual stimuli $\mathbf{I}_A$ and $\mathbf{I}_B$, respectively.

For each response $\mathbf{v}_k$, where $k \in \{A, B\}$, a shared mapping $\mathcal{F} : \mathbb{R}^{d_v} \mapsto \mathbb{R}^{d_1 \times d_2}$ first produces a brain-derived visual representation. Then, task-specific read-outs $\mathcal{T}_{\texttt{what}} : \mathbb{R}^{d_1 \times d_2} \mapsto \mathbb{R}^{|\mathcal{O}|}$ and

[1]$\mathcal{T}_{\texttt{where}} : \mathbb{R}^{d_1 \times d_2} \mapsto \mathbb{R}^4$ then recover object category o_k and spatial location $\mathbf{s}_k$:

$$o_k = \text{argmax}_{o \in \mathcal{O}}(\mathcal{T}_{\texttt{what}} \circ \mathcal{F})(\mathbf{v}_k), \quad \mathbf{s}_k = (\mathcal{T}_{\texttt{where}} \circ \mathcal{F})(\mathbf{v}_k), \tag{1}$$

where $\mathcal{O}$ denotes the set of object categories, and $\mathbf{s}_k \in \mathcal{S} \subseteq \mathbb{R}^4$ represents a 2D bounding box. We define a composition function $\mathcal{H} : \mathcal{O} \times \mathcal{S} \to \mathcal{C}$, where $\mathcal{C}$ denotes the space of constituent combinations. Separately recovering the constituents enables cross-response recombination, which breaks their original association and places each constituent in a novel pairing.

Now, we define compositional recombinations as $\mathbf{c}_{ij} = \mathcal{H}(o_i, \mathbf{s}_j) \in \mathcal{C}$. Within-response combinations correspond to $i = j$, whereas cross-response recombinations correspond to $i \neq j$:

$$\mathbf{c}_{\text{AA}} = \mathcal{H}(o_{\text{A}}, \mathbf{s}_{\text{A}}),\ \mathbf{c}_{\text{BB}} = \mathcal{H}(o_{\text{B}}, \mathbf{s}_{\text{B}}),\ \mathbf{c}_{\text{AB}} = \mathcal{H}(o_{\text{A}}, \mathbf{s}_{\text{B}}),\ \mathbf{c}_{\text{BA}} = \mathcal{H}(o_{\text{B}}, \mathbf{s}_{\text{A}}). \tag{2}$$

While within-response combinations preserve the original `what–where` association, their cross recombinations form new constituent pairings that are not jointly decoded from either source response.

For the cross-response recombinations $\mathbf{c}_{\text{AB}}$ and $\mathbf{c}_{\text{BA}}$, let $\mathbf{I}_{\text{AB}}$ and $\mathbf{I}_{\text{BA}}$ denote the images generated from them, respectively. A successful recombination should preserve the object-category information from one source response while following the spatial information from the other. This requires each decoded constituent to retain its source-specific information through recombination. In Section 4, we describe the architectural details of $\mathcal{F}$, task-specific heads $\mathcal{T}$, and $\mathcal{H}$ used for this purpose.

4 INSIDEOUT

In this section, we detail InsideOut's workflow, as shown in Fig. 2. We describe the parameterizations of the shared mapping $\mathcal{F}$ in Section 4.1, task-specific read-outs $\mathcal{T}$ in Section 4.2, and composition function $\mathcal{H}$ in Section 4.3, followed by the training regularization objective in Section 4.4.

4.1 BRAIN-DERIVED VISUAL REPRESENTATION

Given an fMRI response $\mathbf{v}_i$, InsideOut learns a shared brain-derived visual representation for subsequent constituent decoding. Voxels from the nsdgeneral ROI are processed by an encoder E_θ, which first transforms the voxel sequence into 2D feature maps and then applies CycleMLP-based blocks (Chen et al., 2022). The resulting brain representation conditions a diffusion prior p_θ to recover visual tokens, following the brain-to-visual mapping used in MindEye (Scotti et al., 2023):

$$\mathbf{u}_i = \text{E}_\theta(\mathbf{v}_i), \qquad \mathbf{z}_i := \mathbf{z}^0, \qquad \mathbf{z}^{t-1} \sim p_\theta\left(\mathbf{z}^{t-1} \mid \mathbf{z}^t, \mathbf{u}_i\right), \quad t = T, \ldots, 1, \tag{3}$$

where $\mathbf{z}_i \in \mathbb{R}^{257 \times 1024}$ with $d_1 = 257$ and $d_2 = 1024$, and $\mathbf{z}^T \sim \mathcal{N}(0, \mathbf{I})$. Together, E_θ and p_θ establish the shared mapping $\mathcal{F}$ introduced in Section 3.

Using visual tokens extracted from the corresponding stimulus image $\mathbf{I}_i$ by a frozen pretrained GIT-large-coco image encoder (Wang et al., 2022) as targets, we jointly train E_θ and p_θ to map v_i into the target visual-token space. Once trained, E_θ and p_θ are frozen, and the resulting $\mathbf{z}_i$ serves as the shared input to the task-specific read-outs for object category, caption, and spatial location.

4.2 CONSTITUENT DECODING

For each stimulus, we use the category label and bounding box of a representative salient COCO instance (Lin et al., 2014) to supervise $\mathcal{T}_{\texttt{what}}$ and $\mathcal{T}_{\texttt{where}}$, respectively. Since COCO images often contain multiple annotated objects, we select a single instance among those covering 0.5 to 95 percent of the image by favoring objects that are both large and near the image center, consistent with the central-fixation protocol of NSD.

What Decoding. In the `what` branch $\mathcal{T}_{\texttt{what}}$, with the upstream brain-to-visual mapping frozen, a parameterized category classifier M_θ decodes the explicit category label o_i from the shared brain-derived visual representation $\mathbf{z}_i$, for cross-response recombination.

Object-category prediction is defined as

$$\mathbf{p}_i^{\text{cat}} = \text{Softmax}\left(M_\theta(\mathbf{z}_i)\right), \qquad o_i = \underset{k}{\text{argmax}}\, p_{i,k}^{\text{cat}}. \tag{4}$$

[1]$\mathcal{T}_{\texttt{where}}$ includes an adapter for feature alignment, detailed in Section 4.2.

