# BRAINEXPLORE: LARGE-SCALE DISCOVERY OF INTERPRETABLE VISUAL REPRESENTATIONS IN THE HUMAN BRAIN

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Understanding how the human brain represents visual concepts, and in which brain regions these representations are encoded, remains a long-standing challenge. Decades of work have advanced our understanding of visual representations, yet brain signals remain large and complex, and the space of possible visual concepts is vast. As a result, most studies remain small-scale, rely on manual inspection, focus on specific regions and concepts, and rarely include systematic validation. We present a large-scale, automated framework for discovering and explaining visual representations across the human cortex. Our method comprises two main stages. First, we *discover* candidate interpretable patterns in fMRI activity through unsupervised, data-driven decomposition methods. Next, we *explain* each pattern by identifying the set of images that most strongly elicit it and generating a natural-language description of their shared visual meaning. To scale this process, we introduce an automated pipeline that tests multiple candidate explanations, assigns reliability scores, and selects the best description for each voxel pattern. Our framework reveals thousands of interpretable patterns spanning many visual concepts, including fine-grained representations beyond well-established categories.

## 1 INTRODUCTION

Understanding how the human brain represents visual information is a long-standing challenge. The visual cortex encodes a rich hierarchy of features that support object recognition, scene understanding, and visual reasoning. Yet, the structure of these representations and their underlying organization in the brain remain largely unknown. Functional Magnetic Resonance Imaging (fMRI) has become a dominant tool for studying how the human brain processes visual information, providing a non-invasive window into cortical activity while participants view natural images (Allen et al., 2022; Horikawa & Kamitani, 2017). It measures brain activity across the brain, parceled into tiny volume elements ("voxels"). Over the past decades, fMRI researchers have sought to interpret these complex voxel-level patterns to uncover the structure of visual representations in the brain. Early work targeted retinotopic maps and low-level features (Sereno et al., 1995; DeYoe et al., 1996) , while later studies examined higher-level semantics using category-selective analyses (Downing et al., 2001; Kanwisher & Yovel, 2006) , data-driven fMRI decompositions (Huth et al., 2012; Khosla et al., 2022) and image–brain models (Çukur et al., 2013; Hwang et al., 2025; Luo et al., 2023; Gu et al., 2022).

However, fMRI signals are high-dimensional, comprising tens of thousands of brain voxels per subject that reflect a vast range of possible visual concepts, while available image-to-fMRI datasets remain relatively small. As a result, current studies often focus on specific properties (e.g., faces, places, food), or on particular brain regions, such as face-selective (FFA) or place-selective (PPA). Moreover, analyses typically rely on manual inspection, making it difficult to scale interpretations across many fMRI patterns and visual concepts. These limitations highlight the need for scalable, automated tools capable of discovering and explaining visual representations across the entire cortex.

Recent advances in interpretability of artificial neural networks, such as Language and Vision models, have shown that high-level concepts can be *automatically* discovered from internal activations (Bau et al., 2017; Hernandez et al., 2021; Gandelsman et al., 2023). These results suggest that similar data-driven approaches might offer new insight into the representations that emerge in the brain. Inspired by this, we introduce *BrainExplore*, an unsupervised, data-driven framework for large-scale



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

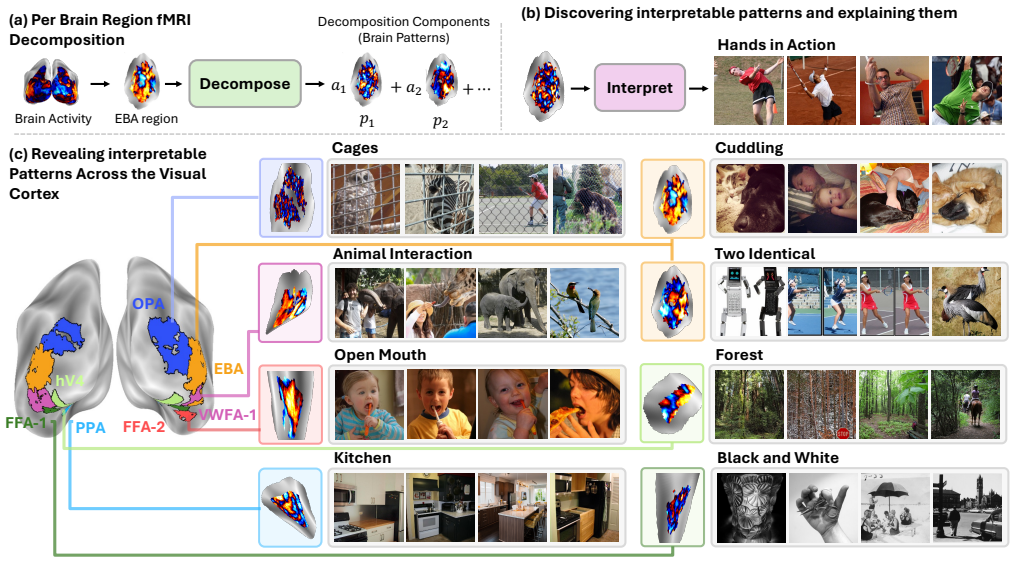


Figure 1: **BrainExplore: Discovering Interpretable Visual Representations in the Human Brain.** **(a)** Per–region fMRI decomposition learns component patterns such that any response is approximated as a linear combination of these patterns. **(b)** For each pattern, we retrieve its top-$N$ activating fMRI responses and corresponding images, then automatically select the best matching concept and assign an alignment score. We then use these scores to identify the most interpretable patterns and their explanations. **(c)** Examples of discovered interpretable patterns across regions, showing pattern activations projected onto cortex, top activating images, and their textual explanations.

discovery and evaluation of visual concept representations across the visual cortex. BrainExplore automatically discovers interpretable patterns of brain activity and explains them in natural language.

Our pipeline comprises two main steps: **(i)** deriving candidate interpretable fMRI patterns via per–brain-region decomposition (Fig. 1a); and **(ii)** discovering interpretable patterns and explaining them (Fig. 1b). The first step performs per brain region decomposition of fMRI activity to extract components that may correspond to meaningful visual representations. Conceptually, each decomposition method learns a set of fMRI patterns (components) such that any fMRI response in a given region of interest ("ROI") can be approximated as a linear combination of these patterns. For each component, we retrieve the top-$N$ fMRI responses with the highest coefficient for that component and collect their corresponding images. This yields a set of images per component that *visualize* the visual concept that most strongly activates it. We then discover interpretable patterns by automatically identifying, for each component, the concept that best aligns with it and assigning an alignment score. This allows us to surface the most interpretable components together with their explanations.

Using *BrainExplore*, we discover thousands of interpretable patterns throughout the brain. Since the method automatically evaluates the quality of each explanation, it enables integration of findings across different decomposition methods rather than relying on a single one. We exploit established techniques (PCA (Hotelling, 1933), NMF (Lee & Seung, 1999), ICA (Comon, 1994)) as well as Sparse Autoencoders (Cunningham et al., 2023) (SAEs), a method widely used to interpret artificial neural networks via projection into high-dimensional space and sparsity constraints. SAE yields a large number of interpretable patterns and reveals visual representations that are complementary to other methods, including representations not captured elsewhere. Moreover, to overcome limited data, we leverage an image-to-fMRI prediction model (Beliy et al., 2024) that synthesizes brain responses for unseen images. This allows us to expand the dataset from approximately 10k images with measured fMRI to over 120k images with measured or predicted responses. This augmentation improves decomposition quality and increases the diversity of discovered representations.

Altogether, *BrainExplore* reveals a large set of nuanced representations across the brain. These include interpretable patterns selective for object identity; for people, body parts, and pose (e.g., open mouths, hands holding objects, bent knees, identical objects); as well as distinct indoor and outdoor scenes (e.g., nature, streets, oceans, rooms, toilets; see Fig. 1). All findings are quantitatively evaluated on measured fMRI not used during interpretation.

**Our contributions are as follows:**

- We propose *BrainExplore*, a large-scale automated framework that discovers thousands of interpretable fMRI patterns, including concepts beyond well-established categories.
- We enrich fMRI decomposition with predicted fMRI, which substantially improves interpretability.
- We propose an SAE fMRI decomposition that finds interpretable patterns beyond standard methods.
- We will release the brain-inspired concept dictionary, the large-scale dataset of image–fMRI-explanations ranking, and the code, providing a benchmark for future work.

## 2 Related Work

**Contrasting predefined stimulus categories.** A large body of work has studied category-selective activations by contrasting responses to hand-picked stimulus classes. For example, researchers have mapped the extrastriate body area (EBA) and related regions using body-selective contrasts (Weiner & Grill-Spector, 2011; Downing et al., 2001), and characterized scene-selective regions such as the parahippocampal place area (PPA) (Epstein & Baker, 2019; Park et al., 2011). Other studies have identified face-selective regions including the fusiform face area (FFA) (Kanwisher & Yovel, 2006), examined overlapping but separable responses to faces and food in fusiform cortex (Adamson & Troiani, 2018; Jain et al., 2023), or investigated object-related properties (Konkle & Oliva, 2010). These approaches have been highly informative, but they are limited by the need to predefine a small set of categories, to use relatively "clean" images dominated by a single concept, and to test each hypothesis separately. Moreover, the same voxels and regions can participate in many concepts, making it difficult to discover nuanced or overlapping representations using only category contrasts.

**Decomposing fMRI signals.** Unsupervised decomposition methods provide an alternative route to fMRI interpretability. Commonly used techniques include Principal Component Analysis (PCA) (Hotelling, 1933), Independent Component Analysis (ICA) (Comon, 1994), and Non-negative Matrix Factorization (NMF) (Lee & Seung, 1999). These have been widely applied to fMRI research such as resting-state analysis and connectivity (Beckmann & Smith, 2004; Smith et al., 2009). Stimulus-driven decompositions for interpretability have also been explored, in domains such as speech and audio (Zeng & Gallant, 2025) and visual stimuli (Ecker et al., 2007; Huth et al., 2012). For visual concept discovery, early work used PCA to capture large-scale semantic maps (Huth et al., 2012) , while later studies leveraged NMF to obtain more interpretable components for categories such as bodies and food (van Dyck et al., 2024; Khosla et al., 2022; Zhao et al., 2024). ICA has seen more limited use, mainly for demonstrating retinotopic organization and low-level visual mapping (van de Ven et al., 2009). Most of these studies focus on specific regions (e.g., early visual cortex, EBA, PPA) or on a small set of concepts (e.g., food, social vs. non-social images), and typically examine a single decomposition method and only the first few components via manual inspection. We systematically compare multiple decomposition methods for brain interpretability, and scale analysis to tens of thousands of components across many regions and methods. We also leverage image-to-fMRI models to greatly expand the effective training set and image pool used for decomposition, and we introduce SAE (Cunningham et al., 2023) for fMRI decomposition.

**Image–fMRI models.** Recent years have brought substantial progress in models that link images and brain activity. This includes encoding models that predict fMRI responses from images (Kay et al., 2008; Beliy et al., 2024) , transformations between subjects or brains (Wasserman et al., 2024; Yamada et al., 2015), and image reconstruction from fMRI (Beliy et al., 2025; Scotti et al., 2024). Beyond pure prediction, several works have used such models to explore cortical representations. Early studies related voxel activations to predefined semantic concepts, for example in face- and place-selective regions (Çukur et al., 2013; 2016). More recent work has learned functional voxel clusters and visualized the kinds of images each cluster corresponds to (Beliy et al., 2024). Another line of work uses image-to-fMRI encoders together with generative image models to synthesize images that strongly activate particular regions or voxels (Hwang et al., 2025; Luo et al., 2023; Gu et al., 2022). However, these approaches still operate at the level of individual voxels or entire regions: regions are too coarse to capture fine-grained mixtures of concepts, while voxels are too local and can participate in many concepts, making nuanced representations difficult to disentangle. Our approach is complementary. We use image-to-fMRI models primarily as a tool for data augmentation and interpretability at the *pattern* level: predicted responses for large image sets greatly expand the pool of images used to learn decompositions and to retrieve maximally activating examples for each component, enabling more robust and fine-grained hypothesis testing.

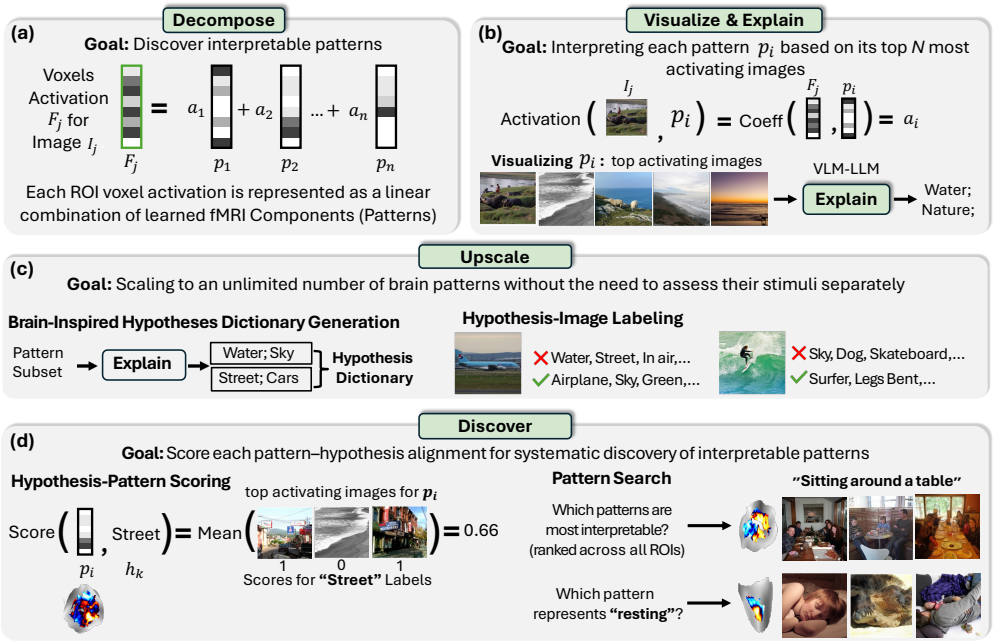


Figure 2: **BrainExplore Framework. (a)** ***Decompose:*** per-region fMRI decomposition to discover interpretable patterns (Sec. 3.2). **(b)** ***Visualize & Explain:*** retrieve the top activating images for each pattern and interpret its semantics (Sec. 3.3). **(c)** ***Upscale:*** scale to an unlimited number of patterns by building a brain-inspired dictionary and labeling each image with respect to each hypothesis (Sec. 3.4). **(d)** ***Discover:*** score each pattern–hypothesis alignment, enabling systematic discovery of the most interpretable patterns and the pattern that best explains any given hypothesis (Sec. 3.5).

**Automated interpretability.** Interpretability of neural networks advanced rapidly in recent years (Bau et al., 2017; Hernandez et al., 2021; Gandelsman et al., 2023; 2024). Recognizing that manual inspection does not scale to modern models, several works have proposed automated interpretability pipelines (Ghorbani et al., 2019; Oikarinen & Weng, 2022; Schwettmann et al., 2023; Shaham et al., 2024). The human brain shares many of the same challenges: it is a large, complex system with many units (voxels) and a vast repertoire of concepts. However, as far as we know, no large-scale, automatic interpretability pipeline has been proposed for the brain. Our work bridges these directions. We combine unsupervised fMRI decomposition with an automatic interpretability pipeline. Together with data augmentation from image-to-fMRI models and the use of SAEs, this yields a scalable framework that discovers many interpretable brain patterns and provides a benchmark for evaluating and improving future decomposition methods.

## 3 Methods

We first describe the experimental setting used in our framework (Sec. 3.1). We then present ***BrainExplore***, our brain interpretability framework with four parts (see Fig. 2): (i) ***Decompose***: per-region fMRI decomposition into pattern components (Sec. 3.2); (ii) ***Visualize & Explain***: retrieve the top activating images for each pattern and provide an interpretation of its semantics (Sec. 3.3); (iii) ***Upscale***: scale to an unlimited number of brain patterns without interpreting each stimulus separately (Sec. 3.4); (iv) ***Discover***: score each pattern–hypothesis alignment for systematic discovery of interpretable patterns (Sec. 3.5).

### 3.1 Experimental Setting

We use the Natural Scenes Dataset (NSD) (Allen et al., 2022), a large publicly available 7-Tesla fMRI dataset that records responses from 8 subjects viewing diverse natural images drawn from COCO (Lin et al., 2014). The dataset contains ∼73k image–fMRI pairs in total, with around 10k images per subject (some images are shared across subjects). We adopt the post-processed version of NSD released by Gifford et al. (2023), which includes ∼40k voxels per subject together with a predefined division into mainly vision-related ROIs. These ROIs are used for our per-ROI decomposition and subsequent analyses. To enrich the training data, we additionally use predicted fMRI responses for images not viewed by any subject. Specifically, we take ∼120K additional natural images from

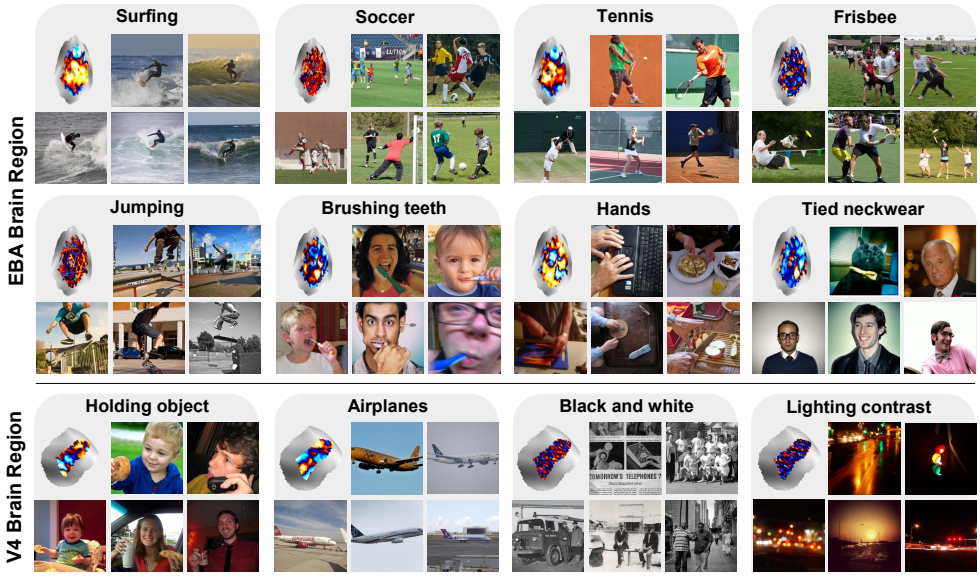


Figure 3: **Discovered Interpretable Patterns (EBA and V4).** We show patterns of subject 1 with their top activating images and selected explanations. EBA is known to process bodies and actions; V4 is known to encode mid-level features (e.g., color, shape).

the unlabeled portion of COCO and employ the image-to-fMRI encoder of Beliy et al. (2024) to predict fMRI responses for each subject. This augmentation produces a substantially larger set of image–fMRI pairs, which we use both for training decompositions and for retrieving maximally activating images. Importantly, all interpretations are evaluated and verified on measured fMRI.

