SCALABLE INTERPRETABLE MAPPING OF HUNDREDS OF VISUAL CONCEPTS IN HUMAN BRAIN ACTIVITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Systems neuroscience and mechanistic interpretability share a central challenge: identifying interpretable components within high-dimensional, distributed representations. Recently, sparse autoencoders (SAEs) have been shown to yield rich concept dictionaries that aid in the interpretation of deep embeddings in vision foundation models. This suggests that SAEs might be a similarly powerful tool for interpreting brain activity patterns in the visual system, where a relatively small set of concept maps has been established by decades of dedicated functional localization experiments. However, limited samples and measurement noise complicate the naive application of SAEs to brain recordings. Here we use SAEs to perform scalable, interpretable mapping of visual concepts in human brain activity by following an SAE dictionary derived from a vision foundation model through a fitted voxelwise fMRI encoder. Correlating these maps with measured brain activity retrieves recognizable stimulus sets, including faces, surfers, black-and-white photographs, and tiled montages. Comparison with functional localizers supports consistency with known cortical organization, while measures of visual coherence and model-brain correspondence reveal preservation of feature identity through the fitted mapping. The same maps support intervention as well as interpretation: off-the-shelf brain decoders turn them into concept-reactivating images, and adding them to recorded neural responses produces targeted changes in reconstructed content and appearance. Across large-scale steering experiments, specific target-feature reactivation increases far beyond random map perturbations. We deliver hundreds of concept maps for the visual cortex, expanding the vocabulary of selectivity and validating their interpretation through complementary methods from neuroscience and mechanistic interpretability.

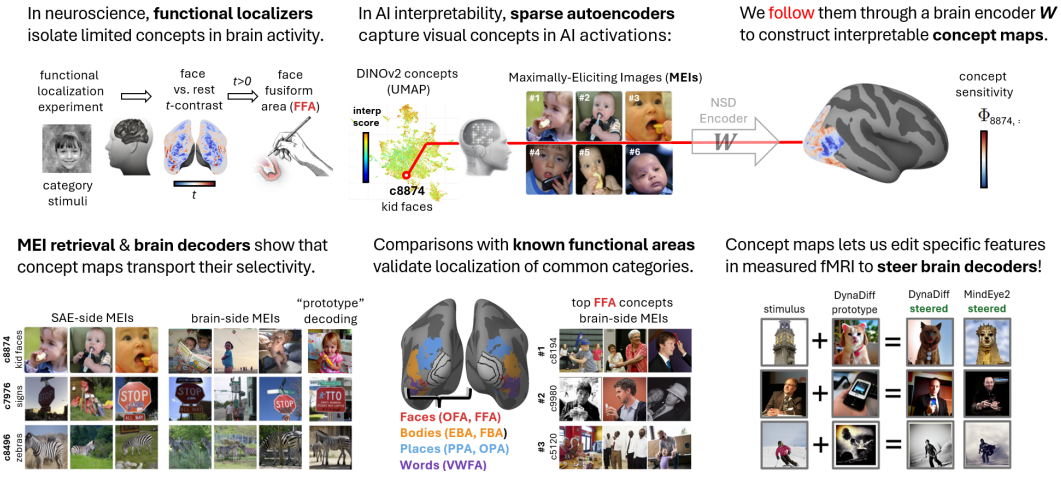

Figure 1: An overview of our framework: concept maps recover established functional localization (e.g. "faces" broadly) at greater granularity ("kid faces", "animal faces", "object held near face") while capturing novel properties like "montage tiling", "black-and-white photography" and "pizza."

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.

1 INTRODUCTION

Systems neuroscience and mechanistic interpretability confront a shared central question: how can distributed representations be understood in terms of interpretable components (Elhage et al., 2022; Olshausen & Field, 1996; Klindt et al., 2026)? In visual neuroscience, the answer has long been pursued through properties of eliciting stimuli, ranging from the orientation of an edge to the presence of a face (Hubel & Wiesel, 1962; Kanwisher et al., 1997; Epstein & Kanwisher, 1998). Functional localizers are experiments designed to elicit stimuli properties that reveal selective cortical regions in brain recordings, enabling comparisons across stimuli, tasks and models. The resulting cortical delineations, or regions of interest (ROIs), provide a shared vocabulary for studying visual representations, and often come pre-packaged with large-scale neural datasets. However, the familiar categories of faces, bodies, places, words, and food cover only a fraction of the distinctions present in natural images. Expanding this vocabulary requires discovering properties that are difficult to pre-specify or whose representations extend across pre-packaged regional boundaries.