Here, the MLP-based classifier M_θ is trained with cross-entropy loss and estimates a distribution over COCO object categories. At inference, we use the highest-probability category as o_i.

Where Decoding. The `where` branch $\mathcal{T}_{\texttt{where}}$ decodes spatial location from the same shared representation $\mathbf{z}_i$. The bounding box is transformed from the COCO coordinate system into the NSD crop coordinate system and represented as normalized coordinates, corresponding approximately to a viewer-centered, retinotopic 2D location under the central-fixation NSD protocol rather than an allocentric or 3D one.

For spatial decoding, $\mathcal{T}_{\texttt{where}}$ exploits pretrained LLaVA-1.5-7B (Liu et al., 2024a) with a separately box-specific BrainAdapter A_θ^{box} to align the representation dimensions. The adapted brain-derived tokens replace the image tokens in a fixed localization instruction that does not specify the object category. This allows the `where` read-out to operate independently of `what`, after which LLaVA generates the coordinates as text. This differs from brain grounding in UMBRAE (Xia et al., 2024b), whose grounding prompt names the object to be localized. The box is then obtained as

$$\mathbf{s}_i = \text{Parse}\left(\text{LLaVA}\left(A_\theta^{\text{box}}(\mathbf{z}_i)\right)\right) = (x_{1,i}, y_{1,i}, x_{2,i}, y_{2,i}), \tag{5}$$

where $\text{Parse}(\cdot)$ takes the first four numbers in LLaVA's greedy output and clips each to $[0, 1]$. We train only A_θ^{box}, using cross-entropy on the ground-truth coordinates to three decimals. The resulting $\mathbf{s}_i$ provides an explicit spatial read-out of the representative object, capturing its position and extent.

4.3 Cross-Response Compositional Generation

For cross-response recombinations, we implement the composition function $\mathcal{H}$ using a pretrained GLIGEN model (Li et al., 2023), denoted by G_θ. Because GLIGEN jointly conditions on an object category and a normalized bounding box, it directly constructs compositions of the decoded `what` and `where` constituents as follows:

$$\mathbf{I}_{\text{AB}} = G_\theta(o_{\text{A}}, \mathbf{s}_{\text{B}}), \qquad \mathbf{I}_{\text{BA}} = G_\theta(o_{\text{B}}, \mathbf{s}_{\text{A}}). \tag{6}$$

Across all recombinations, we fix the generator and vary constituent pairing to isolate the effect of cross-response recombination.

4.4 Training with Improved EFDM

InsideOut is trained in stages. We first optimize the shared mapping $\mathcal{F}$ using the MSE and contrastive objectives of MindEye (Scotti et al., 2023), augmented with our improved EFDM loss for distributional regularization. We then train task-specific read-outs, as detailed in Section 4.2.

We propose a variant of the EFDM loss to complement element-wise MSE, which matches predicted and target tokens. Given two vectors $a, b \in \mathbb{R}^L$, let π denote the permutation that sorts a in ascending order, and let b_k denote the k-th smallest value of b. We define $T(a, b)_{\pi(k)} = b_k$ for $k = 1, \ldots, L$. Thus, $T(a, b)$ preserves the rank order of a while taking the values of b.

We apply T to the predicted visual tokens $\mathbf{z}$ and target visual tokens $\tilde{\mathbf{z}}$ along both token and channel dimensions:

$$\mathcal{L}_{\text{EFDM}} = \frac{1}{B\mathrm{d}_1\mathrm{d}_2}\left(\sum_{b,d=1}^{B,\mathrm{d}_1} \|z_b[:,d] - T(z_b[:,d], \tilde{z}_b[:,d])\|_2^2 + \sum_{b,t=1}^{B,\mathrm{d}_2} \|z_b[t,:] - T(z_b[t,:], \tilde{z}_b[t,:])\|_2^2\right), \tag{7}$$

where B denotes the batch size. The first term matches the distribution over $\mathrm{d}_1 = 257$ token values within each channel, while the second matches the distribution over $\mathrm{d}_2 = 1024$ channel values within each token. Unlike the original EFDM (Zhang et al., 2022), which directly transforms features, we use the sort-matched representation as a detached training target. This formulation preserves the rank structure of the prediction while penalizing contraction of its value distribution.

5 Experiments

5.1 Experimental Setup

Data. We use NSD (Allen et al., 2022) subjects 1, 2, 5 and 7, who completed all 40 scanning sessions, following prior work (Xia et al., 2024a; Scotti et al., 2024). Each model is trained on



Figure 3: **Qualitative examples of cross-response composition.** Subject 1. Each group shows two source stimuli and their two cross-compositions. A@B combines the category decoded from response A with the box decoded from response B. Green denotes the ground-truth box; blue/red dashed boxes are decoded from A/B, and purple dashed boxes are used for composition. Compositions are inpainted into the location-source image for visualization only; Table 2 evaluates renders generated from the decoded category and box alone. *det* is detector confidence (Liu et al., 2024b) multiplied by IoU with the requested box.

Table 1: **Quantitative results.** Cross-subject models are trained on several subjects. † mean of our four single-subject models. Prior caption metrics are as reported, and ours use the 969 of the 982 BrainHub test images whose reference captions match our test set. RefCLIP-S is the harmonic mean of CLIP-S and similarity to the reference captions. Category and box are scored against the salient object of the 959 test images whose BrainHub index we could verify. InsideOut reads both directly with its read-outs. Prior work outputs no category, so its category is the one named in its captions, and its box is the one it returns when given the true category name.