### 3.2 fMRI Decomposing

The first step in our pipeline performs per-ROI decomposition of fMRI activity to extract component patterns that may correspond to meaningful visual representations. Conceptually, each decomposition method learns a set of fMRI patterns (components) such that any fMRI response in a given ROI can be approximated as a linear combination of them (Fig. 2a). Each trained decomposition yields two outputs: *(i) a component matrix*, where each column is a learned fMRI pattern (for PCA, NMF, and ICA this is the standard loading matrix; for SAEs it corresponds to the decoder weights); and *(ii) a coefficient matrix*, which contains the linear coefficients reconstructing each fMRI response from these patterns. Importantly, all decompositions are learned purely from fMRI responses; no image features or labels are used, ensuring that all inferred visual representations arise from brain activity rather than externally imposing image semantics.

We train three standard decomposition methods (PCA, NMF, and ICA) as well as SAEs. For each method, we train two variants: (i) using only measured fMRI responses ($\sim$10k responses per subject), and (ii) using the combined dataset of measured fMRI and 120k predicted responses. For each method, we adopt default parameter settings and vary a few hyperparameters (e.g., variance thresholds for PCA/NMF/ICA and different seeds; sparsity settings and expansion factors for SAE). Each decomposition produces a large pool of candidate patterns per ROI, later used in the *Discover* step. As a baseline, we also treat individual voxel activations as a single-component "decomposition", effectively serving as a one-hot voxel basis. Training details are provided in Appendix G.1.

### 3.3 From fMRI Patterns to Visual Concepts

To interpret each decomposed brain pattern, we first connect it to the set of stimuli that drive it (Fig. 2b). Each fMRI response is directly associated with a stimulus — the natural image viewed by the subject. After decomposition, every fMRI response is represented by a vector of coefficients, one per pattern component, indicating how strongly each pattern contributes to its reconstruction. We assume that the responses with the highest coefficients for a given pattern are the most informative about its underlying visual semantics. Thus, for every pattern, we select its top-$N$ activating responses and collect their corresponding images, producing a set of images that *visualize* what most strongly activates that pattern. We perform this separately for measured-fMRI responses and predicted responses. To *explain* each pattern, we take the six most activating images from the measured-fMRI

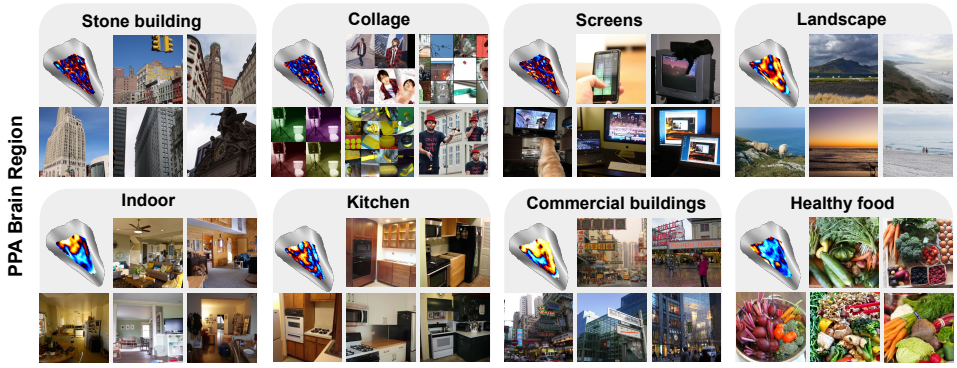


Figure 4: **Discovered Interpretable Patterns (PPA).** We show Subject 1 patterns with top activating images and selected explanations. PPA encodes scenes and places (e.g., indoor/outdoor, landmarks).

pool and the ten most activating images from the predicted-fMRI pool (16 images total). We generate captions for these images using a Vision–Language Model (VLM) and then prompt an LLM to synthesize 5–10 hypotheses capturing what is shared across them (e.g., object identity, pose, texture, scene category). These hypotheses serve as candidate semantic explanations for the pattern.

### 3.4 Scaling to an Unlimited Number of Patterns

Explaining every component in every ROI and decomposition method is costly and inefficient, especially with high-dimensional SAEs, multiple hyperparameters and methods. To scale interpretability, we introduce two steps (Fig. 2c): (i) ***Hypothesis dictionary generation***, which constructs a brain-activity-inspired set of candidate hypotheses; and (ii) ***Hypothesis–image labeling***, which assigns binary labels for each hypothesis and image. Together, these two steps create large-scale associations between visual stimuli and candidate descriptions of brain activity, independent of any specific pattern. Once the dictionary and image labels are in place, any new brain pattern can be evaluated by measuring how consistently its top-activating images express concepts from the dictionary.

**Hypothesis dictionary generation.** We first identify patterns most likely to be interpretable. To do so, we compute a CLIP-based *consistency* score for every pattern: we embed its top activating images, compute pairwise similarities, and use their mean as a proxy to semantic coherence. We then select the top 40 patterns per ROI and decomposition method and run the *Visualize & Explain* step on this subset ($>$ 10k patterns), yielding more than 1k unique explanations. These are aggregated into a brain-inspired dictionary of visual concepts. To remove duplicates we embed each hypothesis with a BGE text encoder (Chen et al., 2024) and merge concepts with high embedding similarity. This produces a dictionary of $\sim$1,300 concepts that we can use to label all stimuli offline with respect to each candidate concept, creating large-scale associations between images and candidate brain activity descriptions, independent of any specific pattern. Further details and prompts appear in Appendix G.

**Hypothesis–image labeling.** Our goal is to obtain, for every image in the dataset (both for images with measured and predicted fMRI), a binary vector indicating which concepts from the dictionary apply to that image. Directly querying a VLM with all 1,300 hypotheses per image is infeasible and often leads to inaccurate results, therefore we use a two-stage procedure. First, we use CLIP to shortlist a small set of candidate hypotheses per image: We embed each hypothesis under multiple prompt templates, compute its similarity to the CLIP image embedding, and retain the top 300 hypotheses per image. Second, we perform VLM-based verification. For each shortlisted candidate hypothesis, we prompt a VLM to output a binary decision (0/1) indicating whether the hypothesis holds for a given image. To improve reliability, we run a second verification pass with a different prompt for all hypotheses initially marked as positive, and keep only those confirmed twice. This produces a sparse but high-quality binary image–hypothesis label matrix, used later for large-scale pattern–hypothesis scoring. Additional details and labeling examples are provided in Appendix G.

### 3.5 Discover Interpretable Brain Patterns

The *Discover* step (Fig. 2d) identifies the visual concept represented by each fMRI pattern. It uses *Upscale*'s hypothesis dictionary and image–hypothesis labels to evaluate any decomposition component pattern by checking whether the images that activate it express consistent visual concepts.

Table 1: **Average Number of Interpretable Hypotheses.** The number/fraction of hypotheses with an alignment score above a threshold (0.5 or 0.8) for at least one pattern, averaged across 8 subjects.

| | Measured fMRI | | + Predicted fMRI | |
|---|---|---|---|---|
| Method | $> 0.5$ | $> 0.8$ | $> 0.5$ | $> 0.8$ |
| Voxels | 42 / 3.17% | 4 / 0.34% | 86 / 6.50% | 15 / 1.19% |
| PCA | 20 / 1.53% | 4 / 0.32% | 78 / 5.90% | 11 / 0.84% |
| NMF | 19 / 1.45% | 3 / 0.23% | 61 / 4.62% | 10 / 0.78% |
| ICA | 11 / 0.88% | 2 / 0.22% | 207 / 15.56% | 34 / 2.62% |
| SAE | **104 / 7.87%** | 18 / 1.33% | 219 / 16.46% | 51 / 3.80% |
| SAE+ICA | 105 / 7.86% | **18 / 1.34%** | **250 / 18.81%** | **58 / 4.36%** |

**Hypothesis–pattern scoring.** For each pattern, we quantify how strongly it aligns with each candidate hypothesis. Every fMRI response has a coefficient indicating how much a given pattern contributes to reconstructing that response. We rank all images by this coefficient and select the top $0.2\%$ most activating images for that pattern, excluding responses that do not meaningfully activate it (SAE coefficients $< 0.01$; non-SAE coefficients $\leq 0$). Since each image has binary labels for all hypotheses, we can directly measure how well a hypothesis matches the pattern. For a given pattern and hypothesis $h$, we compute its score as $\frac{\#\{\text{top-activating images labeled with } h\}}{N}$, where $N$ is the total number of selected top-activating images. This measures how consistently concept $h$ appears among the images that drive the pattern. To account for globally rare hypotheses, we normalize the score by the hypothesis's dataset frequency. Scores are computed independently for the measured and predicted fMRI pools, and the final hypothesis-pattern score is defined as their average.

**Pattern search.** Given the pattern–hypothesis scores, we support two complementary searches: (i) *pattern search*: find the most interpretable pattern(s); and (ii) *hypothesis search*: for a given hypothesis (e.g., "open mouth"), find the best-explained pattern. Both searches can be run *per ROI* (best pattern within an ROI) or *across ROIs* (best pattern over all ROIs), enabling region-specific (*if/where* a hypothesis is represented) and method-level (*which* decompositions capture it) analyses.

# 4 Experiments

## 4.1 Evaluation Protocol

To ensure fair evaluation, we split each pool into disjoint *ranking* and *evaluation* halves: the ranking set is used to assign best pattern-hypothesis pairs; the evaluation set is used to report final metrics at Sec. 4. In the main paper, we present results for Subject 1 from the NSD dataset. Quantitative and qualitative results for the other subjects are provided in Appendices H and I, demonstrating the effectiveness of our pipeline across subjects. Since every pattern receives a score, searches can be conducted not only within a single decomposition, but also *across* methods and hyperparameters.

## 4.2 Quantitative Evaluation

We report two metrics: (i) ***Number of Interpretable Hypotheses*** (Table 1): the number/fraction of hypotheses that achieve an alignment score above a specified threshold (0.5 or 0.8) with at least one pattern. This can be computed either per ROI or across all ROIs. (ii) ***Number of Interpretable Patterns*** (Appendix Table T15): the number of patterns whose best hypothesis alignment exceeds the same threshold (0.5 or 0.8). This table shows that SAE yields far more interpretable patterns than other methods: approximately 9k patterns across the brain at a 0.5 threshold, compared to $\sim$6k for the Voxels baseline and 226 for ICA.

**Predicted fMRI substantially improves interpretability.** Table 1 compares decomposition methods trained on measured fMRI alone versus augmented with predicted fMRI, used both to learn the decomposition and to retrieve top-activating images. Predicted fMRI yields large gains across methods: ICA increases from an average of around 12 to 207 interpretable hypotheses at the 0.5 threshold, and SAE increases from around 105 to 218.

**SAE provides a large number of interpretable hypotheses.** Overall, SAE yields the highest number of interpretable hypotheses among individual methods.

**Combining decompositions yields the best performance.** Integrating across methods (ICA+SAE) yields the strongest performance. This is enabled by our automatic scoring, which evaluates any number of pattern–hypothesis pairs and supports aggregation across decompositions.

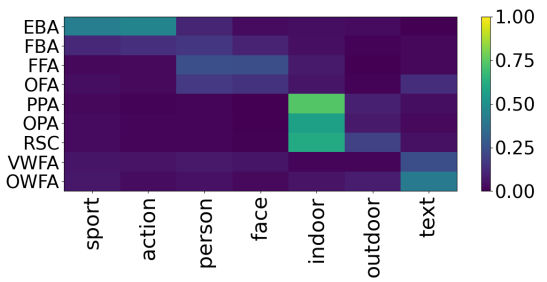


Figure 5: **ROI–concept alignment.** Applying our pipeline to average ROI activation shows scores consistent with the known functional selectivity of visual regions. For example, FFA/OFA align with faces and place regions (e.g. PPA) with indoor/outdoor scenes.

| ROI | $k = 2$ | $k = 4$ | $k = 8$ |
|---|---|---|---|
| EBA | 0.84* | 0.72* | 0.61* |
| FBA | 0.76 | 0.60* | 0.48* |
| FFA | 0.73 | 0.55 | 0.42* |
| OFA | 0.69 | 0.50 | 0.38 |
| PPA | 0.75 | 0.58* | 0.44* |
| OPA | 0.80* | 0.66* | 0.55* |
| RSC | 0.67 | 0.47 | 0.35 |
| VWFA | 0.73 | 0.55 | 0.41 |
| OWFA | 0.70 | 0.51 | 0.38* |

Table 2: **Concept consistency across subjects.** We collected the interpretable concepts discovered in each ROI and measured their overlap across subject subsets of size $k$, compared to a null baseline. Asterisks indicate $p < 0.05$.

### 4.3 Alignment with Established Neuroscientific Knowledge

Our method reveals a substantially larger number of interpretable patterns and visual concepts than prior work. These include previously reported findings, concepts aligned with known functionality of brain regions, but to our knowledge not previously shown, and additional less expected concepts. We next validate our results by showing consistency with established neuroscientific knowledge.

**Alignment with known category selectivity in functional regions.** We conducted two experiments to validate alignment between our discovered concepts and established functional selectivity of visual regions. (i) **ROI–concept alignment:** We applied our scoring procedure at the ROI level, using the average activation of each ROI instead of individual voxels or decomposition components, to compute an ROI–concept alignment score. Because ROIs are polysemantic, we expect their dominant concepts to reflect their known functional role rather than high alignment scores. As shown in Fig. 5, the resulting alignments are consistent with established knowledge: for example, EBA aligns strongly with actions and sports (Weiner & Grill-Spector, 2011; Downing et al., 2001); FFA aligns with people and faces (Kanwisher & Yovel, 2006); PPA, OPA, and RSC align with locations (Epstein & Baker, 2019; Çukur et al., 2016); and OWFA and VWFA align with text-related concepts (Cohen et al., 2000; McCandliss et al., 2003). (ii) **LLM-based relevance labeling:** We further automatically labeled each concept–ROI pair according to whether it is generally consistent with the known functional role of that ROI. We then measured the fraction of relevant concepts found to be interpretable for each ROI. The results, shown in Fig. S1 and discussed in Appendix C, further support the validity of the discovered patterns.

**Consistency across subjects.** To evaluate this, for each ROI and subject, we collected the set of interpretable concepts discovered in that ROI and measured their overlap across subjects. In Table 2, we report the fraction of concepts in each ROI consistent across all subject pairs, and groups of 4 and 8 subjects. As expected, consistency is high but not perfect, reflecting inter-subject variability. We further analyze the cross-subject consistency of patterns in Appendix C.3 and Fig. S2.

### 4.4 Visual Representations in the Human Brain

**BrainExplore reveals fine-grained patterns across the brain.** For any hypothesis of interest, we can retrieve the best aligned pattern within a specific ROI or across ROIs, and visualize its representation via the top activating images. Figs. 1 and 3 show examples from the top 16 images per pattern (8 from measured, 8 from predicted fMRI; full grids appear in Appendix I). Although the predicted fMRI pool is much larger and often yields clearer visualizations, we validate each explanation on measured fMRI as well. Overall, BrainExplore uncovers many fine-grained patterns, including ones not previously identified. For example, in EBA, classically linked to body perception and action, we find patterns whose explanations consistently reference specific sports (surfing, soccer, tennis, frisbee), actions (jumping, tooth brushing), and body oriented concepts (hands, neckwear). In PPA, known to respond to scenes, we observe a more nuanced division than prior indoor-outdoor contrasts, including distinct outdoor concepts for landscapes, commercial buildings, and stone architecture. We further show that our method generalizes to an independent subject from the LAION-fMRI dataset in Appendix F.

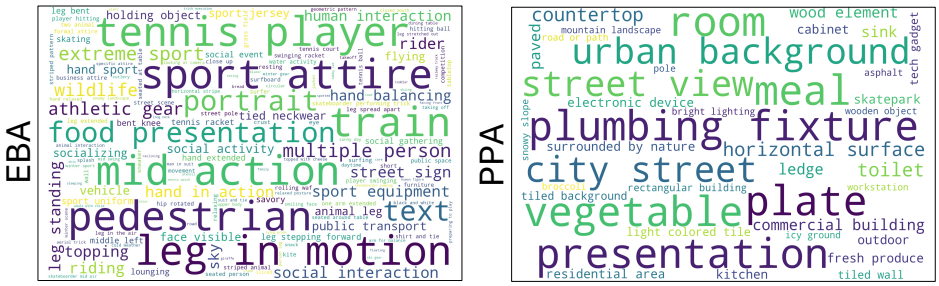


Figure 6: **Concepts best explained by each ROI (EBA, PPA).** Each concept is assigned to a single ROI—the one with the highest alignment score. Only concepts with alignment $> 0.5$ are shown; word size reflects the alignment score within the assigned ROI.

**Different ROIs represent different semantics.** Different ROIs are known to support distinct functions. Many of our discovered patterns align strongly with these known roles, which both validates our approach and lends credibility to novel findings. To examine this, we first find for each hypothesis, the best matching pattern across ROIs; then, for each ROI, we list the hypotheses it best explains. In Fig. 6, we show examples where word size reflects the hypothesis–pattern score in that ROI. As shown, our method recovers many concepts, with different ROIs best explaining different ones. We show additional concepts and ROIs (Appendix I), and joint decompositions of brain regions, which recover additional interpretable representations (Appendix D).

**Interpretable patterns are more localized with SAE.** The brain exhibits meaningful spatial organization at both regional and subregional scales. While low-level maps are well studied, higher-level semantic organization remains less clear. We find that SAEs yield interpretable patterns that are notably more spatially localized, despite receiving only 1D voxel vectors and no spatial information or spatial constraint. In Fig. 3, we show patterns from both ICA and SAE. In EBA, for example, *soccer*, *frisbee*, and *jumping* are derived using ICA, whereas the remaining patterns come from SAE. Additional side-by-side comparisons of pattern localization appear in Appendix I. Greater localization is both biologically plausible and of strong interest to the neuroscience community.

### 4.5 Ablations & Analysis

**Discovered representations depend on structured brain activity.** Voxel shuffling reduces the fraction of interpretable hypotheses on measured fMRI from 6.2% to 0.98% at the 0.5 threshold. Under randomly generated brain signals, no concepts remain significant after multiple-comparison correction, versus 99% for the original responses (Appendix B.7). Additional analyses in Appendix F provide two further validations: **validation of our image labels** through cross-VLM consistency and human annotations; and **robustness to the choice of image-to-fMRI encoder**, showing that most high-scoring concepts are recovered by three independent encoders. Appendix B analyzes additional aspects of our pipeline, including: (i) ***SAE ablations*** - the importance of sparsity, expansion factor and encoder architecture; (ii) ***Predicted fMRI pool size*** - results with varying image–fMRI pool sizes for retrieving top-activating images and their effect on interpretability; (iii) ***Per-ROI quantitative results*** - detailed metrics reported per ROI and (iv) ***Complementarity across decompositions*** - the gains from combining different methods and an analysis of concepts uniquely captured by each.