Mechanistic interpretability approaches the same problem by searching for features or "concepts" within the distributed activations of deep neural networks (Bricken et al., 2023). Sparse autoencoders (SAEs) approximate these activations as sparse combinations of learned atoms in a dictionary $\mathbf{D}$, with unit directions $\mathbf{d}_i$ (Gao et al., 2024). Their semantic selectivity can be examined through preferred stimuli, downstream recruitment (Fel et al., 2026), and steering interventions (Wurgaft et al., 2026). In vision models, such dictionaries supply thousands of candidate eliciting properties without requiring a pre-defined semantic label for each one.

We construct distributed cortical "concept maps" by following concepts from a vision model dictionary through a fitted brain encoder. Starting with a dictionary $\mathbf{D}$ learned from DINOv2 activations and a linear encoder with weights $\mathbf{W}$ mapping the same feature space to voxel responses β from the Natural Scenes Dataset (Allen et al., 2021), we propagate each feature direction through the encoder, weighting its spatial contributions by expected feature activation $\bar{z}$ over a reference image distribution. Each resulting signed pattern over the visual cortex forms a row of our central object, the recruitment matrix Φ, where each row $\Phi_{i,:}$ retains an explicit indexed link to the source concept in feature space. This construction adapts downstream concept recruitment analysis (Fel et al., 2026) to brain responses, making a large vocabulary available across the encoder's cortical coverage without a separate text query, stimulus category, or ROI partitioning.

Each map acts as a hypothesis whose interpretation must be validated against brain measurements. We verify whether the concept maps in Φ capture genuine cortical selectivity by exploring several questions. First, do the concepts remain interpretable through the brain encoder? By combining paired Maximally Eliciting Image (MEI) sets for model and brain with automated perceptual measures (Issa et al., 2026; Fu et al., 2023), we reveal strong visual coherence on both the model and brain sides independently, making a large vocabulary of cortical selectivity accessible to inspection (Section 4.1). Second, is concept identity preserved through brain encoders and decoders? Exact image overlap and perceptual correspondence between paired MEI sets distinguish features whose meanings survive the mapping from coherent features whose stimulus preferences change, making semantic preservation itself measurable (Section 4.2).

Third, do these maps recover established functional organization? Correspondence with independent localizers for faces, bodies, places, and characters grounds the concept maps in the functional-localization literature, with concepts selected in one subject retaining their functional alignment in another (Section 5). Fourth, can the maps support visualizing and steering concepts through "off-the-shelf" brain decoders? Prototype images give the maps a decoded visual expression, while controlled steering increases target-feature reactivation beyond random-map perturbations, connecting the inferred selectivity to its effects on reconstructed images (Section 6).

We deliver hundreds of interpretable, distributed concept maps for the visual cortex, together with the framework for constructing (Section 3) and examining them (Section 4– 6). These maps capture finer distinctions than the familiar localizer categories: faces interacting with handheld objects, locations like train stations, bodies in running motion, and wordforms on red stop signs, alongside black-and-white photography and collage-like compositions. Each map retains its indexed link to the source SAE feature, paired MEI sets, cortical distribution, and decoded visualizations, bringing complementary selectivity evidence together in a reusable resource. Our contribution covers a richer vocabulary of visual concepts and practical framework to discover, inspect, and validate them.