	caption										category	box
	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr	SPICE	CLIP-S	RefCLIP-S	top-1	IoU
single-subject models												
SDRecon-S1 (CVPR'23)	36.21	17.11	7.72	3.43	10.03	25.13	13.83	5.02	61.07	66.36	0.268	–
OneLLM-S1 (CVPR'24)	47.04	26.97	15.49	9.51	13.55	35.05	22.99	6.26	54.80	61.28	0.367	–
UMBRAE-S1 (ECCV'24)	57.63	38.02	25.00	16.76	18.41	42.15	51.93	11.83	66.44	72.12	0.533	–
VINDEX-S1 (ICCV'25)	57.99	39.00	26.04	17.83	18.56	42.27	53.96	11.92	**66.94**	**72.64**	0.519	0.337
InsideOut	**61.94**	**44.11**	**30.92**	**21.80**	**20.84**	**46.75**	**68.24**	**13.39**	65.90	72.11	**0.612**	**0.442**
cross-subject models												
UMBRAE (ECCV'24)	59.44	40.48	27.66	19.03	19.45	43.71	61.06	12.79	67.78	73.54	0.575	0.400
VINDEX (ICCV'25)	60.00	40.72	27.57	18.91	19.51	43.78	60.32	12.84	**68.26**	**73.88**	–	–
InsideOut†	**61.07**	**42.80**	**29.58**	**20.56**	**20.19**	**46.03**	**63.92**	**12.93**	65.25	71.46	**0.595**	**0.427**

one subject's 9,000 training images, corresponding to 27,000 trials, and tested on the 1,000 shared images with the three repetitions averaged. Pretrained components remain frozen.

Evaluations. Category and box are scored against the salient object on the 959 test images whose index in BrainHub (Xia et al., 2024b) we could verify, so that prior decoders are scored on the same images. Captions follow the BrainHub protocol on the 969 of its 982 test images whose reference captions match our test set.

Caption Read-Out. Prior brain decoders output captions rather than categories, so Table 1 reads their category from their captions. To also compare them on their own output, and to check that $\mathbf{z}_i$ keeps semantic scene content beyond the single salient object, we add a caption read-out. It follows the `where` design with its own BrainAdapter A_{cap} and the fixed instruction *"Describe this image in*

Table 2: **Composition, subject 1.** Each cell renders the category read from one response at the location read from another. Identity, location and *joint* give the share of renders with the category, the location or both correct, on the same 1,000 cells in every row. (a) The stage at which the two responses are joined. (b) Ours against references that do not decode the brain, human labels as a ceiling, a random floor, and the ground-truth image given to our read-outs. In (b), category and box are scored as in Table 1 on its 959 images.

(a) level of recombination

stage	method	identity	location	joint
voxel	Brain Algebra	0.779	0.516	0.448
latent	blend the two latents	0.481	0.324	0.188
InsideOut		**0.836**	**0.641**	**0.581**

(b) against references

	category	box	identity	location	joint
ceiling, human labels	1.000	1.000	0.985	0.941	0.928
floor, random	0.098	0.201	0.438	0.319	0.218
GT image	0.843	0.535	0.905	0.705	0.690
InsideOut	0.612	0.442	0.836	0.641	0.581

one sentence." It is trained on the same $\mathbf{z}_i$ with cross-entropy on the COCO captions of the training images, decodes greedily, and is used only for this comparison, not for recombination.

Composition. Compositions are rendered with the text-and-box generation checkpoint of GLIGEN (Li et al., 2023), using the prompt "a photograph of a $\langle o \rangle$.", 50 denoising steps, grounding for the first 30 percent of steps ($\beta = 0.3$) and a seed fixed per pair. No image is given to the generator. We evaluate 500 pairs of subject-1 test stimuli, which give 1,000 cross-response cells. The pairs are drawn at random among stimuli whose ground-truth objects differ in category and each cover 15 to 50 percent of the image, with box centers at least 0.28 apart horizontally and an IoU of at most 0.3. Each stimulus enters at most four pairs. The selection uses ground-truth boxes only and no decoder output.

Scoring a Composition. A cross-response cell renders the category decoded from one response at the box decoded from another (Section 4.3) and is scored against the two source stimuli. Identity is preserved when an open-vocabulary detector (Liu et al., 2024b) prefers the true category of the category source over that of the other stimulus, and location is preserved when the detected box overlaps the true box of the location source more than the other one. *Joint* requires both. The generator, seed rule and detector are fixed, so only the origin of the constituents changes. Three references skip brain decoding. The ceiling renders the human-annotated category and box, the floor a random category at a random box, and the GT image row gives the stimulus image's own embedding to our read-outs.

5.2 Constituent Decoding

On subject 1 (Table 1), the category read-out reaches a top-1 accuracy of 0.612 against 0.533 for the strongest single-subject decoder, UMBRAE-S1, and the box read-out reaches a mean IoU of 0.442 without the category, against 0.337 for VINDEX-S1 (Xia & Oztireli, 2025) given the true category. On captions, the native output of prior decoders, InsideOut leads on eight of ten metrics and trails VINDEX-S1 only on CLIP-S and RefCLIP-S. Our four-subject mean leads the cross-subject models on the same eight metrics and exceeds the cross-subject UMBRAE on category and box (0.595 and 0.427 against 0.575 and 0.400). The comparison is not matched, because those models are trained jointly on all four subjects, whereas each of ours sees the data of one subject only. Figure 4(a) shows examples.

5.3 Cross-Response Recombination

InsideOut preserves the category in 0.836 of the 1,000 cells, the location in 0.641 and both in 0.581 (Table 2(b)). Each score compares the two source stimuli, so a random category at a random box still reaches 0.218, and the human category and box reach 0.928, the loss of the generator and detector alone. Given the stimulus image instead of a brain response, our read-outs reach 0.690. Of the 0.347 gap to the ceiling, 0.238 thus arises in the read-outs and only 0.109 in brain decoding. Location is harder than category in every row. Figures 3 and 4(b) show examples.