## 5 Discussion & Limitations

We introduced *BrainExplore*, a scalable framework for discovering visual concept representations across the visual cortex. By combining multiple decomposition methods within an automatic pipeline and enriching both training and retrieval with a large pool of predicted fMRI signals, BrainExplore reveals thousands of interpretable patterns capturing fine-grained visual concepts across regions. Despite the advantages of our large-scale pipeline, several limitations remain. Our approach uses VLM-based labeling, which can be noisy and may reduce the accuracy of the resulting scores. We mitigate this through a careful two-stage procedure, performing one-concept-at-a-time labeling followed by a second VLM verification step. Moreover, while our extensive evaluation supports the validity of the discovered patterns, definitive validation of newly identified representations will require future targeted neuroscience experiments. We hope that *BrainExplore* will help guide the design of such experiments. Overall, our work advances the large-scale study of visual representations in the brain and provides a practical tool for discovery. It uncovers multiple visual representations beyond well-established categories and can help guide future computational and experimental studies.

## 6 AI USE DISCLOSURE

Generative AI tools were used as components of the BrainExplore pipeline, as described throughout the paper. We additionally used generative AI tools to assist with code development and debugging, and to assist with writing, including polishing and improving the clarity and readability of the manuscript. All AI-assisted work was reviewed by the authors, who take responsibility for the final content of this work.

## REFERENCES

Kayleigh Adamson and Vanessa Troiani. Distinct and overlapping fusiform activation to faces and food. *NeuroImage*, 174:393–406, 2018. doi: 10.1016/j.neuroimage.2018.02.064.

Hossein Adeli, Minni Sun, and Nikolaus Kriegeskorte. Transformer brain encoders explain human high-level visual responses. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.

Christian F. Beckmann and Stephen M. Smith. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans. Med. Imaging*, 23(2):137–152, 2004. doi: 10.1109/TMI.2003.822821.

Roman Beliy, Navve Wasserman, Amit Zalcher, and Michal Irani. The wisdom of a crowd of brains: A universal brain encoder. *arXiv preprint arXiv:2406.12179*, 2024.

Roman Beliy, Amit Zalcher, Jonathan Kogman, Navve Wasserman, and Michal Irani. Brain-it: Image reconstruction from fmri via brain-interaction transformer. *arXiv preprint arXiv:2510.25976*, 2025.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024.

Laurent Cohen, Stanislas Dehaene, Lionel Naccache, Stéphane Lehéricy, Ghislaine Dehaene-Lambertz, Marie-Anne Hénaff, and François Michel. The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain*, 123(2):291–307, 2000.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

Tolga Çukur, Alexander G Huth, Shinji Nishimoto, and Jack L Gallant. Functional subdomains within human ffa. *Journal of Neuroscience*, 33(42):16748–16766, 2013.

Tolga Çukur, Alexander G Huth, Shinji Nishimoto, and Jack L Gallant. Functional subdomains within scene-selective cortex: parahippocampal place area, retrosplenial complex, and occipital place area. *Journal of Neuroscience*, 36(40):10257–10273, 2016.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL `https://arxiv.org/abs/2309.08600`.

E. A. DeYoe, G. J. Carman, P. Bandettini, S. Glickman, J. Wieser, R. Cox, D. Miller, and J. Neitz. Mapping striate and extrastriate visual areas in human cerebral cortex. *Proceedings of the National Academy of Sciences*, 93(6):2382–2386, 1996. doi: 10.1073/pnas.93.6.2382.

Paul E Downing, Yuhong Jiang, Miles Shuman, and Nancy Kanwisher. A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470–2473, 2001.

Christine Ecker, Emanuelle Reynaud, Steven C Williams, and Michael J Brammer. Detecting functional nodes in large-scale cortical networks with functional magnetic resonance imaging: A principal component analysis of the human visual system. *Human brain mapping*, 28(9):817–834, 2007.

Russell A Epstein and Chris I Baker. Scene perception in the human brain. *Annual review of vision science*, 5(1):373–397, 2019.

Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. *arXiv preprint arXiv:2310.05916*, 2023.

Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting the second-order effects of neurons in clip. *arXiv preprint arXiv:2406.04341*, 2024.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

Alessandro T Gifford, Benjamin Lahner, Sari Saba-Sadiya, Martina G Vilas, Alex Lascelles, Aude Oliva, Kendrick Kay, Gemma Roig, and Radoslaw M Cichy. The algonauts project 2023 challenge: How the human brain makes sense of natural scenes. *arXiv preprint arXiv:2301.03198*, 2023.

Zijin Gu, Keith Wakefield Jamison, Meenakshi Khosla, Emily J. Allen, Yihan Wu, Ghislain St-Yves, Thomas Naselaris, Kendrick Kay, Mert R. Sabuncu, and Amy Kuceyeski. Neurogen: Activation optimized image synthesis for discovery neuroscience. *NeuroImage*, 247:118812, 2022. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2021.118812. URL https://www.sciencedirect.com/science/article/pii/S1053811921010831.

Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep visual features. In *International Conference on Learning Representations*, 2021.

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

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Alexander G. Huth, Shinji Nishimoto, An T. Vu, and Jack L. Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012. doi: 10.1016/j.neuron.2012.10.014.

Ethan Hwang, Hossein Adeli, Wenxuan Guo, Andrew Luo, and Nikolaus Kriegeskorte. In silico mapping of visual categorical selectivity across the whole brain. *arXiv preprint arXiv:2510.21142*, 2025.

Nidhi Jain, Aria Wang, Margaret M. Henderson, Ruogu Lin, Jacob S. Prince, Michael J. Tarr, and Leila Wehbe. Selectivity for food in human ventral visual cortex. *Communications Biology*, 6(1):175, 2023. doi: 10.1038/s42003-023-04546-2. URL https://doi.org/10.1038/s42003-023-04546-2.

Nancy Kanwisher and Galit Yovel. The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361 (1476):2109–2128, 2006.

Kendrick N. Kay, Thomas Naselaris, Ryan J. Prenger, and Jack L. Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008. doi: 10.1038/nature06713.

Meenakshi Khosla, N Apurva Ratan Murty, and Nancy Kanwisher. A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition. *Current Biology*, 32(19):4159–4171, 2022.

Talia Konkle and Aude Oliva. Examining how the real-world size of objects is represented in ventral visual cortex. *Journal of Vision*, 10(7):982–982, 2010.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *nature*, 401(6755):788–791, 1999.

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.

Andrew Luo, Maggie Henderson, Leila Wehbe, and Michael Tarr. Brain diffusion for visual exploration: Cortical discovery using large scale generative models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 75740–75781. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ef0c0a23a1a8219c4fc381614664df3e-Paper-Conference.pdf.

Bruce D McCandliss, Laurent Cohen, and Stanislas Dehaene. The visual word form area: expertise for reading in the fusiform gyrus. *Trends in cognitive sciences*, 7(7):293–299, 2003.

Tuomas Oikarinen and Tsui-Wei Weng. Clip-dissect: Automatic description of neuron representations in deep vision networks. *arXiv preprint arXiv:2204.10965*, 2022.

Soojin Park, Timothy F Brady, Michelle R Greene, and Aude Oliva. Disentangling scene content from spatial boundary: complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. *Journal of Neuroscience*, 31(4):1333–1340, 2011.

Sarah Schwettmann, Tamar Shaham, Joanna Materzynska, Neil Chowdhury, Shuang Li, Jacob Andreas, David Bau, and Antonio Torralba. Find: A function description benchmark for evaluating interpretability methods. *Advances in Neural Information Processing Systems*, 36:75688–75715, 2023.

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

Martin I. Sereno, Anders M. Dale, John B. Reppas, Kai K. Kwong, John W. Belliveau, Thomas J. Brady, Bruce R. Rosen, and Roger B. H. Tootell. Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, 268(5212):889–893, 1995. doi: 10.1126/science.7754376.

Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. A multimodal automated interpretability agent. In *Forty-first International Conference on Machine Learning*, 2024.

Stephen M. Smith, Peter T. Fox, Karla L. Miller, David C. Glahn, P. Mickle Fox, Clare E. Mackay, Nicola Filippini, Kate E. Watkins, Roberto Toro, Angela R. Laird, and Christian F. Beckmann. Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences*, 106(31):13040–13045, 2009. doi: 10.1073/pnas.0905267106. URL https://www.pnas.org/doi/abs/10.1073/pnas.0905267106.

VG van de Ven, B Jans, M Been, R Goebel, and P de Weerd. Intrinsic functional organization of retinotopic visual fields in human occipital cortex: A 3t fmri study. *NeuroImage*, 47:S63, 2009.

Leonard van Dyck, Martin N Hebart, and Katharina Dobs. Core neural dimensions of functionally selective areas in the human visual cortex. In *European Conference on Visual Perception (ECVP)*, 2024.

Navve Wasserman, Roman Beliy, Roy Urbach, and Michal Irani. Functional brain-to-brain transformation with no shared data. *arXiv preprint arXiv:2404.11143*, 2024.

Kevin S Weiner and Kalanit Grill-Spector. Not one extrastriate body area: using anatomical landmarks, hmt+, and visual field maps to parcellate limb-selective activations in human lateral occipitotemporal cortex. *Neuroimage*, 56(4):2183–2199, 2011.

Kentaro Yamada, Yoichi Miyawaki, and Yukiyasu Kamitani. Inter-subject neural code converter for visual image representation. *NeuroImage*, 113:289–297, 2015.

Alicia Zeng and Jack L. Gallant. Disentangling superpositions: Interpretable brain encoding model with sparse concept atoms. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=3aNvX9TQTo.

Josefine Zerbe, Johannes Roth, Maggie Mae Mell, Peer Herholz, Tomas Knapen, and Martin N. Hebart. Laion-fmri: A densely sampled 7t-fmri dataset providing broad coverage of natural image diversity, 2026. URL https://www.visionsciences.org/talk-sessions?id=642. Talk 25.11, Scene Perception Talk Session, Vision Sciences Society Annual Meeting, May 16, 2026.

Yuanfang Zhao, Emalie McMahon, and Leyla Isik. Separate neural representations for physical and communicative social interactions along the lateral visual pathway: evidence from data-driven voxel decomposition. *Journal of Vision*, 2024. URL https://api.semanticscholar.org/CorpusID:272937445.

# Appendix

## A Extended Limitation Discussion

The interpretability of our discovered representations is inherently limited by the concept dictionary: a representation is considered interpretable as long as it can be described using concepts contained in the dictionary. Although our dictionary is large, it remains limited in several ways. First, visual representations that are difficult to express verbally, including some low-level visual features or abstract/non-concrete properties, may not be well captured. Second, both NSD and the external image sets used in our analysis are largely based on COCO images, restricting the space of concepts that can be evaluated to those represented in these datasets. Future datasets with broader image distributions may help alleviate this limitation. Third, the dictionary is constructed from a finite set of candidate patterns used in our method and therefore cannot cover the full space of representations potentially encoded in the brain. Expanding and refining this concept space is an important direction for future work.

Our dictionary is seeded from patterns discovered by BrainExplore itself, but is then held fixed: all reported scores, including for patterns that did not contribute to the dictionary, are computed against this fixed set on separate evaluation data. Two additional results support that these scores reflect structured brain activity. Our randomized-activity controls (Appendix B.7) show that applying the same fixed dictionary and scoring pipeline to voxel-shuffled or randomly generated signals yields almost no interpretable hypotheses (0.98% above the 0.5 threshold, versus 6.2% on real measured fMRI, with no concepts surviving a random-signal null after correction). And our held-out generalization analysis (Appendix B.6) shows that hypotheses assigned using only the training split remain predictive of held-out responses to entirely unseen images (median AUC 0.74–0.82). Together, these results indicate the pipeline captures reproducible structure in brain activity. Altogether, the current dictionary provides broad coverage and enables the identification of many interpretable representations, while our methodology, quantitative metrics, and analysis framework provide a basis for systematically extending this space in future work.

Our evaluation combines measured fMRI with external data and multiple image-to-fMRI encoders, and further examines generalization across subjects and datasets, providing evaluation at substantially greater scale and across multiple robustness settings. Nevertheless, further experimental validation is required to more directly establish the reliability and specificity of individual discovered representations. In particular, because measured fMRI data remain limited, the results obtained from measured data alone are already encouraging, while larger and more diverse datasets would likely enable even stronger validation of our findings. We therefore view BrainExplore as a method for generating concrete and quantitatively supported hypotheses about neural representations, which can subsequently be tested through targeted follow-up experiments.

# B Ablations and Additional Results

## B.1 SAE ablations

The two main hyperparameters in our SAEs are (i) the latent-space dimensionality, controlled by the expansion factor, and (ii) the sparsity coefficient, which weights the sparsity regularization on the latent codes. Sparsity encourages each fMRI sample to be reconstructed using only a small number of patterns (i.e., a sparse linear combination of decoder columns). We evaluate four expansion factors: 0.5, 1, 2, and 4. The settings 0.5 and 1 correspond to "regular" autoencoders that reduce or roughly preserve the dimensionality, while 2 and 4 yield overcomplete representations that expand it. We also evaluate four sparsity coefficients: 0, 1, 2, and 4, where 0 corresponds to no sparsity penalty (a standard autoencoder).

In Table T1 we report, for each configuration, the rank score used for model selection (the evaluation results reported in the main paper are computed on a separate held-out set), together with the percentage of interpretable hypotheses above two thresholds (0.5 and 0.8). Sparsity has a clear and consistent effect: higher sparsity generally produces more interpretable patterns. This is expected, since enforcing that each fMRI sample is reconstructed from only a few patterns encourages those patterns to capture meaningful structure rather than overfitting. The expansion factor shows a similar trend: larger expansion factors tend to perform better, with 0.5 performing noticeably worse in most cases. Overall, both hyperparameters are important for SAE interpretability. Expansion factors larger than 4 did not yield consistent additional gains and were computationally expensive. Similarly, increasing the sparsity coefficient beyond 4 did not lead to clear improvements.

Table T1: **SAE Ablation by Expansion Factor and Sparsity.** Percentage of Interpretable Hypotheses. The fraction of hypotheses that achieve an alignment score above a threshold (0.5 or 0.8) with at least one pattern (rank score).

| | | **Sparsity** | | | |
|---|---|---|---|---|---|
| **Threshold** | **Exp. Factor** | 0 | 1 | 2 | 4 |
| $> 0.5$ | 0.5 | 12.1% | 19.2% | 18.2% | 22.0% |
| | 1 | 14.4% | 18.6% | 19.2% | 22.5% |
| | 2 | 16.8% | 19.8% | 20.9% | 22.5% |
| | 4 | 18.2% | 21.1% | 23.3% | 25.0% |
| $> 0.8$ | 0.5 | 1.1% | 3.0% | 2.6% | 3.6% |
| | 1 | 1.4% | 2.6% | 2.9% | 3.6% |
| | 2 | 1.7% | 3.8% | 3.5% | 3.7% |
| | 4 | 2.2% | 2.9% | 3.5% | 3.8% |

**SAE Encoder Architecture.** To determine whether the interpretability of the SAE benefits from increased encoder depth and additional nonlinearities, we compared the original one-layer ReLU encoder with deeper multilayer perceptron (MLP) encoders while keeping the decoder linear. We evaluated two- and three-layer MLP encoders and varied encoder weight decay across $0$, $10^{-4}$, and $10^{-3}$, using the same training and interpretability-scoring procedure as in the main experiments. As shown in Table T2, all deeper MLP variants performed substantially worse than the original one-layer ReLU encoder. These results suggest that increasing nonlinear depth does not improve interpretability.

Table T2: SAE encoder architecture ablation. Number and percentage of hypotheses with alignment scores above 0.5 and 0.8 for different encoder architectures and weight-decay values.

| Model | Weight decay | $> 0.5$ | $> 0.8$ |
|---|---|---|---|
| One-layer ReLU | 0 | 316/1331 (23.7%) | 85/1331 (6.4%) |
| MLP2 | 0 | 180 (13.5%) | 35 (2.6%) |
| MLP2 | $10^{-4}$ | 169 (12.7%) | 28 (2.1%) |
| MLP2 | $10^{-3}$ | 214 (16.1%) | 49 (3.7%) |
| MLP3 | 0 | 191 (14.4%) | 31 (2.3%) |
| MLP3 | $10^{-4}$ | 197 (14.8%) | 34 (2.6%) |
| MLP3 | $10^{-3}$ | 186 (14.0%) | 30 (2.3%) |

### B.2 Predicted fMRI pool size

We aim to disentangle the influence of the retrieval pool size (the pool from which the top-*N* most activating images are retrieved) on interpretability performance. In Table T3 we report results for retrieval pools enriched with different amounts of predicted fMRI. All decomposition methods are trained on the same combined data (measured + predicted fMRI). Unlike the main paper, where we also compare models trained only on measured fMRI, here we fix the training data and vary only the retrieval pool. Importantly, increasing the pool size also increases the number of images used to compute the hypothesis scores, since the number of top activating images is defined as a fraction of the pool (0.2%). Therefore, larger pools not only tend to yield higher interpretability scores, but also make the scores more robust. Even when scores are similar, we therefore prefer larger pools. As shown in the table, enriching the pool with predicted fMRI leads to a clear improvement compared to using measured fMRI alone, and for the best-performing model the largest pool yields the highest and most stable interpretability scores.

Table T3: **Percentage of Interpretable Hypotheses (different retrieval pools).** All decomposition methods are trained on the same combined data (measured + predicted fMRI). Unlike the main paper, where we also compare models trained only on measured fMRI, here we fix the training data and vary only the retrieval pool. We report results for pools enriched with different amounts of predicted fMRI (threshold $> 0.5$).

| Method | Measured fMRI | +Pred 30k | +Pred 60k | +Pred 90k | +Pred 120k |
|---|---|---|---|---|---|
| Voxels | 3.8% | 7.1% | 6.8% | 7.1% | 6.7% |
| PCA (Single) | 1.8% | 7.1% | 7.7% | 7.7% | 7.4% |
| NMF (Single) | 2.4% | 5.6% | 5.8% | 5.9% | 5.5% |
| ICA (Single) | 10.4% | 18.0% | 18.5% | 17.7% | 18.1% |
| ICA (Multiple) | 10.0% | 17.4% | 17.6% | 17.9% | 18.3% |
| SAE (Single) | 5.6% | 16.0% | 14.6% | 15.5% | 15.7% |
| SAE (Multiple) | 8.2% | 17.8% | 16.8% | 17.4% | 17.4% |
| SAE+ICA | 9.8% | 20.7% | 20.6% | 20.9% | 21.5% |

### B.3 Per-ROI quantitative results

We report, for each ROI, the percentage of explained hypotheses above a 0.5 threshold for the averaged results shown in the paper. V1–V3 are averaged across ventral and dorsal partitions, and FBA, FFA, and VWFA are averaged across their sub-areas. Higher-level visual regions show greater interpretability than early visual areas—likely because they encode higher-level semantics that are both easier to decompose and to describe using natural language within our pipeline.