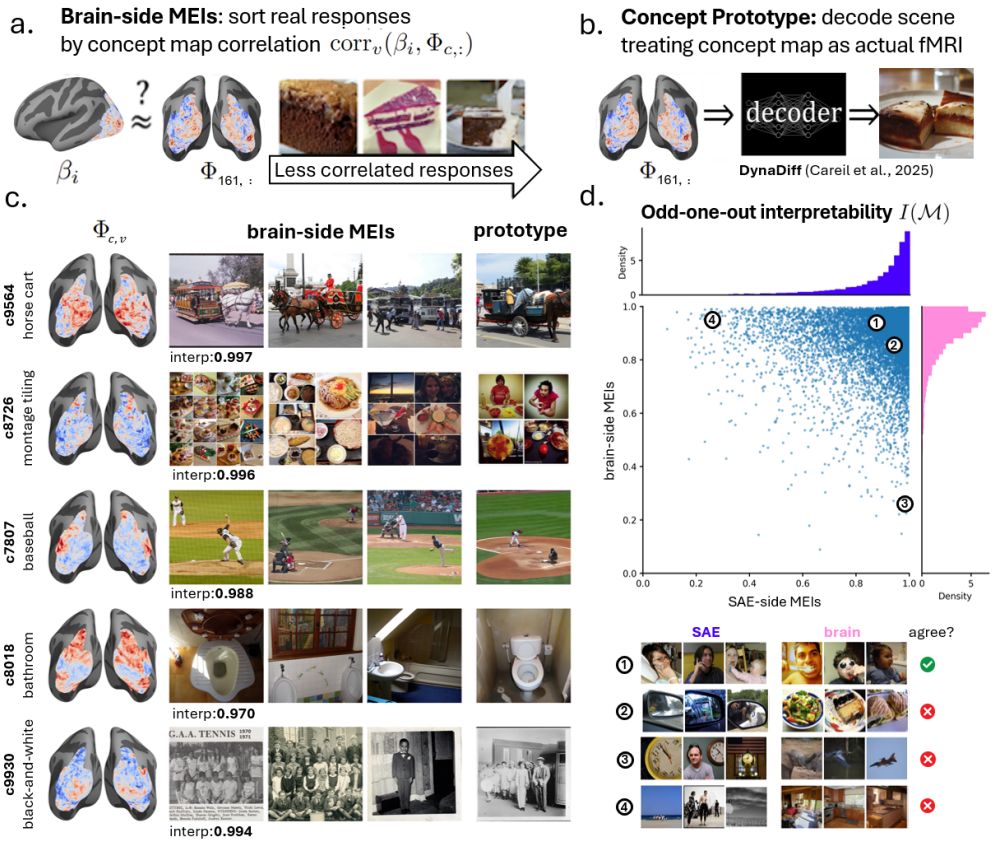

Figure 2: **Concept maps capture interpretable semantics.** (a.) We correlate neural responses to concept maps to find brain-side Maximally-Eliciting Images (MEIs). (b.) Treating concept maps as measured fMRI enables direct visualization via brain decoders. (c.) Diverse sample of interpretable examples; from left to right, columns show $\Phi_{i,:}$ concept maps, their brain-side MEI sets with their interpretability scores, and their decoded prototypes. (d.) Interpretability I scores for each concept's paired sets of $K = 8$ images. The vast majority of MEI sets have high interpretability on both the SAE- and brain-side. Concepts may become more or less coherent through the fitted encoder, and model-brain MEIs sets may not "agree" but instead reflect coherent yet different concepts.

2 Related Work

Functional localization and its vocabulary. Category-selective regions have provided durable experimental targets in visual neuroscience, including the fusiform face area (Kanwisher et al., 1997), parahippocampal place area (Epstein & Kanwisher, 1998), extrastriate body area (Downing et al., 2001), and visual word form area (Cohen et al., 2000), with food selectivity a recent addition (Khosla et al., 2022; Jain et al., 2023). Standardized localizers make these regions comparable across studies and available in major datasets (Allen et al., 2021), but the resulting vocabulary remains tied to a limited set of experimental contrasts requiring dedicated experiments. Recent approaches broaden selectivity discovery by synthesizing activating stimuli (Luo et al., 2023; Gu et al., 2022), describing voxel tuning in language (Luo et al., 2024; Matsuyama et al., 2026), or grounding it in spatial features (Luo et al., 2025; Sarch et al., 2023). We contribute distributed concept maps derived from a model dictionary, extending the vocabulary of selectivity while testing correspondence with established localizers in the visual domain.

Mechanistic interpretability and concept dictionaries. Sparse autoencoders recover candidate feature directions from distributed activations (Bricken et al., 2023; Huben et al., 2024) motivated by accounts of superposition in which features occupy overlapping directions (Elhage et al., 2022). Their use for concept-level analysis and steering (Templeton et al., 2024) extends to vision dictio-

naries that capture materials, lighting, and implied motion alongside object and scene categories (Thasarathan et al., 2025). Particularly relevant is the analysis by Fel et al. (2026) of how downstream linear readouts recruit dictionary features, which we adapt to cortical response prediction by treating brain responses as a "probe." These dictionaries are nonetheless hypotheses about representation: their features can depend on random seed (Paulo & Belrose, 2025) and only partially capture more complex latent structure (Bhalla et al., 2026; Fel et al., 2026). We therefore examine which feature interpretations survive brain mapping, and note that an analogous Φ could be constructed from any dictionary basis compatible with a neural encoder.