Level of Recombination. Table 2(a) varies only where the two responses are combined. Brain Algebra (Ferrante et al., 2025) adds to the location-source voxels a pattern for the category decoded

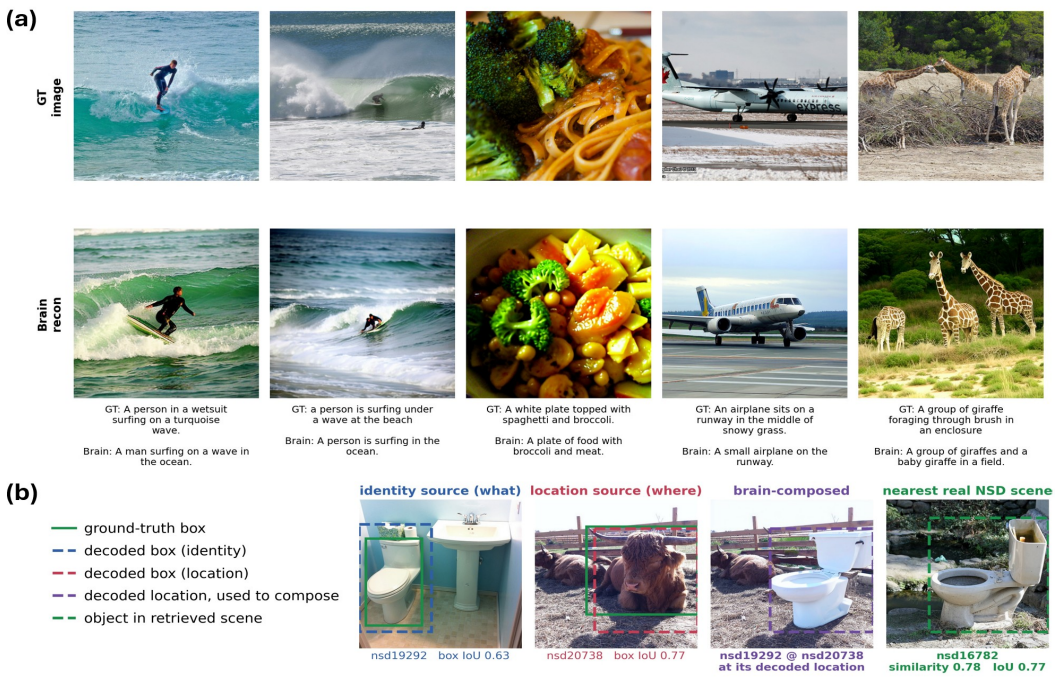


Figure 4: **Single-response decoding and a composition next to a real scene. (a)** Test stimuli, reconstructions with Versatile Diffusion (Xu et al., 2023), and ground-truth (GT) and decoded (Brain) captions for subjects 2 and 5, chosen for display. **(b)** A subject-1 composition next to its most similar real NSD stimulus by SwAV (Caron et al., 2020) among those holding the same category at a box of IoU at least 0.5, so the pairing occurs in real scenes. Compositions are inpainted into the location-source photograph for visualization.

from the other response and reaches 0.448 joint, and blending the two latents falls below the floor at 0.188. Recombining the read-outs is highest on all three scores, mostly through location.

Swapping the Encoder. With the single-subject MindEye1 representation (Scotti et al., 2023) in place of ours (Table 3 in Appendix), joint preservation falls from 0.581 to 0.395, again mostly through location (0.641 to 0.452). There, category and box are measured on the source stimuli of the 1,000 cells, hence 0.625 and 0.466 for InsideOut.

5.4 Cortical Analysis

Constituent Decodability. The searchlight of Figure 1(a) uses only recorded voxels and COCO labels. Each searchlight, of radius 3 voxels, decodes category by nearest centroid (one-vs-rest AUC) and box center by ridge regression after removing its category-predictable part, trained on the training images and tested on the 981 shared images. In 90.4 to 97.0 percent of sampled centers both exceed a permutation threshold, so the constituents are not divided between regions and InsideOut separates them only at the read-outs.

Regional Contributions. Figure 5 edits subject-1 test responses before they enter the fixed model. Removing early visual cortex lowers the box IoU most (0.447 to 0.385) but category little (0.590 to 0.538), whereas removing face and body regions lowers category most (to 0.412) but the box little (to 0.414). Kept alone, early visual cortex gives a box IoU of 0.361 at a category accuracy of 0.124, and face and body regions give 0.410 category at 0.348 box. Since the model receives one undivided voxel vector, this division arises in training. Whole-brain values use a re-sampled prior and differ slightly from Table 1.

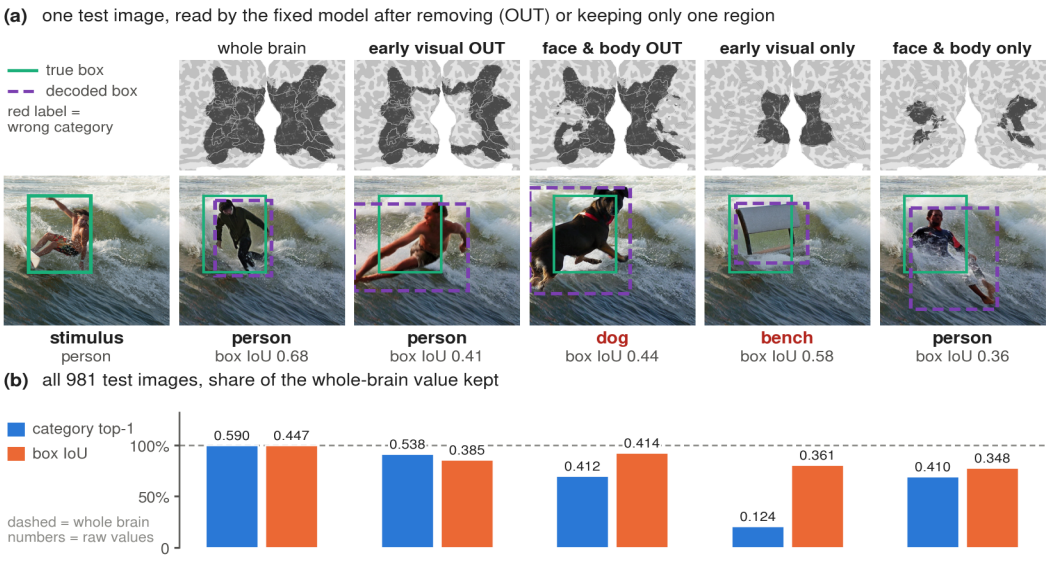


Figure 5: **Cortical ablations differentially affect category and spatial decoding.** Subject 1. Voxels inside a region group (OUT) or outside it (only) are set to their training mean before the fixed model reads the response. Early visual is V1 to hV4, and face and body is OFA, FFA, EBA and FBA. **(a)** One test image, chosen among the 16 of 326 correctly read images that show all effects. Green, ground truth. Purple dashed, decoded box. Red, wrong category. **(b)** Category top-1 and box IoU over 981 test images relative to the whole brain, with raw values above the bars.