Table T4: **Per-ROI quantitative results.** Per-ROI percentage of explained hypotheses ($> 0.5$) on the pool enriched with all predicted fMRI (+120k predicted fMRI). V1–V3 are averaged across ventral/dorsal; FBA, FFA, and VWFA are averages of their sub-areas.

| Method | V1 | V2 | V3 | hV4 | EBA | OFA | OPA | OWFA | PPA | FFA | RSC | FBA | VWFA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Voxels | 0.2% | 0.5% | 0.6% | 1.2% | 5.8% | 1.1% | 2.8% | 3.1% | 2.1% | 2.7% | 1.0% | 2.8% | 3.0% |
| PCA (Single) | 0.3% | 0.5% | 0.5% | 0.5% | 4.4% | 0.7% | 2.3% | 2.7% | 2.9% | 2.0% | 1.8% | 2.6% | 3.2% |
| NMF (Single) | 0.2% | 0.1% | 0.4% | 0.4% | 2.0% | 0.8% | 1.1% | 1.2% | 1.7% | 1.6% | 2.2% | 2.4% | 1.1% |
| ICA (Single) | 1.9% | 2.7% | 5.7% | 11.4% | 15.3% | 7.1% | 13.9% | 14.5% | 12.8% | 8.2% | 6.2% | 9.7% | 0.2% |
| SAE (Single) | 0.6% | 0.9% | 1.3% | 4.8% | 14.5% | 1.9% | 9.2% | 10.1% | 6.8% | 5.7% | 2.9% | 7.6% | 7.6% |
| SAE+ICA | 1.8% | 2.8% | 6.8% | 11.0% | 18.8% | 7.5% | 16.8% | 17.2% | 13.6% | 8.6% | 5.2% | 10.7% | 11.8% |

### B.4 Complementarity across decompositions

We aim to understand which decomposition methods are most beneficial to combine, i.e., which ones provide complementary patterns that explain hypotheses the others do not. To this end, we measure, for every pair of decompositions, the change in the percentage of explained hypotheses (threshold 0.5) when combining their patterns. "Combining" a pair means pooling the patterns from both methods and recomputing the fraction of explained hypotheses; for reference, combining a method with itself corresponds to using multiple runs (different seeds) of the same decomposition. The reported value is the difference between the combined score and the maximum score of the two methods when used separately. As shown in Table T5, SAE and ICA are the most complementary pair, each substantially improving the other.

Table T5: Pairwise complementarity between decompositions. Values show the gain in explained hypotheses (threshold 0.5) when combining each pair, relative to the better single method.

| | Voxels | PCA | NMF | ICA | SAE |
|---|---|---|---|---|---|
| Voxels | 0.0% | 0.5% | 1.4% | 0.1% | -0.2% |
| PCA | 0.5% | 0.0% | 0.0% | 0.4% | 0.0% |
| NMF | 1.4% | 0.0% | 0.0% | 0.6% | 0.1% |
| ICA | 0.1% | 0.4% | 0.5% | 0.2% | 2.1% |
| SAE | -0.2% | 0.0% | 0.1% | 2.1% | 0.7% |

### B.5 Matched Component-Budget Comparison

To separate decomposition quality from the number of candidate components, we conducted a matched component-budget experiment for PCA, NMF, ICA, and SAE. For each ROI, the number of components produced by SAE with an expansion factor of 4 defined the target component budget. To obtain comparable numbers of candidate components for NMF and ICA, we pooled components from multiple runs with different random seeds, while PCA was evaluated using the closest feasible component count. When a method still produced fewer components than the SAE target, we used its largest available component pool. Results are averaged over Subjects 1 and 8.

As shown in Table T6, SAE performs best on measured-fMRI-only data, identifying 74 hypotheses above 0.5 and 10 above 0.8, compared with 28/8 for PCA, 14/2 for NMF, and 14/2 for ICA. With predicted-fMRI augmentation, ICA performs best at the 0.5 threshold, while SAE performs better at the 0.8 threshold, with 34 and 32 hypotheses, respectively. We further examined whether the additional candidate patterns learned by SAE provide information complementary to ICA. Under the same total component budget as ICA alone, the combined ICA+SAE decomposition identifies 236 hypotheses above 0.5 and 42 above 0.8, compared with 216 and 32 for ICA alone. Thus, combining SAE and ICA improves interpretability even when the total number of candidate components is controlled, indicating that the two decompositions recover complementary candidate representations.

Results on LAION-fMRI (Table T9) were also obtained under an equal component budget and show that SAE outperforms the other methods in both the measured-fMRI-only and predicted-fMRI-

augmented settings. Thus, SAE's stronger performance and its complementarity with other methods are not simply consequences of having more candidate components available for selection.

Overall, these results highlight three important advantages of the SAE decomposition. First, under a matched component budget, SAE recovers substantially more interpretable concepts from measured fMRI than the standard decomposition methods, which is particularly important for validation that does not depend on predicted responses. Second, SAE identifies interpretable patterns beyond those found by standard decomposition methods. Although the relative performance of SAE and ICA varies across settings and subjects, and ICA can outperform SAE at the 0.5 threshold when predicted fMRI is used, the two methods recover complementary representations, as demonstrated by the improvement obtained when combining them. Importantly, the ICA+SAE results use the same total component budget as ICA alone, showing that this improvement is not due simply to a larger candidate pool. Third, beyond quantitative performance, SAE patterns are substantially more spatially localized than ICA patterns (Sec. 4.4; Fig. S3). Because SAE receives only one-dimensional voxel vectors and no spatial information, this localization emerges from the structure of the brain activity and makes the resulting patterns more informative for neuroscientific analysis.

Table T6: Matched component-budget comparison across decomposition methods. Results are averaged over Subjects 1 and 8. The component budget is matched to the number of components produced by SAE with expansion factor 4.

| Method | Measured $> 0.5$ | Measured $> 0.8$ | With predicted $> 0.5$ | With predicted $> 0.8$ |
|---|---|---|---|---|
| PCA | 28 | 8 | 76 | 8 |
| NMF | 14 | 2 | 54 | 11 |
| ICA | 14 | 2 | 216 | 32 |
| SAE | 74 | 10 | 168 | 34 |
| ICA + SAE | 73 | 10 | 236 | 42 |

### B.6 Held-Out Generalization of the SAE

To test whether the learned decomposition and its semantic interpretations generalize beyond the data used to identify them, we conducted two complementary experiments: held-out reconstruction of brain responses and held-out semantic-selectivity evaluation.

**Generalization of the learned decomposition.** We trained an SAE using only part of Subject 1's fMRI responses and then applied the learned SAE, without retraining, to responses elicited by held-out images. The reconstructed responses correlated strongly with the corresponding measured responses, with a median correlation of 0.78 and a mean correlation of 0.75. Thus, the learned decomposition generalizes to brain responses that were not used during training.

**Generalization of the discovered semantic patterns.** We used only the training split to identify each interpretable pattern and assign its corresponding hypothesis. We then kept both the pattern and hypothesis fixed and applied the learned decomposition to held-out fMRI responses. For each pattern–hypothesis pair, we asked whether the pattern received higher coefficients for held-out images containing the assigned concept than for held-out images not containing it. We quantified this separation using one-vs-rest ROC-AUC, which measures how well the pattern coefficient distinguishes images containing the assigned concept from all other images without requiring an arbitrary decision threshold. The median AUC was 0.74 for patterns with an interpretability score above 0.5 and 0.82 for patterns above 0.8. This experiment therefore tests more than reconstruction: it shows that a pattern interpreted from the training data remains selectively activated by its assigned concept in entirely new stimuli.

Together, these two experiments demonstrate generalization at both the reconstruction and semantic-selectivity levels.

### B.7 RANDOMIZED-ACTIVITY CONTROLS

To evaluate whether the discovered patterns depend on structured brain activity, we conducted two randomized-activity controls and a concept-specific statistical test.

**Voxel-shuffling control.** We independently shuffled the voxel values within each fMRI response for Subject 1 before training the SAE and then applied the full BrainExplore pipeline to the resulting activity. We measured the number of interpretable hypotheses using the same evaluation protocol as in the main experiments. The number of interpretable hypotheses decreases substantially after voxel shuffling. Using measured fMRI only, the fraction above the 0.5 threshold decreased from 6.2% (82 hypotheses) to 0.98% (13 hypotheses), while the fraction above 0.8 decreased from 1.3% (17 hypotheses) to 0.075% (1 hypothesis). Using measured and predicted fMRI, the corresponding values decreased from 17.4% (231 hypotheses) to 1.43% (19 hypotheses) at 0.5 and from 3.8% (51 hypotheses) to 0.15% (2 hypotheses) at 0.8. Voxel shuffling preserves the distribution of voxel values within each response while disrupting their correspondence with voxel locations. Consequently, distributed or global properties of the original responses can remain after shuffling. For example, when a concept is represented across many voxels, independently shuffling voxel values can preserve weak associations between that concept and individual voxel positions across images.

**Random-signal control.** We next generated random brain signals from the same distribution as the original brain responses and repeated the full BrainExplore pipeline using three random seeds. Only two concepts exceeded the original interpretability thresholds across these experiments: *"middle center"* above 0.8 and *"middle left"* above 0.5.

To quantify statistical significance directly, we constructed a concept-specific null distribution for each hypothesis by sampling 10,000 random image sets and computing the corresponding BrainExplore score for each set. We then compared the observed score with its null distribution and applied Benjamini–Hochberg correction for multiple comparisons. Neither of the two concepts identified in the random-signal experiments remained significant after correction, resulting in zero significant discoveries under the randomized-signal control. In comparison, 99% of the concepts identified using the original brain responses remained significant after correction.

# C Extended Neuroscientific Validations

## C.1 Alignment between ROI-discovered representations and known functionality

In our method, we can ask, for each ROI, which concepts are represented in it, meaning which concepts have a brain pattern in this ROI with a high interpretability score. In parallel, we can also ask which concepts from our hypothesis dictionary align with the ROI's known functionality. We can then measure, for each ROI, what fraction of these functionally aligned concepts are indeed found to be interpretable by our method.

To create the ROI–concept alignment list, we use a strong LLM (`GPT-5-mini`[1]). We first ask it to describe, for each ROI, the general known functionality according to neuroscientific knowledge, considering only clear and widely accepted functional roles. Then, for each ROI and concept in our dictionary, we ask whether the concept is well aligned with the functionality of that ROI. The model is instructed to assign a concept to an ROI only when the concept clearly matches the ROI's functional category. For example, the known functionality of FFA is faces, so both general concepts such as "face" or "human face" are included in the relevant concept list for this ROI, as well as more nuanced ones such as "smiling face". EBA is known to correspond to body parts and actions, so concepts such as "hands," "running," and "skateboarding" are all aligned as relevant. Concepts that do not clearly correspond to a known functional domain are not assigned to any ROI. This procedure provides, for each ROI, a set of *functionally relevant* concepts, i.e., concepts that are assigned to that ROI by the LLM based on known functional selectivity.

We then use the score produced by our pipeline for each ROI and concept, namely the alignment score of the best-matching pattern in that ROI for that concept. We examine the relevant concept list for each ROI and measure what fraction of those concepts receives a score greater than 0.5 from our pipeline. To ensure robustness across subjects, we restrict the analysis to *globally interpretable concepts*, defined as concepts with interpretability score $> 0.5$ in at least three subjects (regardless of ROI). For each subject and ROI, we then define the set of *interpretable concepts* as those with score $> 0.5$ in that specific subject and ROI. Both the relevant and interpretable sets are restricted to the global concept set. Finally, for each subject and ROI, we compute the alignment score as the fraction of relevant concepts (as defined above) that are also interpretable. ROIs with fewer than five relevant concepts are excluded from the analysis. Results (Fig. S1) show consistently high values across subjects and ROIs, indicating that concepts identified as interpretable within an ROI tend to match its known functional specialization.

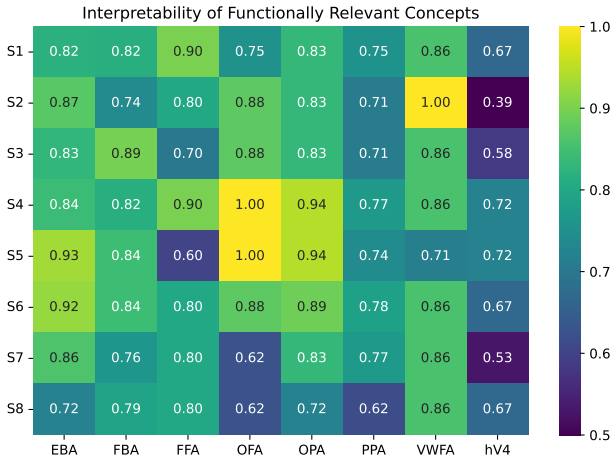


Figure S1: **Alignment between ROI-discovered representations and known functionality.** For each subject (rows) and ROI (columns), we report the proportion of concepts labeled as functionally relevant to that ROI (based on automatic LLM assignment) that are also found to be represented in that ROI (score $> 0.5$). Higher values indicate stronger agreement between the known functional role of an ROI and the concepts identified by our method.

[1] `https://developers.openai.com/api/docs/models/gpt-5-mini`

### C.2 CROSS-SUBJECT CONSISTENCY OF INTERPRETABLE CONCEPTS

To assess the robustness of discovered concepts across subjects, we measure the overlap of interpretable concepts identified in each ROI using an intersection-over-union (IoU) metric. For each ROI, we consider subsets of $k$ subjects and compute the IoU of their corresponding concept sets, defined as the size of the intersection divided by the size of the union. We report the average IoU across subsets of size $k = 2, 4, 8$ (Table 2 in the main text).

To evaluate whether the observed overlap exceeds chance, we compare against a permutation-based null model, where for each subset we randomly assign ROIs to subjects and recompute the IoU. We use this null distribution to assess statistical significance.

### C.3 CROSS-SUBJECT CONSISTENCY OF INTERPRETABLE PATTERNS

To further validate the biological relevance of the discovered SAE patterns, we asked whether patterns found in one ROI tend to have similar counterparts in the same or functionally related ROIs, both within and across subjects. For this analysis, we considered each SAE pattern in Subject 1 and measured its overlap with patterns from other ROIs using the intersection-over-union (IoU) between their sets of top-activating images. For every pattern in Subject 1, and for each target ROI separately, we identified the pattern with the highest IoU in that ROI. We then averaged these best-match scores across all patterns belonging to a given ROI in Subject 1, producing an ROI-to-ROI similarity matrix.

We visualize these results as heatmaps. Since within-Subject-1 comparisons contain especially high values on the diagonal, reflecting similarity between different SAE patterns learned in the same ROI, we do not show the Subject-1-to-Subject-1 diagonal.

Although the absolute IoU values are modest, since having exactly the same top-activating images is unlikely, the resulting structure is informative. First, the heatmaps are broadly consistent across subjects, suggesting that the discovered SAE patterns capture reproducible aspects of functional organization. Second, ROIs with similar functionality tend to show higher mutual similarity. In particular, early visual ventral regions cluster more strongly with each other than with dorsal regions, and additional ROIs with related functionality also exhibit stronger overlap. This indicates that the SAE-derived patterns are not only interpretable in terms of fine-grained visual concepts, but also capture the broader functional organization of the cortex, both within individual subjects and across subjects.

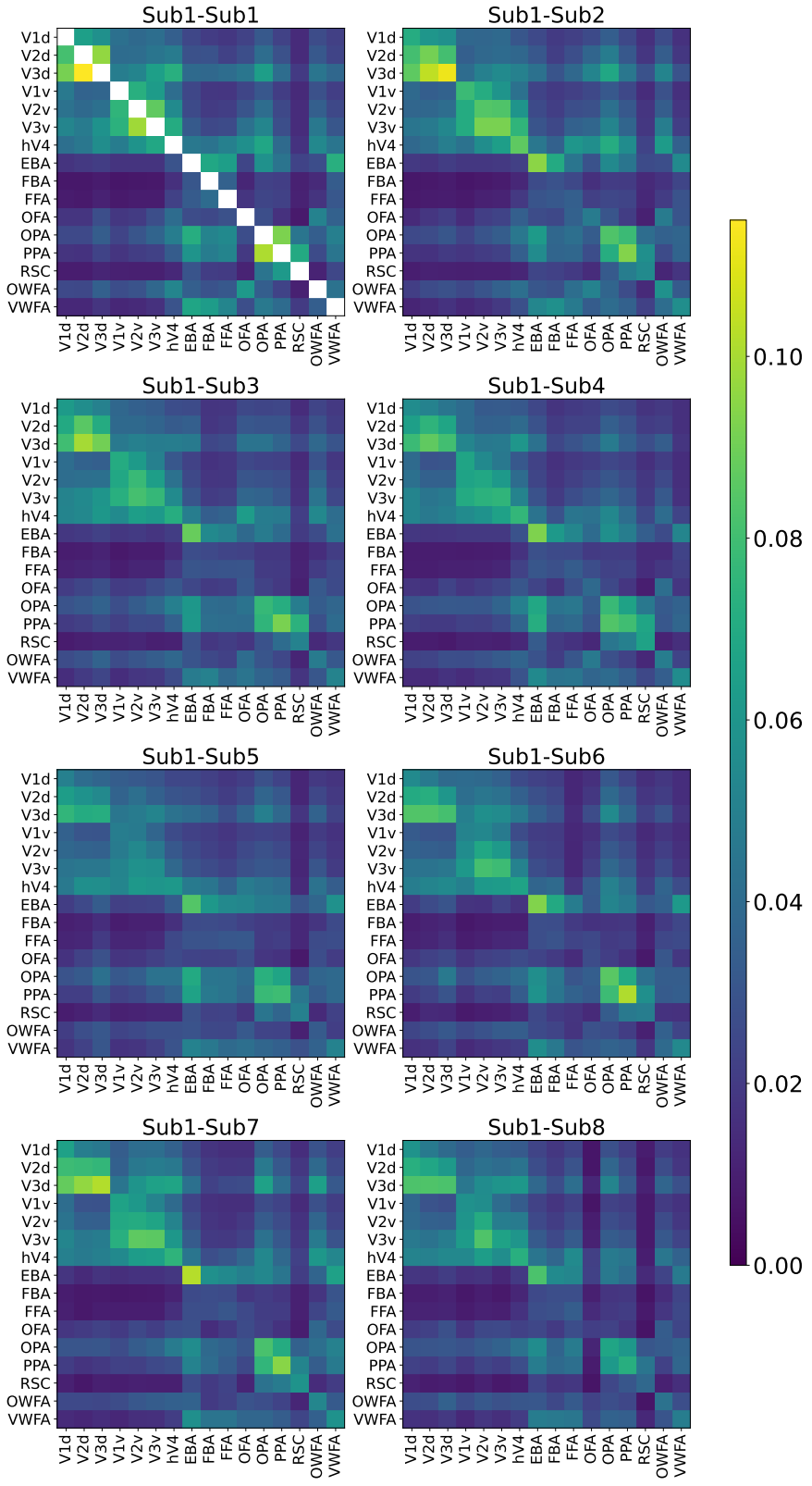


Figure S2: **Cross-subject consistency of SAE patterns.** For each SAE pattern in Subject 1, we find the pattern with the highest IoU of top-activating images in each target ROI, within the same subject and across other subjects, and average these scores across patterns from each source ROI. Functionally related ROIs show higher similarity, and the overall structure is consistent across subjects.

# D JOINT-ROI DECOMPOSITION

To examine whether conventional ROI boundaries can limit the discovery of distributed representations, we evaluated whether jointly decomposing activity across multiple ROIs improves interpretability. BrainExplore can be applied to arbitrary sets of voxels, and our ROI-wise decomposition is primarily motivated by computational tractability rather than an assumption that representations are confined to individual ROIs.