Discovering and testing brain concepts. Recent years have seen several attempts at scaling concept discovery on measured brain responses. SAE-BrainMap associates model features with voxels through activation similarity (Mao et al., 2025), whereas BrainExplore decomposes fMRI responses with ROIs to discover interpretable features (Wasserman et al., 2025). Related dictionary approaches operate over word embeddings (Zeng & Gallant, 2025), contextual language states (Lepori et al., 2026) and CNN representations (Longon et al., 2026). Brain Algebra obtains decoder-intervention directions by averaging responses to images selected through text-image similarity using specified prompt-categories (Ferrante et al., 2025). Our maps derive from a non-linguistic model dictionary and independently fitted encoder, retaining an explicit link to each source concept without requiring feature-specific text queries or ROI partitions. This link allows stimulus preference, cortical distribution, and decoder effects to be examined together, and makes changes in interpretation through the fitted encoder and decoder mappings available for quantitative study.

Encoding and decoding as tools for interpretation: Encoding models relate stimuli to measured responses, while decoding models recover stimulus information from activity (Naselaris et al., 2011). Earlier "hand-engineered" feature spaces, such as Gabor pyramids (Kay et al., 2008) and motion energy (Nishimoto et al., 2011) made explicit hypotheses about neural selectivity; deep representations offer strong natural-scene prediction (St-Yves et al., 2023) but less direct access to the properties responsible. The observation that divergent models can achieve similar brain predictivity motivates analysis beyond predictive fit (Storrs et al., 2021). In parallel, generative decoders reconstruct recognizable image content and appearance from fMRI (Scotti et al., 2024; Careil et al., 2025), providing a way to inspect the effects of perturbing a response representation. We leverage linear encoding models to construct concept maps, and generative decoder models to isolate their selectivity.

3 Concept recruitment through the brain encoder-as-probe

Figure 1 summarizes the construction of Φ and its use in image retrieval, prototype decoding, and steering. We construct the Φ matrix by asking which concepts in the SAE dictionary are recruited by our fitted ridge encoder to predict voxel responses. Following Fel et al. (2026), we measure concept recruitment by weighting the alignment between dictionary directions and a downstream linear readout by expected concept activation. We adapt their construction to an encoder in which the entire embedding field produces a scalar response for each voxel, retaining location-specific weights across patch positions. For model patch embeddings $\mathbf{a}_s \in \mathbb{R}^d$, the encoder predicts $\widehat{\beta} = \mathbf{b} + \sum_s \mathbf{a}_s^\top \mathbf{W}_s$, where $\mathbf{W}_s \in \mathbb{R}^{d\times V}$ is the position-s block of $\mathbf{W}$, mapping to V voxels, and $\mathbf{b}$ is the intercept. The SAE supplies sparse coefficients $z_{s,c}$ over C dictionary atoms $\mathbf{D} \in \mathbb{R}^{C\times d}$, learned in standardized embedding coordinates with feature standard deviations $\sigma \in \mathbb{R}^d$. The unit atom direction is $\mathbf{d}_c = \mathbf{D}_{c,:}/\|\mathbf{D}_{c,:}\|_2$; the construction retains the atom's learned scale.

Defining $\bar{z}_{s,c} = \mathbb{E}_{\text{ref}}[z_{s,c}]$ over a reference image distribution, we construct the concept recruitment matrix $\Phi \in \mathbb{R}^{C\times V}$ tracking the contribution of each concept c to voxel v as

$$\Phi_{c,v} = \sum_s \bar{z}_{s,c} \cdot \left[(\mathbf{D}_{c,:} \odot \sigma)\mathbf{W}_s\right]_v . \tag{1}$$

Each row $\Phi_{c,:}$ is a signed cortical "concept map": the exact change in the linear encoder's prediction induced by the reference-weighted concept delta field $\bar{z}_{s,c}(\mathbf{D}_{c,:} \odot \sigma)$ across patch positions.