5.5 Discussion

Spatial location is both the weaker constituent and the one that separates the alternatives. Voxel-level recombination keeps most of the category but loses location, suggesting that a category pattern places the concept but not a specific box, and a weaker encoder again costs mostly location. With our representation, location from brain responses (0.641) nears that from the stimulus image (0.705), so the remaining gap to the ceiling (0.941) lies mostly in the `where` read-out. That early visual cortex carries more of the box, without being assigned to it, fits this picture.

6 Conclusion

InsideOut decodes object category and spatial location from single fMRI responses as separate read-outs of one shared representation and recombines them across responses. On subject 1 both read-outs exceed the published decoders we compare against, and its captions lead on eight of ten metrics among single-subject models. Recombined, the two constituents are jointly preserved in 0.581 of the cells, between a floor of 0.218 and a ceiling of 0.928, and better than when voxels or latents are recombined. Location is the harder constituent, and its remaining loss lies mostly in the `where` read-out. Although the model receives one undivided response, removing early visual or face and body cortex impairs the box and the category, respectively. Two limitations remain. Each scene is reduced to one salient object, leaving attributes, relations and multi-object scenes untested, and preservation is scored by a fixed generator and detector against COCO labels, which measures whether constituents can be rendered and recognized rather than how images are perceived.

AI use statement

Generative AI tools were used for translation, language editing, drafting and revising parts of the manuscript, and improving its clarity and organization. They were also used for literature retrieval. All research decisions, implementations, analyses, and conclusions were independently verified by the authors. The authors reviewed all AI-assisted content and take full responsibility for the final manuscript.

REFERENCES

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

Andrei Barbu, N. Siddharth, Caiming Xiong, Jason J. Corso, Christiane D. Fellbaum, Catherine Hanson, Stephen José Hanson, Sébastien Hélie, Evguenia Malaia, Barak A. Pearlmutter, Jeffrey Mark Siskind, Thomas Michael Talavage, and Ronnie B. Wilbur. The compositional nature of event representations in the human brain. *arXiv preprint arXiv:1505.06670*, 2015.

William H Bosking, Justin C Crowley, and David Fitzpatrick. Spatial coding of position and orientation in primary visual cortex. *Nature neuroscience*, 5(9):874–882, 2002.

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

Shoufa Chen, Enze Xie, Chongjian Ge, Runjian Chen, Ding Liang, and Ping Luo. Cyclemlp: A mlp-like architecture for dense prediction. In *International Conference on Learning Representations*, 2022.

Charles E Connor and James J Knierim. Integration of objects and space in perception and memory. *Nature neuroscience*, 20(11):1493–1503, 2017.

Tao Fang, Yu Qi, and Gang Pan. Reconstructing perceptive images from brain activity by shape-semantic gan. *Advances in Neural Information Processing Systems*, 33:13038–13048, 2020.

Matteo Ferrante, Furkan Ozcelik, Tommaso Boccato, Rufin VanRullen, and Nicola Toschi. Brain captioning: Decoding human brain activity into images and text. *arXiv preprint arXiv:2305.11560*, 2023.

Matteo Ferrante, Tommaso Boccato, Nicola Toschi, and Rufin VanRullen. Evidence for compositionality in fMRI visual representations via Brain Algebra. *Communications Biology*, 8(1):1263, 2025. doi: 10.1038/s42003-025-08706-4.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26574–26585. IEEE, 2024.

Wanqi He, Jin Wang, Hui Li, Hanyang Chi, and Bingfeng Zhang. Balancing semantic and structural decoding for fMRI-to-image reconstruction. *Expert Systems with Applications*, 298:129836, 2026. doi: 10.1016/j.eswa.2025.129836.

Ha Hong, Daniel LK Yamins, Najib J Majaj, and James J DiCarlo. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature neuroscience*, 19(4):613–622, 2016.

Jingyang Huo, Yikai Wang, Yun Wang, Xuelin Qian, Chong Li, Yanwei Fu, and Jianfeng Feng. NeuroPictor: Refining fMRI-to-image reconstruction via multi-individual pretraining and multi-level modulation. In *European Conference on Computer Vision (ECCV)*, pp. 56–73, 2024. doi: 10.1007/978-3-031-72983-6_4.

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

Nikolaus Kriegeskorte, Rainer Goebel, and Peter Bandettini. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, 103(10):3863–3868, 2006.

Chong Li, Jingyang Huo, Weikang Gong, Yanwei Fu, Xiangyang Xue, and Jianfeng Feng. DecoFuse: Decomposing and fusing the "what", "where", and "how" for brain-inspired fMRI-to-video decoding. *arXiv preprint arXiv:2504.00432*, 2025.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22511–22521, 2023.

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 *European Conference on Computer Vision*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55. Springer, 2024b.

Lu Meng, Zengyu Song, and Jiayang Lu. Brain-Imager: A multimodal framework for image reconstruction and captioning from human brain activity. *Brain Informatics*, 12:32, 2025. doi: 10.1186/s40708-025-00282-x.

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

Paul S. Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J. Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth A. Norman, and Tanishq Mathew Abraham. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36, 2023.

Paul S. Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, 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 (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44038–44059, 2024.

Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS Computational Biology*, 15(1):e1006633, 2019. doi: 10.1371/journal.pcbi.1006633.

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

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research*, 2022.

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

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

Weihao Xia, Raoul de Charette, Cengiz Öztireli, and Jing-Hao Xue. DREAM: Visual decoding from reversing human visual system. In *Winter Conference on Applications of Computer Vision (WACV)*, 2024a.

Weihao Xia, Raoul de Charette, Cengiz Öztireli, and Jing-Hao Xue. UMBRAE: Unified multimodal brain decoding. In *European Conference on Computer Vision (ECCV)*, 2024b.

Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7754–7765, 2023.

Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8035–8045, 2022.