**Joint decomposition across higher-level visual cortex.** For Subject 1, we trained each decomposition method jointly across all analyzed higher-level visual-cortex voxels, excluding V1–V3 because including the full voxel set remains computationally prohibitive. We compared the resulting joint decomposition with the ROI-wise analysis, where the best component for each hypothesis is selected across ROIs. As shown in Table T7, joint decomposition improves SAE performance from 203 to 235 hypotheses above 0.5 and from 42 to 57 above 0.8. ICA also improves substantially, particularly at the stricter threshold, while voxels remain similar and PCA and NMF do not improve. The method-dependent effect suggests that some representations are better captured when decomposition is allowed to span conventional ROI boundaries. For SAE, examples of concepts with low scores in the ROI-wise analysis but high scores in the joint model include wings, mouth, mid-air, public space, bent knees, and pets.

**Joint SAE decomposition within functionally related ROI groups.** To further isolate the effect of grouping ROIs, we trained a single SAE jointly within each functional-localizer (fLoc) group (body, face, place, and word) rather than training separate SAEs for each individual ROI. Joint SAE decomposition improves concept recovery across all four functional groups at the 0.5 threshold. When considering all fLoc systems, the number of interpretable hypotheses increases from 196 to 238 above 0.5 and from 42 to 50 above 0.8 (Table T8). These results further support the conclusion that some visual representations span conventional ROI boundaries and can be better captured when related regions are jointly decomposed.

Table T7: Joint decomposition across higher-level visual cortex for Subject 1. We compare ROI-wise decomposition, where the best component is selected across ROIs for each hypothesis, with a single joint decomposition across all analyzed higher-level visual-cortex voxels.

| Method | Per-ROI $> 0.5$ | Joint $> 0.5$ | Per-ROI $> 0.8$ | Joint $> 0.8$ |
|---|---|---|---|---|
| Voxels | 87 | 96 | 19 | 19 |
| PCA | 100 | 71 | 13 | 10 |
| NMF | 68 | 72 | 8 | 8 |
| ICA | 242 | 298 | 57 | 115 |
| SAE | 203 | 235 | 42 | 57 |

Table T8: Joint SAE decomposition within functionally related ROI groups. Results compare separate ROI-wise SAEs with a single SAE jointly trained within each functional-localizer (fLoc) system.

| Training strategy | Body | | Face | | Place | | Word | |
|---|---|---|---|---|---|---|---|---|
| | $> 0.5$ | $> 0.8$ | $> 0.5$ | $> 0.8$ | $> 0.5$ | $> 0.8$ | $> 0.5$ | $> 0.8$ |
| Separate ROI SAEs | 172 | 39 | 73 | 18 | 134 | 22 | 90 | 13 |
| Joint SAE per fLoc system | 211 | 46 | 96 | 18 | 160 | 34 | 111 | 15 |

# E GENERALIZATION TO LAION-FMRI

To evaluate whether our findings generalize beyond the NSD dataset, we applied BrainExplore to Subject 1 of the LAION-fMRI dataset (Zerbe et al., 2026), which contains approximately 6,000 image–fMRI pairs. We trained a new image-to-fMRI encoder on the LAION-fMRI data and applied the same concept discovery and interpretability pipeline used for NSD. For concept discovery, we used an additional pool of approximately 60,000 LAION images that were not part of the fMRI dataset, replacing the additional COCO images used in the NSD experiments. All decomposition methods were run with a similar computational budget.

Table T9 reports the number of interpretable concepts discovered by each method. SAE identifies substantially more interpretable concepts than the other single decomposition methods in the predicted image pool, with 214 concepts above the 0.5 threshold and 45 above the 0.75 threshold. The combination of ICA and SAE performs best overall, identifying 218 and 48 concepts at the two thresholds, respectively. When requiring interpretability on the combined measured and predicted data, the same ordering is largely preserved: SAE identifies 98 concepts above 0.5 and 21 above 0.75, while ICA+SAE identifies 101 and 22, respectively.

The lower number of concepts under the combined criterion compared with the predicted-pool criterion is consistent with the smaller amount of measured fMRI data available in LAION-fMRI, which provides less per-concept coverage than NSD. Nevertheless, the substantial number of interpretable concepts discovered on this independent dataset, together with the consistent advantage of SAE and SAE+ICA, provides evidence that the proposed pipeline is not specific to the NSD dataset and can recover interpretable brain representations under a different imaging dataset and image distribution.

Table T9: Number of interpretable concepts discovered on Subject 1 of LAION-fMRI. "Predicted Pool" denotes evaluation using the larger pool of predicted fMRI responses, while "Combined" denotes the criterion requiring interpretability using both measured and predicted responses.

| Method | Predicted Pool $\geq 0.5$ | Predicted Pool $\geq 0.75$ | Combined $\geq 0.5$ | Combined $\geq 0.75$ |
|---|---|---|---|---|
| Voxels | 144 | 15 | 57 | 11 |
| PCA | 57 | 7 | 36 | 4 |
| NMF | 83 | 10 | 33 | 6 |
| ICA | 86 | 20 | 48 | 12 |
| SAE | 214 | 45 | 98 | 21 |
| ICA+SAE | 218 | 48 | 101 | 22 |

# F Robustness and Additional Validations

## F.1 VLM and Human Validation

To assess the reliability of the image–hypothesis labels used by BrainExplore, we evaluated their agreement across independent VLMs and against human annotations. The main pipeline uses a two-pass VLM verification procedure; here we test whether the resulting labels are robust to the choice of VLM and agree with human judgments.

**Cross-VLM validation.** We relabeled the same 1,000 images used by the main pipeline with two independent VLMs, InternVL3.5 and Llama-3.2-Vision, in addition to the Qwen model used in BrainExplore. As shown in Table T10, the labels produced by the independent VLMs closely agree with those produced by Qwen, with 99.36% and 98.65% agreement for InternVL3.5 and Llama-3.2-Vision, respectively. To further evaluate agreement on positive labels, which are more informative given the sparse image–hypothesis label matrix, we measured the fraction of Qwen-positive labels that were also labeled positive by each independent VLM. InternVL3.5 reproduced 79.04% of Qwen-positive labels, while Llama-3.2-Vision reproduced 75.04%. Together, these results indicate that the image–hypothesis labels are robust to the choice of VLM.

Table T10: Agreement of independent VLMs with Qwen on image–hypothesis labels. Overall agreement considers all labels, while positive-label agreement measures the fraction of Qwen-positive labels reproduced by each independent VLM.

| VLM | Overall agreement | Positive-label agreement |
|---|---|---|
| InternVL3.5 | 99.36% | 79.04% |
| Llama-3.2-Vision | 98.65% | 75.04% |

**Human evaluation.** We additionally evaluated the correctness of the generated image–hypothesis labels using human annotations. Three annotators independently evaluated a total of 1,500 image–concept pairs containing both positive and negative examples. The annotators were not shown the model-generated labels and were asked to indicate whether the concept appeared in each image. Qwen achieved 82.80% agreement with the human annotations on positive labels and 93.15% agreement on negative labels, as shown in Table T11. These results show substantial agreement with independent human judgments, providing additional validation of the semantic labeling stage while some dependence on model-generated semantics remains. While agreement is not perfect, it is high, and continued improvements in vision models are likely to further improve labeling reliability and, consequently, the robustness of our method.

Table T11: Agreement between Qwen image–hypothesis labels and blinded human annotations.

| Label type | Agreement with humans |
|---|---|
| Positive labels | 82.80% |
| Negative labels | 93.15% |

## F.2 Image to fMRI Encoder Robustness

To test whether the discovered representations are dependent on a particular image-to-fMRI encoder, we repeated the Subject 1 analysis using independent image-to-fMRI models while keeping the SAE configuration and interpretability-scoring procedure fixed. We evaluated the transformer-based brain encoder of Adeli et al. (2025) and a CLIP-backbone variant of the original encoder architecture.

**Comparison across encoders.** The independent transformer-based encoder recovered fewer interpretable concepts than the original encoder, but still identified many concepts with high alignment scores. As shown in Table T12, the original encoder identified 256 concepts above 0.5 and 55 above 0.8, while the transformer-based encoder identified 192 and 30, respectively.

Additionally, we measured how many concepts discovered by the alternative transformer-based encoder were also recovered by the original encoder. Of the 192 concepts identified above 0.5 by the transformer encoder, 162 (84.4%) also exceeded 0.5 with the original encoder. At the 0.8 threshold, 29 of 30 concepts (96.7%) identified by the transformer encoder were also recovered by the original encoder.

Table T12: Number of interpretable concepts identified using different image-to-fMRI encoders on Subject 1.

| Image-to-fMRI encoder | $> 0.5$ | $> 0.8$ |
|---|---|---|
| Original encoder | 256 | 55 |
| Transformer brain encoder (Adeli et al., 2025) | 192 | 30 |

**Robustness of original-encoder discoveries.** To directly assess whether the original encoder introduces additional encoder-specific concepts, we performed the complementary analysis in the opposite direction. Among the 256 concepts scoring above 0.5 with the original encoder, 186 (72.7%) were recovered above 0.5 by the transformer-based encoder, 176 (68.8%) were recovered above 0.5 by the CLIP-based encoder, and 212 (82.8%) were recovered by at least one of the two alternative encoders. Among the 61 concepts scoring above 0.8 with the original encoder, 33 (54.1%) also exceeded 0.8 with the transformer encoder and 32 (52.5%) exceeded 0.8 with the CLIP-based encoder, while 37 (60.7%) exceeded 0.8 with at least one of the alternative encoders. When relaxing the alternative threshold to 0.5, 59 of the 61 high-scoring concepts (96.7%) were recovered by at least one alternative encoder.

The alternative encoders produced fewer high-scoring concepts overall, consistent with their lower normalized encoding correlations (0.58 and 0.61 for the transformer-based and CLIP-based encoders, respectively, compared with 0.69 for the original encoder). These results provide evidence for robustness to encoder choice, although encoder-specific effects may remain for concepts not recovered by either alternative encoder.

Table T13: Recovery of concepts discovered by the original image-to-fMRI encoder using two independent alternative encoders.

| Original threshold | Alternative threshold | Number of original concepts | Transformer-based encoder | CLIP-based encoder | Either encoder |
|---|---|---|---|---|---|
| $> 0.5$ | $> 0.5$ | 256 | 186/256 (72.7%) | 176/256 (68.8%) | 212/256 (82.8%) |
| $> 0.8$ | $> 0.8$ | 61 | 33/61 (54.1%) | 32/61 (52.5%) | 37/61 (60.7%) |
| $> 0.8$ | $> 0.5$ | 61 | 56/61 (91.8%) | 57/61 (93.4%) | 59/61 (96.7%) |

### F.3 Prevalence-Adjusted Concept Analysis

Frequent concepts may obtain high pattern scores simply because they occur in many images in the COCO dataset. To assess whether our results are driven by such prevalence effects, we performed a prevalence-adjusted concept analysis. For each concept, we defined its prevalence-adjusted score as

$$\Delta = \text{pattern score} - \text{global concept probability},$$

where the global concept probability is the fraction of images in the full image pool associated with that concept. We counted a concept only when its pattern score was at least 0.5 and its prevalence-adjusted score $\Delta$ exceeded a specified threshold. Thus, a concept was considered significant only when it occurred substantially more often among a pattern's top-activating images than expected from its overall frequency in the image pool.

Table T14 reports the percentage of concepts satisfying these criteria for increasing $\Delta$ thresholds. The relative ordering of methods remains stable as the threshold becomes more stringent. SAE and ICA substantially outperform the voxel, PCA, and NMF baselines at all thresholds, while the combination of SAE and ICA consistently achieves the highest proportion of prevalence-adjusted concepts. At the strictest threshold of $\Delta \geq 0.7$, SAE+ICA identifies 5.8% of concepts, compared with 4.6% for SAE and 3.9% for ICA, whereas the three baseline methods identify less than 1%. These results indicate that the performance differences are not explained primarily by frequent or otherwise easy-to-label concepts, and instead persist when concept prevalence is explicitly taken into account.

Table T14: Percentage of concepts with pattern score $\geq 0.5$ satisfying increasing prevalence-adjusted score thresholds on Subject 1.

| Method | $\Delta \geq 0.1$ | $\Delta \geq 0.3$ | $\Delta \geq 0.5$ | $\Delta \geq 0.7$ |
|---|---|---|---|---|
| Voxels | 6.5% | 6.4% | 4.6% | 1.0% |
| PCA | 5.9% | 5.8% | 3.8% | 0.7% |
| NMF | 4.4% | 4.3% | 2.6% | 0.4% |
| ICA | 16.4% | 16.3% | 13.0% | 3.9% |
| SAE | 16.5% | 16.4% | 13.7% | 4.6% |
| SAE+ICA | 19.3% | 19.3% | 16.3% | 5.8% |

# G Additional Details

## G.1 SAE Training Details

The Sparse Autoencoder (SAE) was trained to learn a compact and interpretable latent representation of fMRI activations, enabling decomposition into meaningful spatial patterns.

**Architecture.** We used a simple autoencoder with a linear decoder and a ReLU encoder, where both encoder and decoder are implemented as matrix multiplications , with the ReLU enforcing non-negativity in the latent activations. Sparsity is encouraged via an $\ell_1$ loss applied to the latent space, added to the reconstruction mean-squared error (MSE) loss. The *expansion factor* controls the latent dimensionality (the ratio between latent and input size), while the *sparsity coefficient* balances the sparsity loss relative to the reconstruction loss.

**Combining with predicted fMRI.** We combine the real measured fMRI data ($\sim$10k samples) with predicted fMRI activations from the image-to-fMRI encoder ($\sim$120k samples). Real fMRI measurements are inherently noisy, and many voxels exhibit low signal-to-noise ratio (SNR). Consequently, the predicted encoder often outputs values close to zero for those voxels, leading to a mismatch in distribution between measured and predicted fMRI. Training a single SAE on this mixed data causes an imbalance, resulting in higher sparsity for predicted samples and reduced sparsity for measured ones. To address this, we use separate encoder networks for the measured and predicted data, while sharing a common decoder. This allows both data types to be decomposed into the same set of components (dictionary patterns) while enabling independent control over their activations. During training, each batch contained an equal number of measured and predicted samples to ensure balanced gradients from both data sources.

**Hyperparameters.** As reported in the ablation studies, we experimented with multiple combinations of sparsity coefficients and expansion factors. For the main hyperparameter search, we considered values $\{1, 2, 4\}$ for both. The top four models selected according to the rank score threshold of 0.8 on Subject 1, were used in the *multiple model* setup: sparsity coefficient 4 with expansion factors $\{1, 2, 4\}$ and sparsity coefficient 1 with expansion factor 2. The *single model* configuration used sparsity coefficient 4 and expansion factor 4. The same hyperparameters were applied across all subjects.

## G.2 Baseline Decomposition Training Details

All baseline decomposition methods were trained in two configurations: (i) using only the measured fMRI data ($\sim$10k samples), and (ii) using the combined pool of measured and predicted fMRI activations ($\sim$120k samples). This setup enables consistent comparison between the baselines and the Sparse Autoencoder (SAE), both when relying solely on measured data and when leveraging the large scale predicted data.

**Voxels.** The voxel baseline uses the raw voxel activations without any decomposition or learned transformation. Each voxel is treated as an independent component. For visualization and interpretability, every voxel's activation pattern was duplicated into two components (the original and its negation), allowing retrieval of top activating images for both positive and negative responses. This ensures interpretability is consistent across all decomposition methods, including those that produce only nonnegative activations.

**PCA.** Principal Component Analysis (PCA) serves as a deterministic, orthogonal decomposition baseline. It was trained on both measured only and combined fMRI datasets, with the number of components determined by the cumulative explained variance. To ensure consistency for interpretability visualization, each PCA component was duplicated into two versions, positive and negative, allowing activation maps to always correspond to positively responsive image sets.

**ICA.** Independent Component Analysis (ICA) was applied using numbers of components corresponding to the dimensionalities that achieve 90%, 95%, and 98% explained variance in PCA. Higher explained variance targets ($>$98%) were found to be unstable or unnecessary due to overfitting. ICA was run multiple times with different random seeds to capture variability arising from its nondeterministic optimization process. Similar to PCA, each independent component was duplicated into its positive and negative versions for interpretability visualization.

**NMF.** Nonnegative Matrix Factorization (NMF) was trained with the same component counts as ICA (matching the 0.9, 0.95, and 0.98 explained variance configurations). Since NMF produces strictly nonnegative components, no positive or negative duplication was needed. NMF was also

trained on both measured and combined fMRI data. Different hyperparameter settings (for example, initialization and regularization) showed minimal impact on the final interpretability results, consistent with its stable convergence properties under nonnegativity constraints.

### G.3 VISUALIZE & EXPLAIN

As described in the paper, in the *Visualize* step we select, for every pattern, its top-$N$ activating responses and collect their corresponding images, producing a set of images that *visualize* what most strongly activates that pattern. This is done separately for measured fMRI responses and for predicted responses. We take the six most activating images from the measured fMRI pool and the ten most activating images from the predicted fMRI pool (16 images in total).

For each image, we first generate a detailed caption using `Qwen2.5-VLM-32B`[2]. The VLM is instructed (see prompt in Fig. S17) to produce rich descriptions including objects, scene or room type, dominant colors, ongoing activities, and body postures.

Given the set of $N$ captions, we then ask a language model `Qwen2.5-32B`[3] to generate between 3 and 12 hypotheses that may explain what is shared among at least half of those images (Fig. S18). Requesting multiple hypotheses provides diversity, since both the captions and the LLM can be biased toward a dominant concept, whereas smaller but meaningful shared attributes might otherwise be missed.

We intentionally separate the VLM and LLM stages. We found that current vision–language models often struggle to infer what is shared across multiple images, frequently producing incorrect or overly general hypotheses. By contrast, separating the process allows each VLM to focus on a single image, producing high-quality detailed captions, while the LLM operates only on textual descriptions—an easier and more reliable input for identifying cross-image commonalities.

### G.4 SCALING TO AN UNLIMITED NUMBER OF PATTERNS

Explaining every component in every ROI and for every decomposition method is costly and inefficient. To scale interpretability, we introduce two steps: (i) ***Hypothesis dictionary generation***, and (ii) ***Hypothesis–image labeling***. Once the dictionary and the image labels are in place, any new brain pattern can be evaluated by measuring how consistently its top-activating images express concepts from the dictionary. The prompts used for hypothesis–image labeling are shown in Fig. S19, and examples of images with positively labeled hypotheses appear in Fig. S20.

### G.5 REGION OF INTEREST (ROI)

We analyze activity within predefined Regions of Interest (ROIs), which are standard cortical parcels used to summarize responses within broader anatomical and functional subdivisions of the visual system. In the Algonauts/NSD data, these ROIs are supplied by the dataset releases based on independent localizers and anatomical parcellations. The full set used here is: V1v, V1d, V2v, V2d, V3v, V3d, hV4, VWFA-1, VWFA-2, EBA, FBA-1, FBA-2, FFA-1, FFA-2, OFA, OPA, OWFA, PPA, and RSC. Early visual areas (V1–V3, separated into ventral and dorsal subregions) correspond to the low-level visual cortex and are primarily involved in processing basic visual attributes such as orientation, color, and spatial frequency. Area hV4 is associated with color and shape processing. Higher visual regions show more specialized selectivity: EBA, FBA, and OFA are body- and face-selective regions; OPA, PPA, and RSC respond to scenes, places, and navigation-related information; VWFA and OWFA are selective to word and object forms, respectively. Together, these ROIs span the visual hierarchy from low-level visual representations to high-level semantic and categorical processing.

### G.6 COMPUTATIONAL RESOURENCES

The hypothesis generation over image sets was performed using large LLM and VLM models for the upscaling stage on an H200 GPU, requiring around 100 GPU hours. The image labeling was performed on an L40S GPU using a smaller 8B VLM model, and required around 100 GPU hours.

[2] `https://huggingface.co/Qwen/Qwen2.5-VL-32B-Instruct`
[3] `https://huggingface.co/Qwen/Qwen2.5-32B-Instruct`

After the upscaling stage, the pipeline itself was run on L40S GPUs; the full pipeline over all decompositions and concepts took around 4 hours.

## H Additional Quantitative Evaluation

Table T15: **Number of Interpretable Patterns.** The number of patterns whose best hypothesis alignment exceeds the same threshold (0.5 or 0.8). To avoid counting near-duplicate patterns, we remove any component whose voxel-wise correlation with a higher-scoring pattern exceeds 0.5.

| | Measured fMRI | | + Predicted fMRI | |
|---|---|---|---|---|
| Method | $> 0.5$ | $> 0.8$ | $> 0.5$ | $> 0.8$ |
| Voxels | 5234 | 355 | 5905 | 291 |
| PCA (Single) | 1045 | 6 | 76 | 6 |
| NMF (Single) | 27 | 2 | 19 | 2 |
| ICA (Single) | 580 | 4 | 226 | 44 |
| ICA (Multiple) | 424 | 41 | 305 | 61 |
| SAE (Single) | 17242 | 586 | 8858 | 286 |
| SAE (Multiple) | 30005 | 1074 | 15748 | 617 |
| SAE+ICA | 30583 | 1077 | 16051 | 679 |

Table T16: **Average across subjects: Interpretable Hypotheses (count) with standard error.**

| | Measured fMRI | | + Predicted fMRI | |
|---|---|---|---|---|
| Method | $> 0.5$ | $> 0.8$ | $> 0.5$ | $> 0.8$ |
| Voxels | $42.1 \pm 4.9$ | $4.5 \pm 0.8$ | $86.5 \pm 7.0$ | $15.9 \pm 1.9$ |
| PCA | $20.4 \pm 4.8$ | $4.2 \pm 1.5$ | $78.5 \pm 7.0$ | $11.1 \pm 2.1$ |
| NMF | $19.2 \pm 4.2$ | $3.0 \pm 0.9$ | $61.5 \pm 9.0$ | $10.4 \pm 2.2$ |
| ICA | $11.8 \pm 3.7$ | $2.9 \pm 1.5$ | $207.0 \pm 19.8$ | $34.9 \pm 9.4$ |
| SAE | $104.6 \pm 10.5$ | $17.8 \pm 4.0$ | $218.9 \pm 14.0$ | $50.5 \pm 6.3$ |
| SAE+ICA | $104.5 \pm 10.4$ | $17.9 \pm 4.1$ | $250.1 \pm 18.2$ | $58.0 \pm 9.7$ |

Table T17: **Average across subjects: Interpretable Hypotheses (percentage) with standard error.**

| | Measured fMRI | | + Predicted fMRI | |
|---|---|---|---|---|
| Method | $> 0.5$ | $> 0.8$ | $> 0.5$ | $> 0.8$ |
| Voxels | $3.17 \pm 0.37\%$ | $0.34 \pm 0.06\%$ | $6.50 \pm 0.52\%$ | $1.19 \pm 0.15\%$ |
| PCA | $1.53 \pm 0.36\%$ | $0.32 \pm 0.12\%$ | $5.90 \pm 0.53\%$ | $0.84 \pm 0.16\%$ |
| NMF | $1.45 \pm 0.32\%$ | $0.23 \pm 0.07\%$ | $4.62 \pm 0.68\%$ | $0.78 \pm 0.16\%$ |
| ICA | $0.88 \pm 0.28\%$ | $0.22 \pm 0.11\%$ | $15.56 \pm 1.49\%$ | $2.62 \pm 0.71\%$ |
| SAE | $7.87 \pm 0.79\%$ | $1.33 \pm 0.30\%$ | $16.46 \pm 1.05\%$ | $3.80 \pm 0.48\%$ |
| SAE+ICA | $7.86 \pm 0.78\%$ | $1.34 \pm 0.31\%$ | $18.81 \pm 1.37\%$ | $4.36 \pm 0.73\%$ |

Table T18: **Interpretable hypotheses across all subjects (count / percentage).**

| Subject | Method | Measured fMRI | | + Predicted fMRI | |
|---|---|---|---|---|---|
| | | >0.5 | >0.8 | >0.5 | >0.8 |
| S1 | Voxels | 39 / 2.9% | 5 / 0.4% | 88 / 6.6% | 18 / 1.4% |
| | PCA | 32 / 2.4% | 11 / 0.8% | 97 / 7.3% | 13 / 1.0% |
| | NMF | 18 / 1.4% | 3 / 0.2% | 73 / 5.5% | 10 / 0.8% |
| | ICA | 11 / 0.8% | 0 / 0.0% | 244 / 18.3% | 72 / 5.4% |
| | SAE | 82 / 6.2% | 17 / 1.3% | 231 / 17.4% | 51 / 3.8% |
| | SAE+ICA | 83 / 6.2% | 16 / 1.2% | 286 / 21.5% | 78 / 5.9% |
| S2 | Voxels | 54 / 4.1% | 5 / 0.4% | 107 / 8.0% | 19 / 1.4% |
| | PCA | 30 / 2.3% | 5 / 0.4% | 96 / 7.2% | 18 / 1.4% |
| | NMF | 28 / 2.1% | 5 / 0.4% | 84 / 6.3% | 17 / 1.3% |
| | ICA | 15 / 1.1% | 2 / 0.2% | 282 / 21.2% | 48 / 3.6% |
| | SAE | 115 / 8.6% | 13 / 1.0% | 247 / 18.6% | 61 / 4.6% |
| | SAE+ICA | 115 / 8.6% | 13 / 1.0% | 310 / 23.3% | 70 / 5.3% |
| S3 | Voxels | 31 / 2.3% | 4 / 0.3% | 72 / 5.4% | 13 / 1.0% |
| | PCA | 17 / 1.3% | 2 / 0.2% | 72 / 5.4% | 8 / 0.6% |
| | NMF | 14 / 1.1% | 1 / 0.1% | 44 / 3.3% | 5 / 0.4% |
| | ICA | 5 / 0.4% | 1 / 0.1% | 175 / 13.1% | 14 / 1.1% |
| | SAE | 82 / 6.2% | 7 / 0.5% | 191 / 14.4% | 46 / 3.5% |
| | SAE+ICA | 82 / 6.2% | 7 / 0.5% | 218 / 16.4% | 45 / 3.4% |
| S4 | Voxels | 41 / 3.1% | 4 / 0.3% | 83 / 6.2% | 13 / 1.0% |
| | PCA | 20 / 1.5% | 2 / 0.2% | 60 / 4.5% | 9 / 0.7% |
| | NMF | 20 / 1.5% | 2 / 0.2% | 36 / 2.7% | 6 / 0.5% |
| | ICA | 7 / 0.5% | 2 / 0.2% | 168 / 12.6% | 31 / 2.3% |
| | SAE | 103 / 7.7% | 15 / 1.1% | 189 / 14.2% | 40 / 3.0% |
| | SAE+ICA | 103 / 7.7% | 16 / 1.2% | 202 / 15.2% | 42 / 3.2% |
| S5 | Voxels | 71 / 5.3% | 4 / 0.3% | 122 / 9.2% | 25 / 1.9% |
| | PCA | 43 / 3.2% | 11 / 0.8% | 109 / 8.2% | 17 / 1.3% |
| | NMF | 43 / 3.2% | 7 / 0.5% | 107 / 8.0% | 21 / 1.6% |
| | ICA | 36 / 2.7% | 13 / 1.0% | 269 / 20.2% | 74 / 5.6% |
| | SAE | 159 / 11.9% | 42 / 3.2% | 262 / 19.7% | 85 / 6.4% |
| | SAE+ICA | 159 / 11.9% | 43 / 3.2% | 303 / 22.8% | 110 / 8.3% |
| S6 | Voxels | 30 / 2.3% | 4 / 0.3% | 66 / 5.0% | 17 / 1.3% |
| | PCA | 8 / 0.6% | 1 / 0.1% | 66 / 5.0% | 4 / 0.3% |
| | NMF | 15 / 1.1% | 5 / 0.4% | 59 / 4.4% | 5 / 0.4% |
| | ICA | 6 / 0.5% | 1 / 0.1% | 210 / 15.8% | 14 / 1.1% |
| | SAE | 94 / 7.1% | 12 / 0.9% | 226 / 17.0% | 36 / 2.7% |
| | SAE+ICA | 93 / 7.0% | 12 / 0.9% | 251 / 18.9% | 36 / 2.7% |
| S7 | Voxels | 38 / 2.9% | 9 / 0.7% | 87 / 6.5% | 16 / 1.2% |
| | PCA | 3 / 0.2% | 1 / 0.1% | 75 / 5.6% | 17 / 1.3% |
| | NMF | 3 / 0.2% | 0 / 0.0% | 57 / 4.3% | 13 / 1.0% |
| | ICA | 10 / 0.8% | 3 / 0.2% | 193 / 14.5% | 19 / 1.4% |
| | SAE | 132 / 9.9% | 26 / 2.0% | 257 / 19.3% | 58 / 4.4% |
| | SAE+ICA | 131 / 9.8% | 26 / 2.0% | 266 / 20.0% | 58 / 4.4% |
| S8 | Voxels | 33 / 2.5% | 1 / 0.1% | 67 / 5.0% | 6 / 0.5% |
| | PCA | 10 / 0.8% | 1 / 0.1% | 53 / 4.0% | 3 / 0.2% |
| | NMF | 13 / 1.0% | 1 / 0.1% | 32 / 2.4% | 6 / 0.5% |
| | ICA | 4 / 0.3% | 1 / 0.1% | 115 / 8.6% | 7 / 0.5% |
| | SAE | 70 / 5.3% | 10 / 0.8% | 148 / 11.1% | 27 / 2.0% |
| | SAE+ICA | 70 / 5.3% | 10 / 0.8% | 165 / 12.4% | 25 / 1.9% |

# I ADDITIONAL VISUALIZATIONS

**Spatial localization of interpretable patterns (SAE vs. ICA).** Extending the main-text observation that SAE patterns are more spatially localized, Fig. S3 presents side-by-side brain activation maps for multiple ROIs, comparing patterns derived from ICA and from SAE. Despite being trained on one-dimensional voxel vectors with no spatial priors, SAE produces compact, clustered patterns, whereas ICA frequently yields more diffuse or scattered maps. All maps were generated using identical preprocessing, color scaling, and selection criteria (top-activating images per pattern), ensuring that visual differences reflect the decomposition method rather than visualization settings.

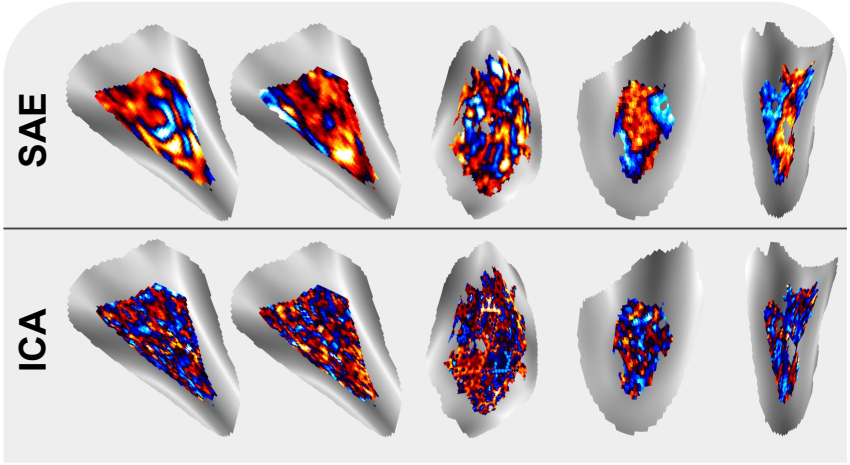


Figure S3: **Brain activation map (SAE vs ICA)**. We show different activation patterns derived from a Sparse Autoencoder (SAE) and Independent Component Analysis (ICA) for different ROIs. SAE patterns are notably more compact and clustered and spatially localized, despite both receiving only 1D voxel vectors and no spatial information.

**Additional concepts across ROIs.** We present additional interpretable patterns from Subject 1 together with representative top-activating images. We show further results for EBA and RSC in Fig. S4, for OPA in Fig. S5, and for FBA-1 and FBA-2 in Fig. S6.

**Additional subjects visualization.** We present discovered interpretable patterns and concepts for two additional subjects with a larger number of examples (Subjects 2 and 5) in Figs. S7 to S10. All subjects' trained decompositions and discovered patterns will be publicly available, enabling further research and discoveries.

**Full top 16 images per pattern.** We present full visual grids showing the top 16 activating images for the concepts from EBA discussed in the main paper (Figs. S11 and S12). For each concept, the top row corresponds to predicted fMRI responses and the bottom row to measured fMRI responses. Images are ordered by activation strength, from left to right. We also show all 16 top-activating images for the patterns and concepts presented in the main paper for PPA (Figs. S13 and S14).

**Concepts best explained by each ROI.** We summarize which semantic concepts are most strongly represented across different Regions of Interest (ROIs). Fig. S15 shows, for each ROI, the concepts achieving an alignment score above 0.5, allowing a concept to appear in multiple ROIs if it is represented across regions. In contrast, Fig. S16 presents an exclusive version in which each concept is assigned only to the ROI where it achieves its highest alignment score. Together, these visualizations reveal both the distributed and the region-specific organization of conceptual representations across the visual cortex.

**Prompts.** Figures S17 to S19 depict the prompts used in our pipeline: first, detailed per-image captions are elicited (Fig. S17); second, a language model infers shared hypotheses across the image set (Fig. S18); and, distinct from the two-stage explanation, the image–hypothesis labeling prompt is used in the *upscale* stage to decide whether each hypothesis is supported by each image (Fig. S19).

**Labeled examples.** Figure S20 presents representative images together with the hypotheses labeled as positive for each, demonstrating the output produced by the hypothesis–image labeling stage.

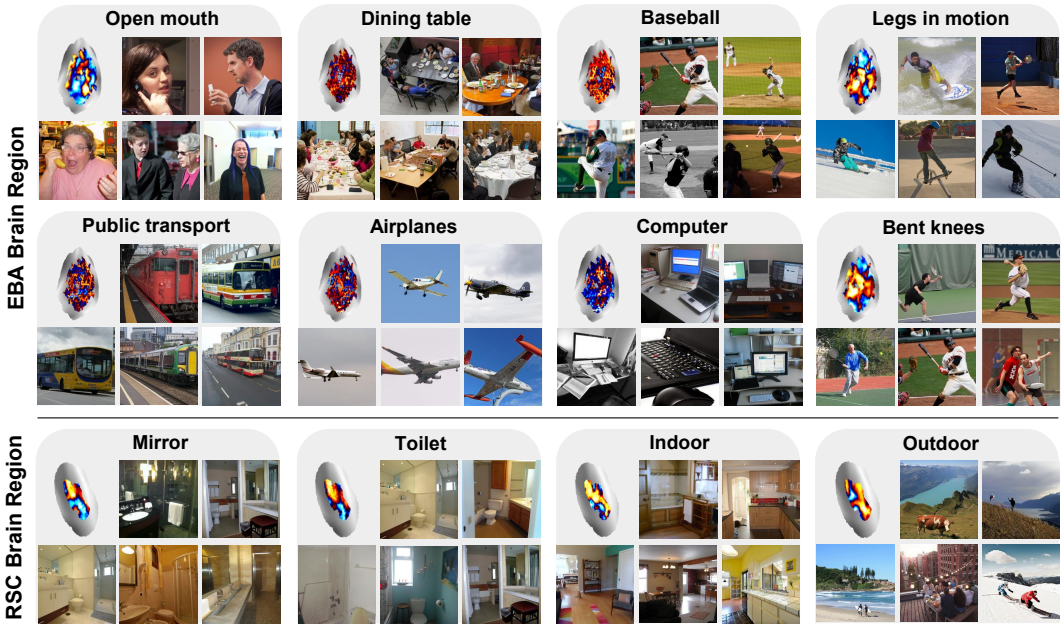


Figure S4: **Discovered Interpretable Patterns (EBA and RSC)**. We show additional patterns for Subject 1 with top activating images and selected explanations. EBA is known to encode bodies and actions, whereas RSC processes scene information.

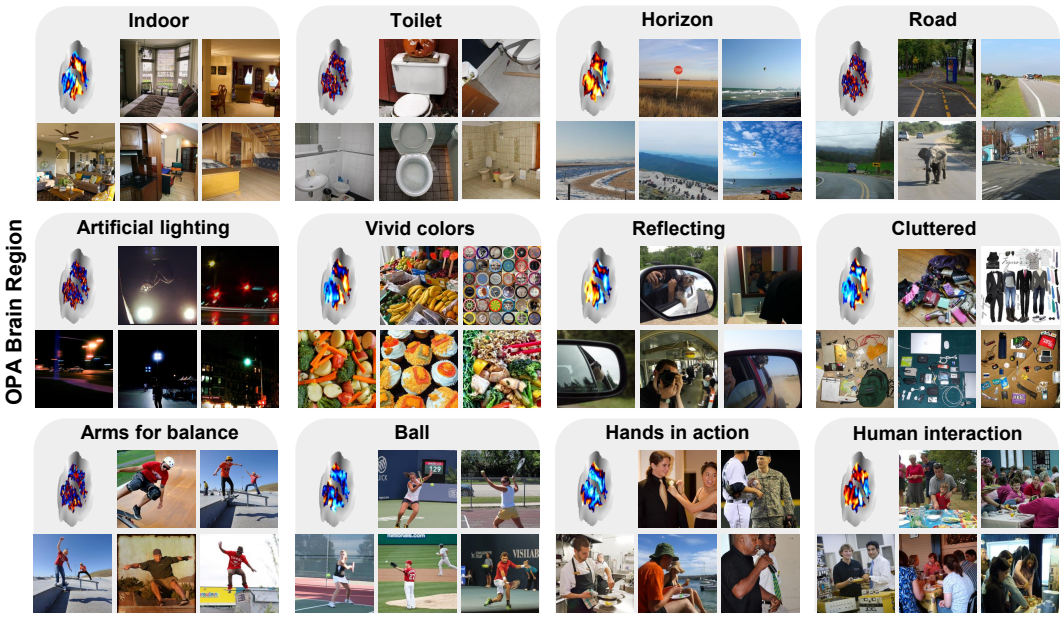


Figure S5: **Discovered Interpretable Patterns (OPA)**. We show additional patterns for Subject 1 with top activating images and selected explanations. OPA is known to process scene layout and navigability.