APPENDIX CONTENTS

A Searchlight analysis of Figure 1(a) 13
B Implementation details 13
C Composition details 14
D Composition across four subjects 15
E Decoding results of each subject 15
F Additional results 15
F.1 Searchlight maps of the other subjects 15
F.2 All cortical ablations 15
F.3 Whether the two read-outs point to the same object 16
F.4 Correspondence with real scenes 16
F.5 More single-response decoding 17
F.6 More compositions 17

A SEARCHLIGHT ANALYSIS OF FIGURE 1(A)

Data. We use single-trial NSD betas in 1.8 mm space within nsdgeneral (15,724 voxels in subject 1 and 12,682 to 14,278 in the others), z-scored per voxel within each session and averaged per image. Training uses the non-shared images after removing all trials of the shared images (8,821 images with a salient object in subject 1). Testing uses the 981 shared images with a salient object. Labels are the category and box center of the salient object (Section 4.2).

Searchlight. Each nsdgeneral voxel is the center of a sphere of radius 3 voxels (about 5.4 mm), holding 9 to 117 voxels (median 77 in subject 1).

Category. Unit-length patterns are compared by cosine to the unit-length mean training pattern of each category. The score is the one-vs-rest AUC averaged over the 12 categories with at least 20 test and 10 training images (chance 0.5). We use AUC because person alone covers 29 percent of the test images.

Location. Category predicts vertical position in COCO, so we subtract the training-set mean box center of each category from (c_x, c_y). Ridge regression ($\alpha = 1000$, not tuned) predicts each residual coordinate, and the score is the mean of the two test Pearson correlations (chance 0). Results barely change between $\alpha = 1$ and 1000.

Test. Overlapping searchlights are not independent, so we test the map as a whole. At 500 centers per subject, drawn in advance, both scores are recomputed under 1,000 permutations of the training labels shared across centers. A center passes a score when it exceeds the 99th percentile of its own null. In subjects 1, 2, 5 and 7, 94.2, 97.0, 90.4 and 91.2 percent of centers pass both, against at most 0.8 percent in any permuted map ($p = 0.001$ for each subject).

Color. Each score is scaled between its 2nd and 98th percentiles over all searchlights of the four subjects (AUC 0.560 to 0.775, r 0.034 to 0.261). The color interpolates between white, blue (category), red (location) and purple (both).

B IMPLEMENTATION DETAILS

Input. Betas of the nsdgeneral voxels (15,724 in subject 1) are z-scored per session, and each response is min-max scaled to $[0, 1]$. Training uses the 27,000 single trials of the 9,000 training images, and testing uses the 1,000 shared images averaged over three repetitions.

Encoder. Four 1D convolutions (kernel 7, 64 channels) turn the voxels into a 64×64 map with 64 channels, which feeds two CycleMLP-B1 branches. Their outputs are concatenated, mapped to 257 tokens and mixed by ten CycleMLP blocks into the 257×1024 condition $\mathbf{u}_i$. All weights are randomly initialized.

Prior and target. A transformer prior (depth 6, width 1024, 19 heads) predicts the clean tokens over 100 diffusion steps, with the condition dropped at a rate of 0.2. The target is the 257×1024 last hidden state of the GIT-large-COCO image encoder (Wang et al., 2022) for the stimulus resized to 224×224. At test time we draw one sample with the full 100-step sampler and no guidance.

Objective and optimization. The loss is $30\,\mathcal{L}_{\text{prior}} + 10\,\mathcal{L}_{\text{SoftCLIP}} + \mathcal{L}_{\text{EFDM}}$, where $\mathcal{L}_{\text{prior}}$ is the MSE of the predicted tokens and $\mathcal{L}_{\text{SoftCLIP}}$ is the soft contrastive loss of MindEye (Scotti et al., 2023) on a projection of $\mathbf{u}_i$, without MixCo. We use AdamW (learning rate 10^{-4}, weight decay 0.05), a one-cycle schedule, gradient clipping at 1.0 and seed 42, with 22 samples per GPU. A first stage runs 196 epochs, on one GPU for the first 80 and three afterwards. A second stage restarts the schedule from it for up to 200 epochs on three GPUs, and we use its epoch 153.

Category read-out. Mean pooling over tokens, LayerNorm, a 1024-to-512 layer with GELU and dropout 0.1, and a linear layer over the 79 categories present in the subject's labels. It is trained on $\mathbf{z}_i$ of the training trials with cross-entropy (AdamW, learning rate 10^{-4}, weight decay 0.01, batch 128, 20 epochs), and the epoch is chosen by top-1 on 10 percent of the training images held out.

Adapters. A_{box} and A_{cap} compute $\mathbf{x} + W_2\,\text{GELU}(W_1\,\text{LN}(\mathbf{x}))$ with hidden size 2048 and W_2 initialized to zero. All 257 adapted tokens pass through the frozen projector of LLaVA-1.5-7B and replace its image tokens. LLaVA stays frozen, and only the adapter is trained, with cross-entropy on the answer tokens.

Where read-out. The instruction is *"Locate the most salient object and output its bounding box as (x1,y1,x2,y2) normalized 0-1."* and the answer is "(x_1, y_1, x_2, y_2)" with three decimals. Trials without a salient object (2 percent) are trained on captions instead. We use AdamW (learning rate 10^{-4}, weight decay 0.01), a cosine schedule and 32 samples per step on two GPUs, and keep the adapter after five epochs. Decoding is greedy with at most 40 new tokens.

Caption read-out. The answer is one COCO caption per image, the first that names the salient category or otherwise the longest. We train for five epochs with AdamW (learning rate 10^{-4}, weight decay 0.01), a cosine schedule and 16 samples per step, keep the last epoch, and decode greedily with at most 30 new tokens.

Salient object. Among the COCO instances inside the NSD crop that cover 0.5 to 95 percent of it, we take the one maximizing $a\,(1 - d/0.7071)$, where a is its area and d the distance of its box center from the image center. Images without such an instance are left out (198 of 10,000 in subject 1).

C Composition Details

Rendering. GLIGEN (`gligen-1-4-generation-text-box`) renders one 512×512 image per cell with 50 PNDM steps, guidance scale 7.5, grounding for the first 30 percent of steps and seed $42 + n$ for pair n, shared by its two cells and by every method. No image is given to the generator.

Scoring. Grounding DINO (`grounding-dino-base`, box threshold 0.25, text threshold 0.20) is queried with each category name alone, and its top detection is kept. Identity is preserved when the true category of the category source scores higher than that of the other stimulus. Location is preserved when the detection for the true category overlaps the ground-truth box of the location source more than that of the other stimulus. A missing detection counts as a failure.

References. The floor draws a category and a box independently from the ground-truth categories and target boxes of the 1,000 cells (seed 42). The GT image row gives the GIT tokens of the stimulus to our read-outs.

Brain Algebra. For each category, the concept pattern is the mean z-scored training response of the 100 training images that CLIP ViT-B/32 assigns most strongly to it, keeping the half of the voxels with the largest magnitude. It is scaled by the voxel standard deviation, added with $\alpha = 4$ to the location-source response, min-max scaled as our encoder expects, and decoded by our model. The category is the one our read-out decodes from the category source.

Blend and MindEye1. The blend averages the two latents of a pair and decodes one render, which is scored as both cells, so its identity and joint cannot exceed 0.5. For MindEye1 (Scotti et al., 2023), we retrain our category head and box adapter on the 257×768 output of its subject-1 model (Table 3).

Table 3: **Swapping the encoder, subject 1.** Our read-outs on MindEye1's single-subject representation, on the same 1,000 cells as Table 2. Category and box describe the constituents handed to the generator.

encoder	category	box	identity	location	joint
MindEye1	0.544	0.326	0.771	0.452	0.395
InsideOut	**0.625**	**0.466**	**0.836**	**0.641**	**0.581**

D Composition Across Four Subjects

For each subject we draw 100 pairs of test stimuli (200 cross-response cells) whose objects differ in category, cover at most half of the image and overlap with an IoU of at most 0.3 (seed 42). This pool has no minimum area or spacing rule, so its scores are not directly comparable to Table 2. Subjects 1, 5 and 7 share the same pairs, and those of subject 2 differ because one of its outputs yields no box. Each subject uses its own model. On subject 1, scoring each render against the targets of another cell gives a null joint of 0.187 against the observed 0.490 ($p = 0.001$, 1,000 permutations).

Table 4: **Composition in four subjects.** 200 cells per subject.

subject	identity	location	joint	floor	ceiling
1	0.785	0.550	0.490	0.165	0.875
2	0.730	0.530	0.465	0.210	0.915
5	0.740	0.510	0.450	0.180	0.870
7	0.715	0.525	0.470	0.135	0.875
mean	0.743	0.529	0.469	0.173	0.884

E Decoding Results of Each Subject

Table 5 lists the four models behind the † row of Table 1.

Table 5: **Decoding results of each subject.** The rows behind the † row of Table 1, on the same 969 caption images and 959 category and box images. One output of subject 2 yields no box and scores an IoU of zero.

	caption										category	box
subject	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr	SPICE	CLIP-S	RefCLIP-S	top-1	IoU
1	61.94	44.11	30.92	21.80	20.84	46.75	68.24	13.39	65.90	72.11	0.612	0.442
2	59.78	41.49	28.31	19.36	19.63	45.40	59.72	12.28	64.43	70.68	0.591	0.455
5	62.52	44.03	30.60	21.40	20.88	46.71	68.28	13.79	66.95	72.94	0.630	0.412
7	60.05	41.55	28.49	19.70	19.40	45.26	59.43	12.25	63.75	70.12	0.547	0.398
mean	61.07	42.80	29.58	20.56	20.19	46.03	63.92	12.93	65.25	71.46	0.595	0.427

F Additional Results

F.1 Searchlight Maps of the Other Subjects

The mixed coloring of Figure 1(a) holds in every subject (Figure 6).

F.2 All Cortical Ablations

Table 6 adds the scene group omitted from Figure 5. Removing it changes little, and keeping it alone gives the lowest box IoU.

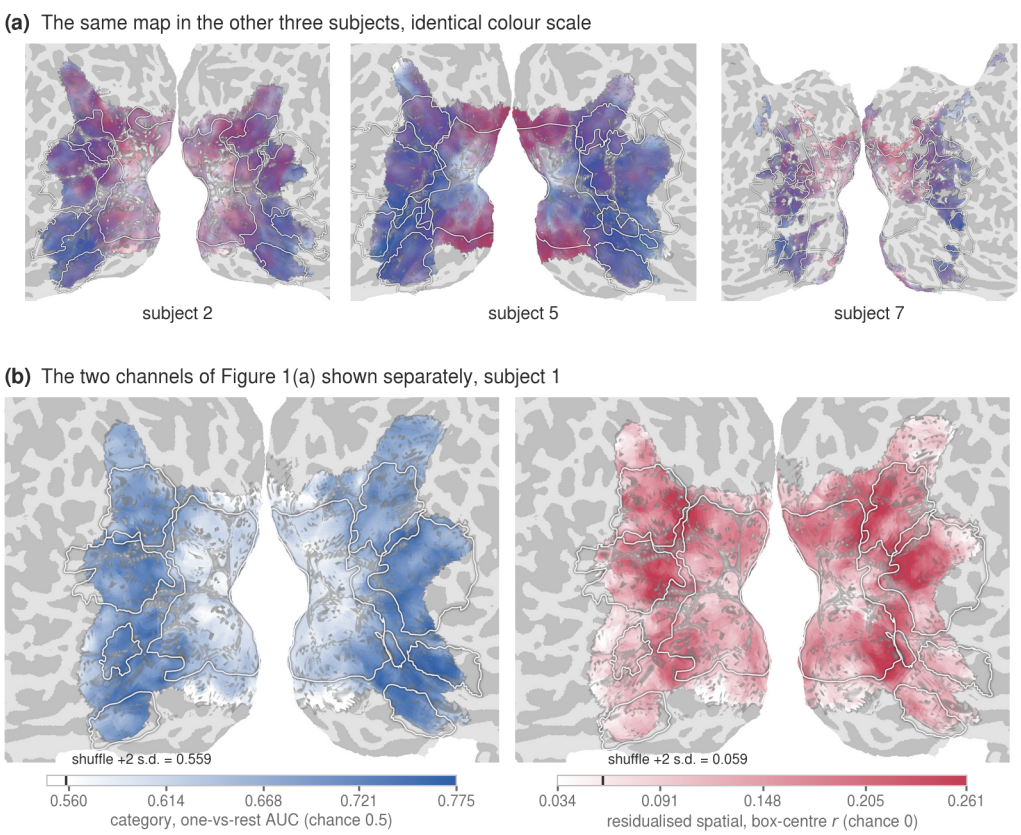


Figure 6: **Searchlight maps.** **(a)** Subjects 2, 5 and 7 on the color scale of Figure 1(a). **(b)** The category and location scores of subject 1 shown separately. Ticks on the color bars mark the mean plus two standard deviations of a label-shuffled null and are for display only.