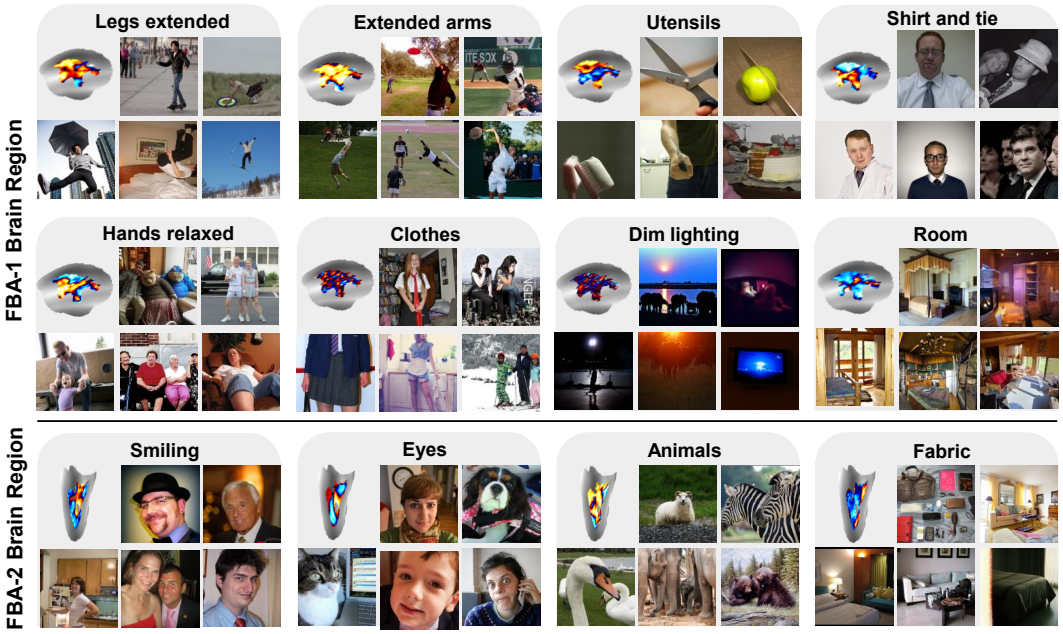


Figure S6: **Discovered Interpretable Patterns (FBA-1 and FBA-2)**. We show additional patterns for Subject 1 with top activating images and selected explanations. FBA is a body-selective area, encompassing two sub-areas shown here, FBA-1 and FBA-2.

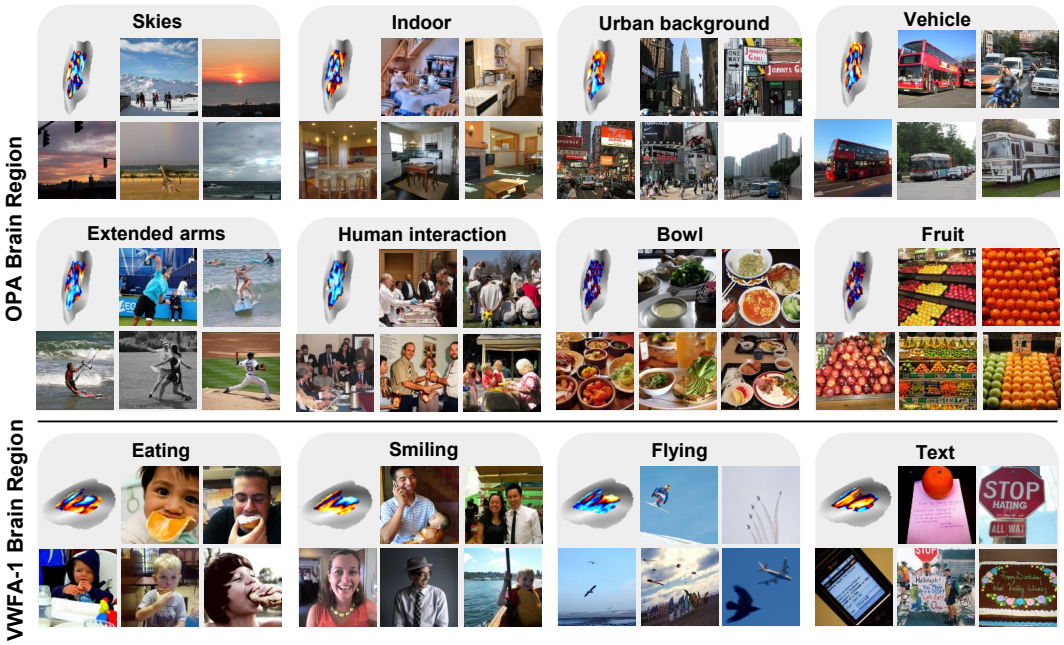


Figure S7: **Discovered Interpretable Patterns of Subject 2 (OPA and VWFA-1)**. We show patterns for Subject 2 with top activating images and selected explanations. OPA is known to process scene layout and navigability, and VWFA is involved in processing word shapes and visual text.

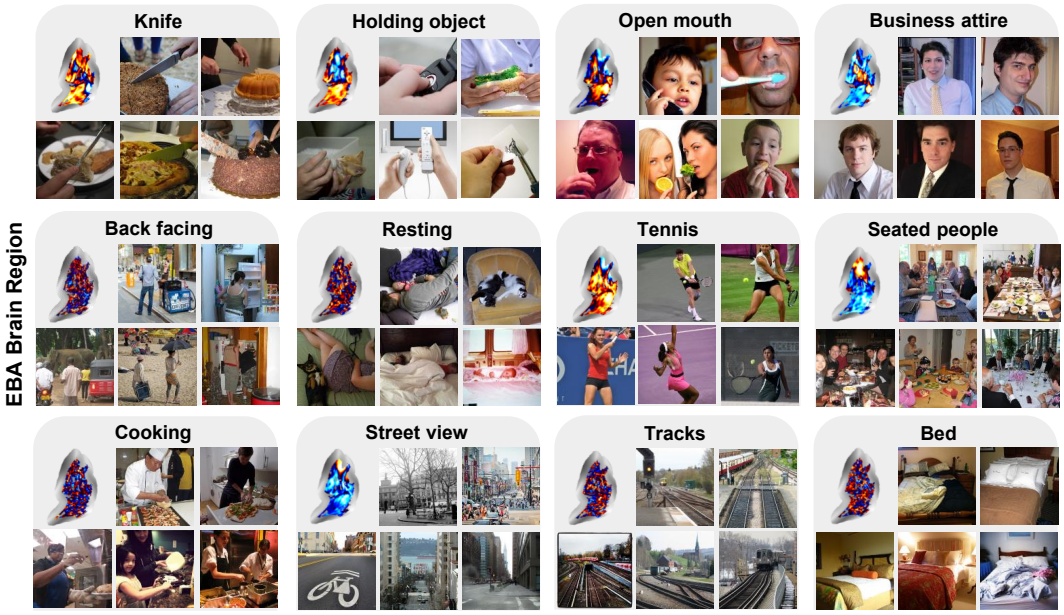


Figure S8: **Discovered Interpretable Patterns of Subject 2 (EBA)**. We show patterns for subject 2 with top activating images and selected explanations. EBA is known to encode bodies and actions.

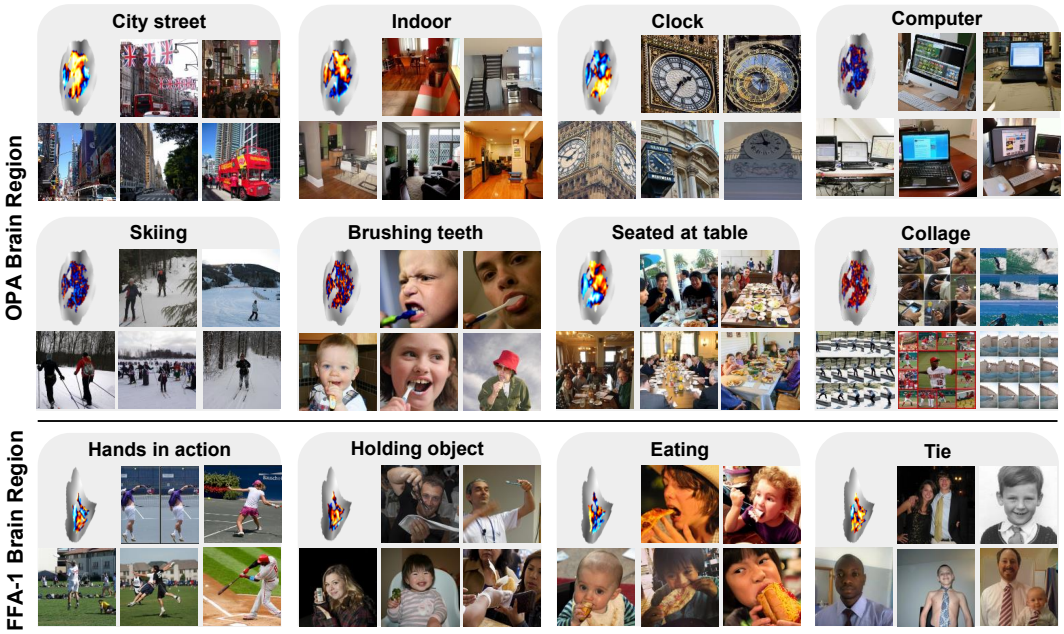


Figure S9: **Discovered Interpretable Patterns of Subject 5 (OPA and FFA-1)**. We show patterns for subject 5 with top activating images and selected explanations. OPA is known to process scene layout and navigability, and FFA is primarily known for face processing.

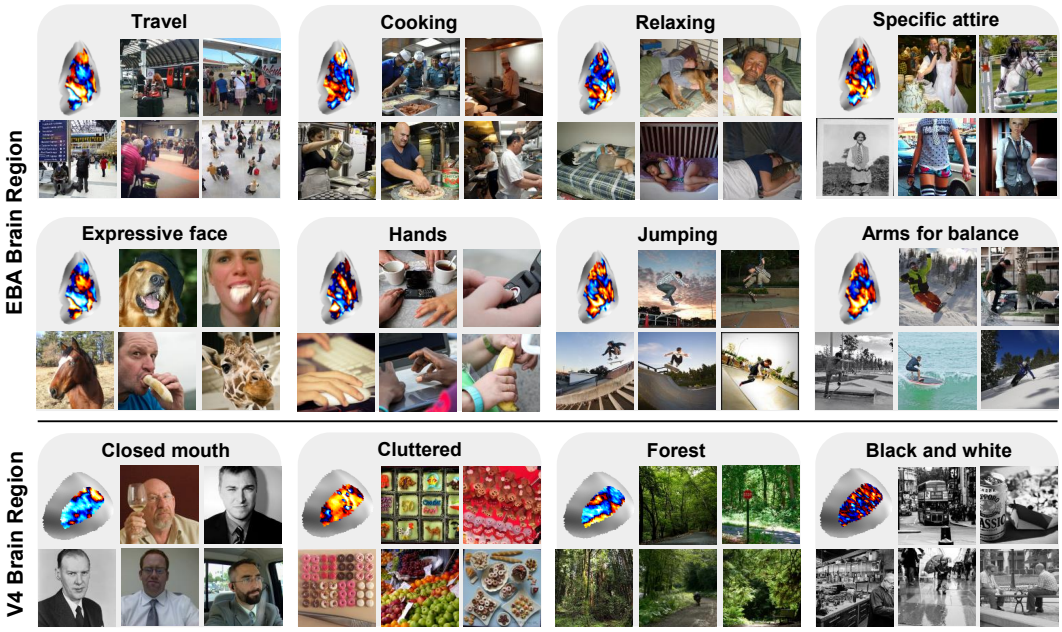


Figure S10: **Discovered Interpretable Patterns of Subject 5 (EBA and hV4)**. We show patterns for subject 5 with top activating images and selected explanations. EBA is known to encode bodies and actions and V4 is known to encode mid-level features (e.g., color, shape).

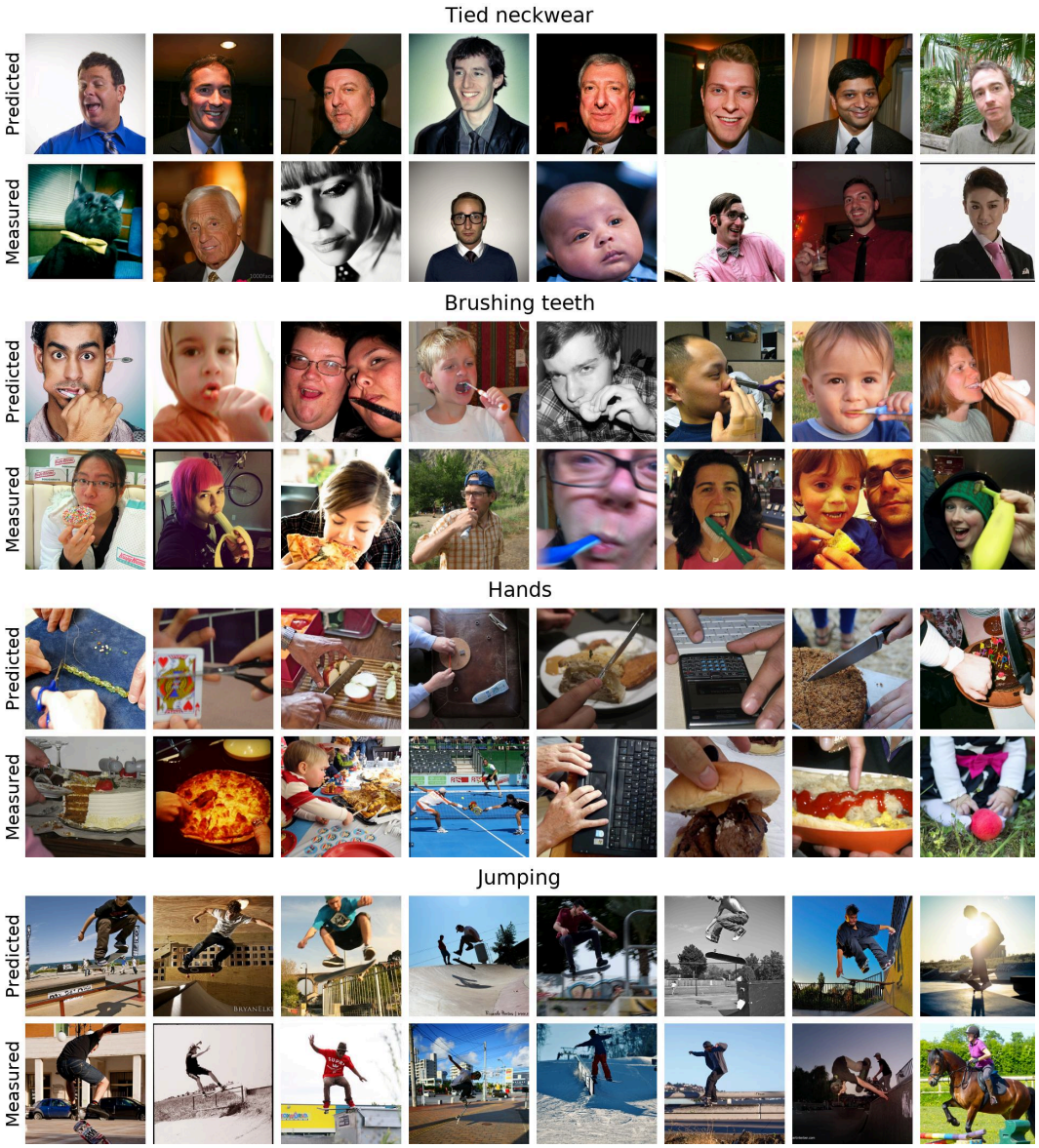


Figure S11: **Image grid EBA (Part 1).** Full visual grid with top 16 activating images for concepts *Tied neckwear*, *Brushing teeth*, *Hands*, and *Jumping* found in the EBA region. For each concept, the top 8 images corresponding to **Measured** and **Predicted** fMRI responses are shown.

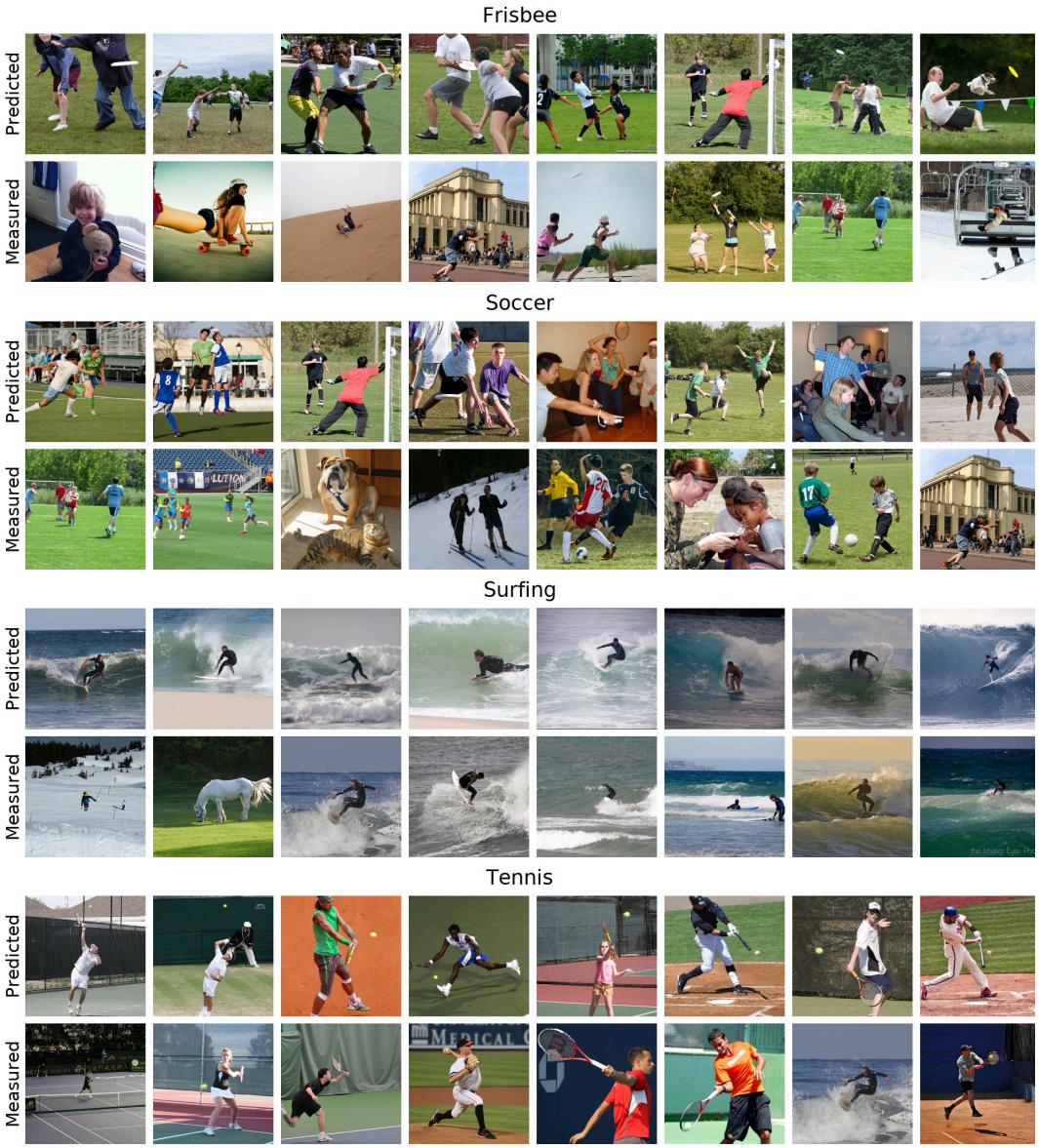


Figure S12: **Image grid EBA (Part 2).** Full visual grid with top 16 activating images for concepts *Frisbee*, *Soccer*, *Surfing*, and *Tennis* found in the EBA region. For each concept, the top 8 images corresponding to **Measured** and **Predicted** fMRI responses are shown.

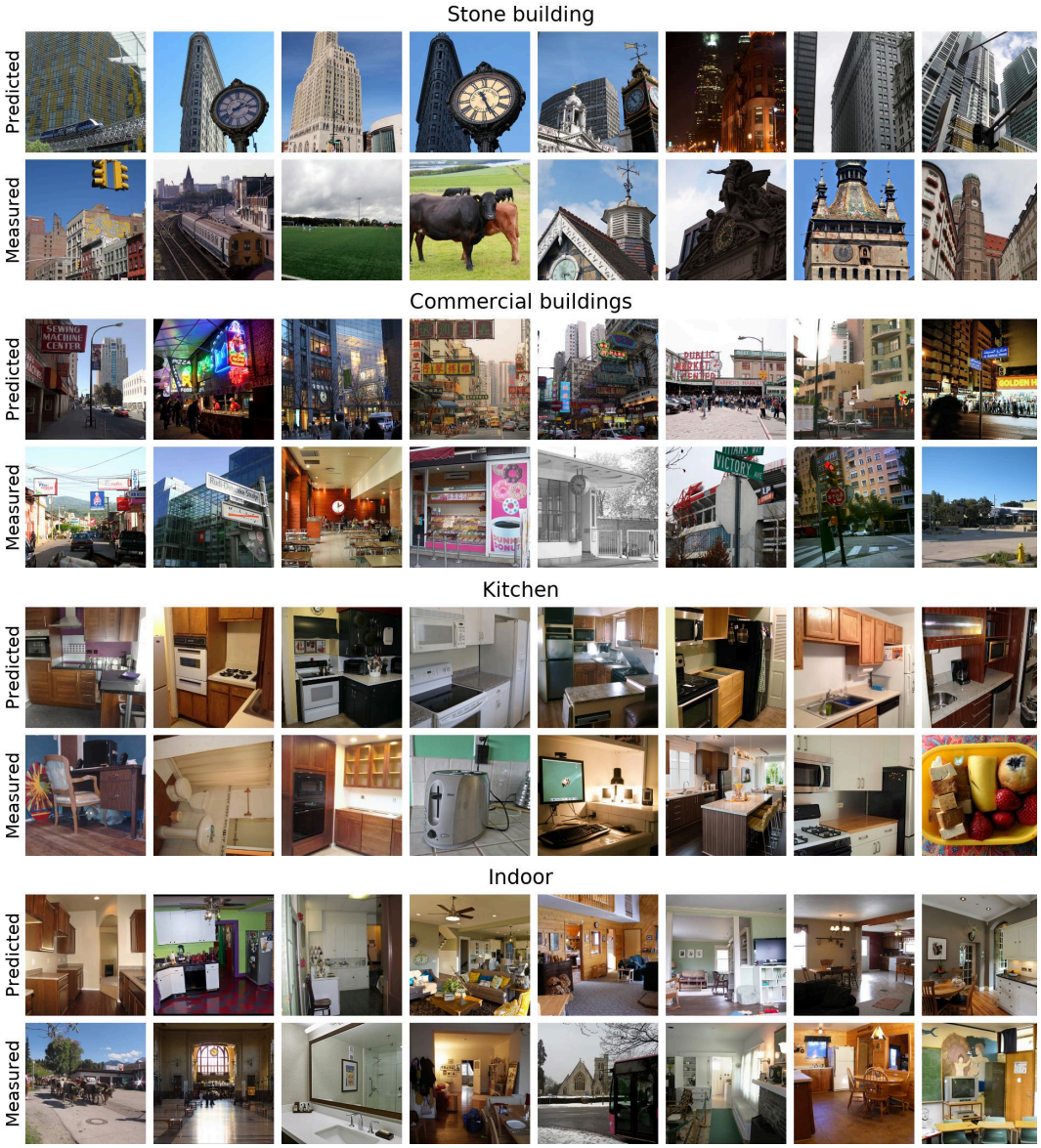


Figure S13: **Image grid PPA (Part 1).** Full visual grid with top 16 activating images for concepts *Stone building*, *Commercial buildings*, *Kitchen*, and *Indoor* found in the PPA region. For each concept, the top 8 images corresponding to **Measured** and **Predicted** fMRI responses are shown.

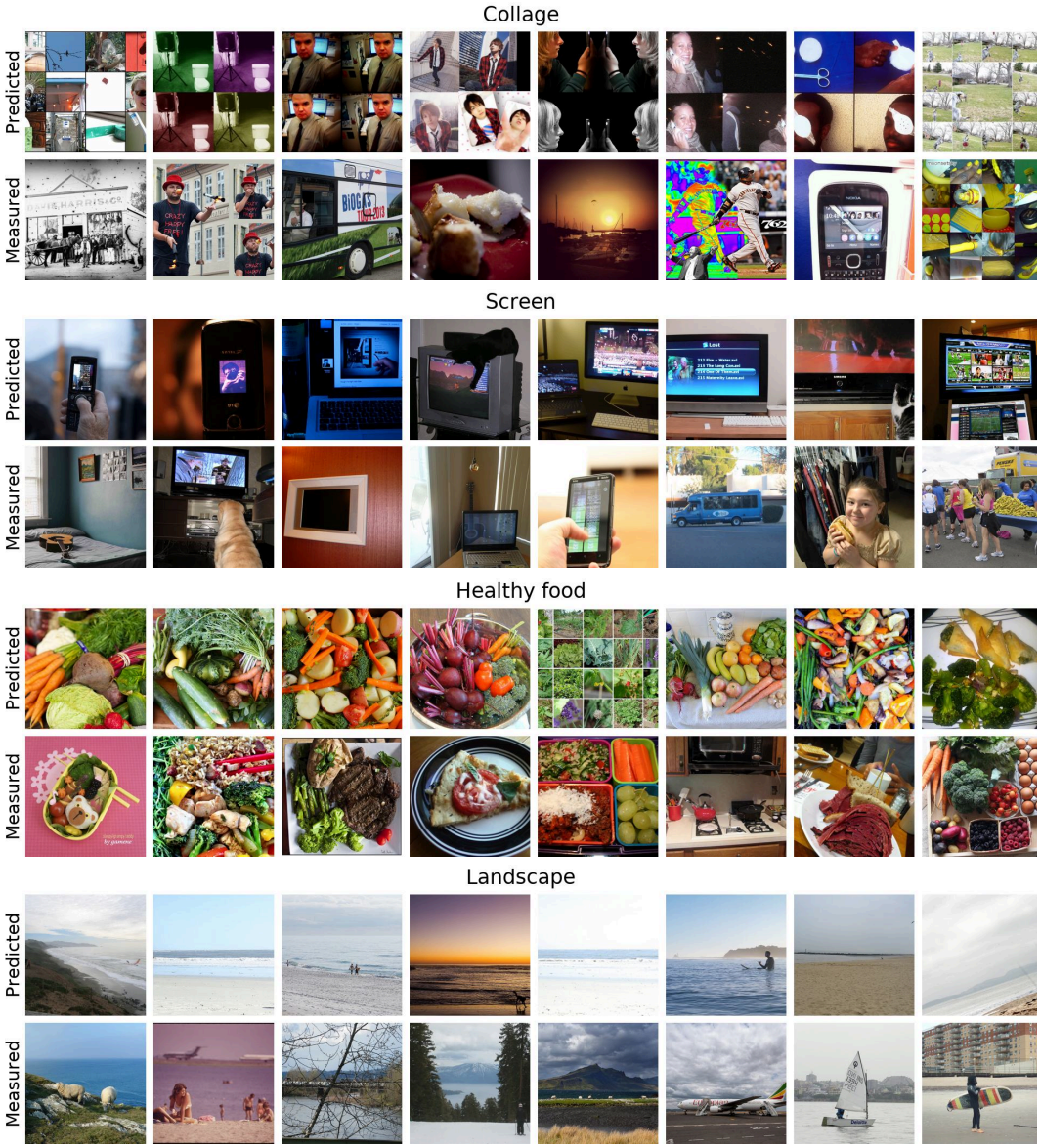


Figure S14: **Image grid PPA (Part 2).** Full visual grid with top 16 activating images for concepts *Collage*, *Screen*, *Healthy food*, and *Landscape* found in the PPA region. For each concept, the top 8 images corresponding to **Measured** and **Predicted** fMRI responses are shown.




Figure S15: **Concepts best explained by every ROI (Non-Exclusive)**. Concepts represented in each Region of Interest (ROI) that achieve an alignment score $> 0.5$. A concept may appear in multiple ROI word clouds, reflecting its representation across different brain regions. Word size reflects the alignment score of the concept within the assigned ROI.

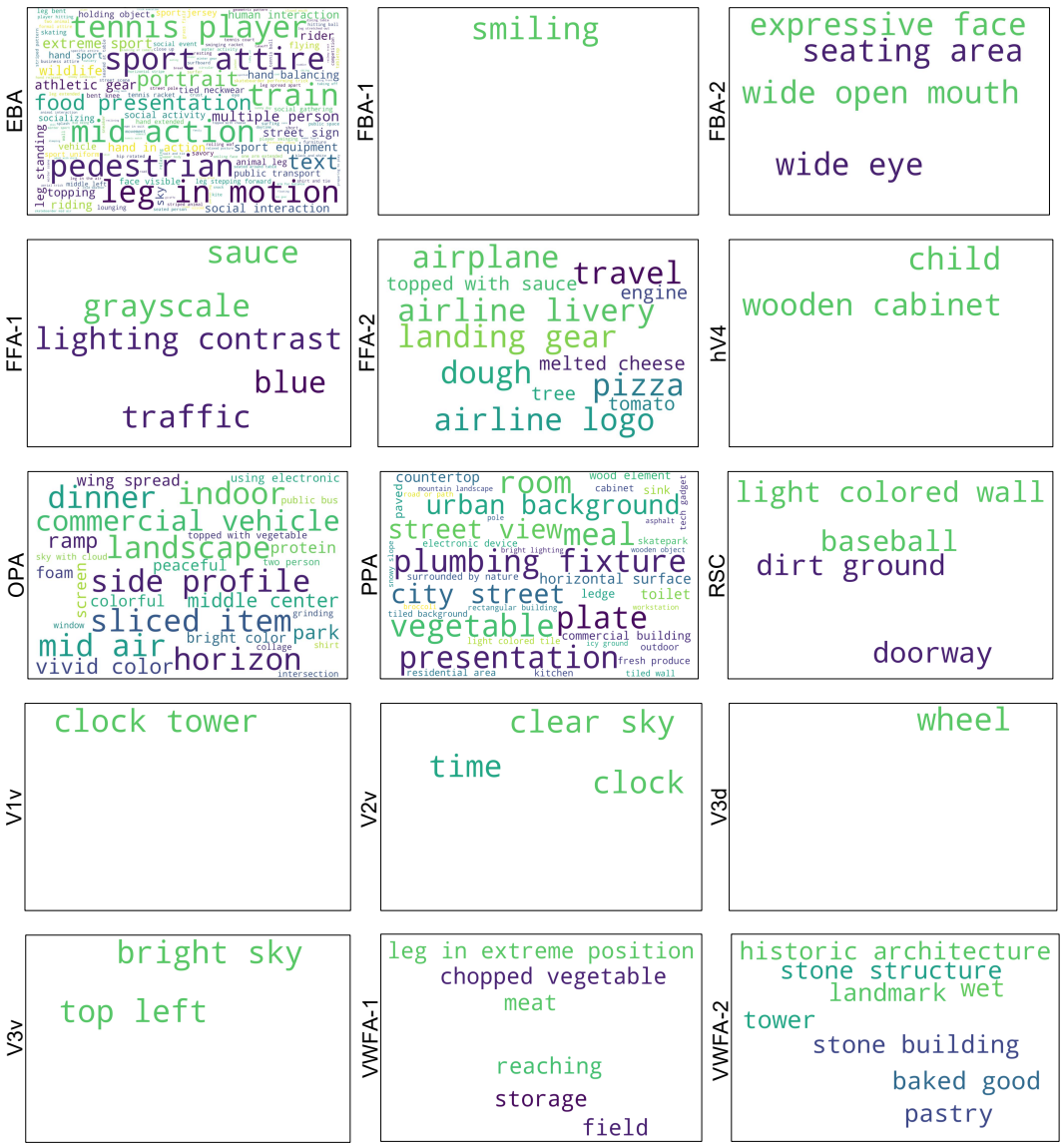


Figure S16: **Concepts best explained by every ROI (Exclusive)**. Each concept is assigned to a single ROI—the one with the highest alignment score. Only concepts with alignment $> 0.5$ are shown. Word size reflects the alignment score within the assigned ROI.

**Image Captioning Prompt**

```
You are given one image.

Output: return ONLY a single JSON object (no prose, no code fences). Keep all fields SHORT strings.
If unknown, use "" (empty string). Emit "shapes" only when clearly visible.

Schema:
{
"caption": "2–3 short sentences. Include key objects, activities, background/room, zoom, and any visible TEXT/LOGO with its color.",
"objects": "comma-separated items like: girl — smiling, holding apple; stove — metal, black, shiny",
"number_of_objects": "none | one | two | multiple | many | two identical <object>",
"room_type": "kitchen | bedroom | bathroom | living room | office | gym | none",
"scene_type": "indoor | outdoor | mixed | none",
"dominant_colors": "comma-separated color words",


"activity": "high/mid/low-level actions in a short phrase, e.g., cooking — stirring; sport/football — running; commute — walking",


"body_postures": "legs:<detailed specific posture>; hands:<- please detailed posture >; mouth:<detailed state>; in_air:<true/false>;
stretching:<true/false>",


"background": "very short phrase (e.g., ocean horizon; tile backsplash; stadium crowd)",


"shapes": "only if clear: straight lines — fence/train track/shore line; clear edges — plate edge/road curb; round — plate/ball/wheel",


"food": "if present, note food types/tableware/body interaction, e.g., desert, Asian food, bread type food, pizza, vegetables; tableware —
plate, fork; body — hand",


"main_object_location": "one of: left/right/center + up/middle/down (e.g., left down, center middle, right up)"
}
```

Figure S17: **Image Captioning Prompt.**

**Hypothesis Generation Prompt**

```
You are given structured captions for {N} images of the SAME SAE feature (its top activations).
Images are NOT ordered by significance.
Your goal: propose concise, shared hypotheses about what this feature likes.
Think at multiple levels and combine cues when common:
- Low-level: colors (≤2 shared colors), textures/materials, lighting, edges/lines, shapes (round objects, squared etc.)
- Mid-level: objects & types, counts (one/two/multiple), actions/verbs, body parts & POSTURES,
relationships (X on Y, X in Y), scene elements that co-occur
- High-level: style/design, scene type, overall mood/context

STRICT RULES
- Do NOT output the location "center middle" (it's too common). Other locations are allowed.
- Colors: if you include colors, list at most TWO (e.g., "blue", "white and black"), not more. If 2 colors are always together in all images
list them as one ( e.g. "black and white" for Zebra).
- Return between 3 and 12 hypotheses.
- Be BOTH general and nuanced: mix simple terms with specific combos when truly common (e.g., "airplane", "flying", "airplane flying"; "cows",
"grass", "cows on grass", "farm animals"; "person", "skateboard", "jumping", "skateboarder jumping").
- Don't stay only generic: if you see "pizza", also consider "round things"; if you see "sky", note what's IN the sky; for "toilet", also
notice "sink", "tiles", "counter", "wood".
- For sports/people: mention body posture and role (serving? running? legs crossed/bent/straight), hand vs leg sport, field/court type,
uniform/common wear, count of people, in-action vs static.
- You may infer obvious shared facts even if not spelled out in every caption (e.g., pizza is round; cows eat grass; zebra is black and white).
- No placeholders, no explanations, no evidence lists.
COUNTING RULE
- Only include a hypothesis if it appears in at least {support} of the {N} images (~{ratio}%).
FORMAT
- Phrases only, no sentences, no explanations, no evidence lists, no duplicates.
NOW USE THIS DATA:
CAPTIONS:
{captions}
First, silently think about the shared concepts across these images.
Then, OUTPUT (JSON only):
{{"hypotheses": ["...", "..."]}}
```

Figure S18: **Hypothesis Generation Prompt.** Given a set of 10 images and their detailed captions, an LLM is instructed to identify what is common across the images and to generate hypotheses that may explain what is shared among them.

**Image-Hypothesis Labeling Prompt 1**

```
You are given ONE image and ONE short visual concept (hypothesis).\n
Answer with EXACTLY one word: true or false (lowercase).\n\n
Be STRICT:\n
- true = clearly visible and unmistakably present.\n
- false = absent, ambiguous, weakly implied, too small/occluded/background to be sure, or inferred from text only.\n
- ACTIONS must be clearly occurring (e.g., 'washing hands' is false if only a sink is visible).\n
- VAGUE terms (e.g., 'play') are false unless the action is clearly happening, not just a toy present.\n
- COLORS: answer true only if this is a DOMINANT color region, not a tiny patch.\n\n
Return ONLY: true OR false.
```

**Image-Hypothesis Labeling Prompt 2**

```
You are given ONE image and ONE short visual concept (hypothesis).\n
This is a SECOND PASS to confirm a previous 'true' decision. Be EXTRA STRICT.\n
Answer with EXACTLY one word: true or false (lowercase).\n\n
Rules (even stricter than before):\n
- true = the concept is CLEAR, DOMINANT, and UNAMBIGUOUS in the image.\n
- false = any ambiguity, partial/occluded/tiny evidence, inference from context,
"or near-match that does not exactly satisfy the concept.\n
- For actions: the action must be obviously occurring (not implied by objects).\n
- For colors: only if the color is clearly dominant (not a small patch).\n\n
Return ONLY: true OR false.
```

Figure S19: **Image–Hypothesis Labeling Prompt.** For each image–hypothesis pair, we label whether the hypothesis is supported by the image or not.




Figure S20: **Images with per-hypothesis labeling.** We show six example images and list all hypotheses labeled as positive for each. The hypotheses are taken from our brain-inspired hypothesis dictionary, and the images are drawn from the set used with predicted fMRI.