Table 6: **Cortical ablations, subject 1, 981 test images.** The whole brain (15,724 voxels) gives a category top-1 of 0.590 and a box IoU of 0.447, and always answering person gives a top-1 of 0.290. Scene is OPA and PPA.

		OUT		only	
region	voxels	category	box	category	box
early visual	4,657	0.538	0.385	0.124	0.361
face and body	3,557	0.412	0.414	0.410	0.348
scene	2,252	0.553	0.446	0.162	0.286

F.3 Whether the Two Read-Outs Point to the Same Object

The `where` instruction names no category, and 76 percent of the test images contain more than one object, so the box may land on an object other than the decoded one. We match each decoded box of subject 1 to the COCO instance it overlaps most, when the IoU is at least 0.5. The box lands on the salient object in 44.1 percent of the 981 images, on another object in 9.6 percent and on no single object in 46.3 percent. When the category is decoded correctly, the box lands on an object of another category in only 3.8 percent of the images.

F.4 Correspondence with Real Scenes

Figure 4(b) shows one composition next to a real scene. Over all 1,000 cells, we ask whether real test scenes that hold the composed category at the composed box (IoU at least 0.5) are closer to the

composition than scenes that differ in location (same category, IoU at most 0.1), in category (another object at the box) or in both. Closeness is the correlation between SwAV features in image space, and in brain space between the measured response to each scene and the response predicted for the composition by a ridge model ($\alpha = 10^4$) from SwAV features to voxels, fitted on the 9,000 training images. The two sources are excluded. A null that pairs compositions with the scenes of other cells gives a location AUC of 0.534 in brain space ($p < 0.005$, 200 permutations).

Table 7: **Correspondence with real scenes, subject 1.** AUC of scenes matching the composition against scenes that differ, averaged over the cells where both sets exist (chance 0.5).

differs in	cells	brain space	image space
location	688	0.652	0.662
category	713	0.768	0.821
both	715	0.778	0.776

F.5 More Single-Response Decoding

Figure 7 shows more examples in the format of Figure 4(a).



Figure 7: **Additional single-response decoding.** Subject 5, except the second and last columns (subject 2). Top, test stimuli. Bottom, reconstructions as in Figure 4(a), with the ground-truth caption (GT) and the output of the caption read-out (Brain). Examples are chosen for display.

F.6 More Compositions

Figure 8 shows ten pairs, including the six of Figure 3.

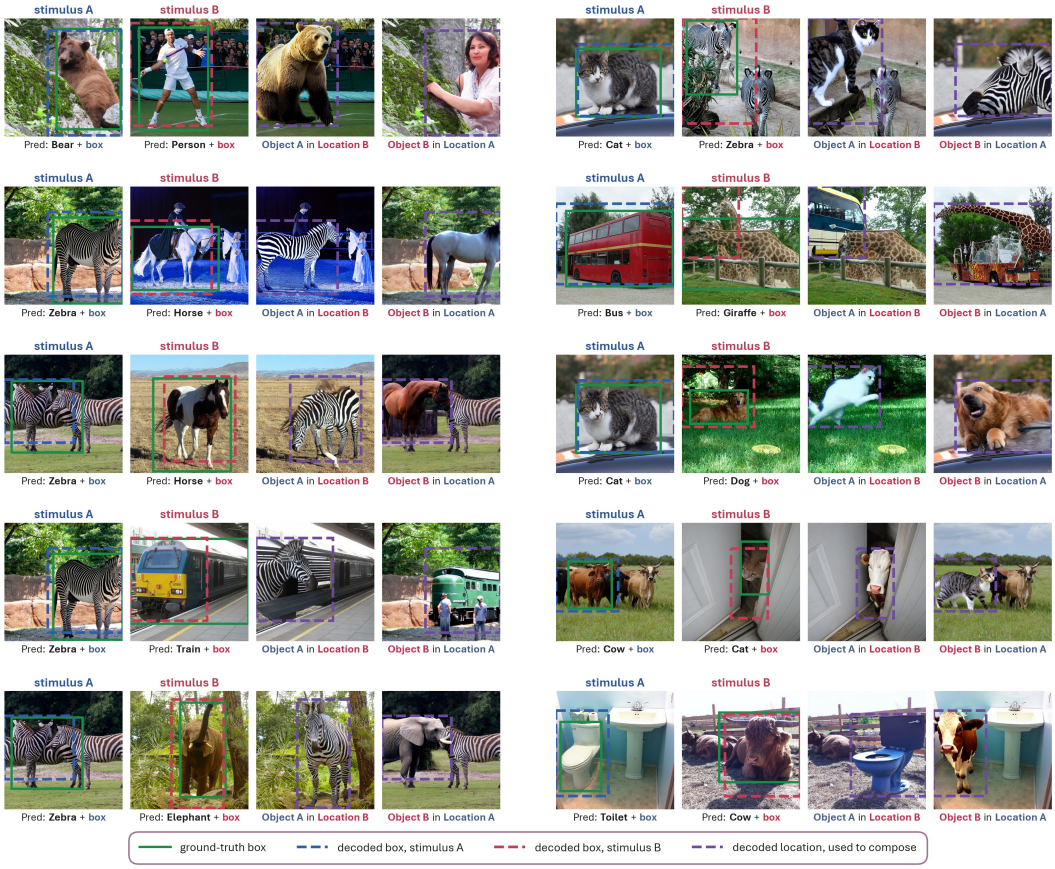


Figure 8: **More cross-response compositions, subject 1.** Notation as in Figure 3. For visualization only, each composition is inpainted into the photograph of its location source, so the pose and lighting of the drawn object follow the photograph and are not decoded. The pairs are chosen for display, and every cell shows the same one of five samples.