Φ interface for Natural Scenes Dataset (NSD). We instantiate the above construction with the final normalized patch embeddings of DINOv2 ViT-B/14 (Oquab et al., 2024) for ImageNet (Deng et al., 2009): each image supplies $S = 256$ patch embeddings of dimension $d = 768$. Our fixed SAE

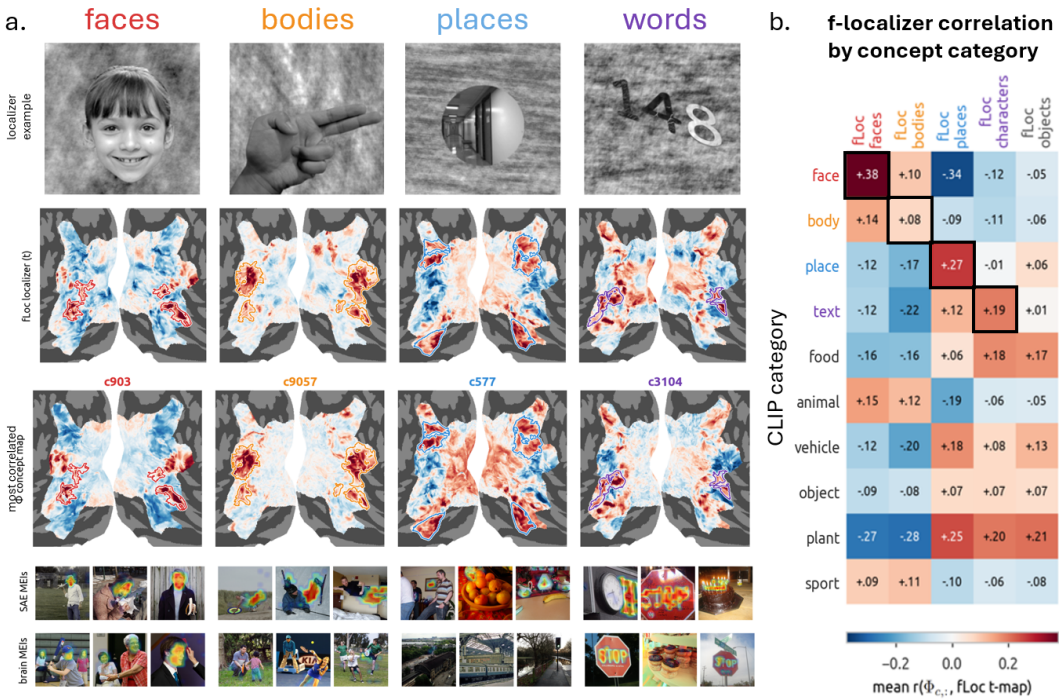

Figure 3: **Concept maps recover established localization.** (a) Stimuli examples from NSD localizer experiments, resulting t-contrasts and ROI, most-correlated Φ (surface) concept map, and paired MEIs with the concept's SAE patch activations. Per-ROI concept rankings reveal appropriate specificity. (b) CLIP concept category by localizer correlation matrix. We construct independent categories using CLIP labels, then group concepts in each category and correlate them with localizer contrasts, showing that the concept maps align with established ROIs for the appropriate selectivity .

dictionary contains $C = 10,000$ atoms, with top-100 sparsity applied independently at each patch. We flatten dense embeddings across positions and fit a ridge encoder to trial-averaged fMRI betas from the Natural Scenes Dataset (Allen et al., 2021) or NSD, restricting its outputs to the V voxels in the `nsdgeneral` mask of voxels responsive to the NSD experiment. We include an extensive analysis of our SAE, its training and concept quality in Appendix B.

Unless otherwise specified, we display results for NSD Subject 1 ($V = 15{,}724$). For each subject, 8,000 images determine the encoder weights and reference activity $\bar{z}$; 1,000 additional development images select a single subject-wide ridge penalty from 20 logarithmically spaced values between 10^{-2} and 10^{8}. We do not refit on the combined 9,000-image "development pool". Response normalization uses fitting-only statistics, and encoder weights are restored to beta units (accounting for normalization for our encoder inputs) before computing Φ. Mean voxelwise prediction correlations on the shared 1,000-image test pool are 0.372 for Subject 1 and 0.377 for Subject 2 ($V = 14,278$). The shared images are held out from encoder fitting and penalty selection. Additional encoder details and comparisons to non-linear encoders are provided in Appendix A.

4 Do model features retain their interpretable identity through encoders and decoders of brain activity?

If Φ captures genuine cortical selectivity, then measured fMRI responses should resemble concept profiles for concepts present in the viewed image. We characterize each concept through paired sets of **Maximally Eliciting Images (MEIs)**, retrieved from a common pool $\mathcal{I}$ of stimuli with measured responses. We use the 9,000-image development pool for discovery and the shared 1,000-image

pool for all of the quantitative MEI results reported below. Concepts maps that underflow, zero out or do not produce enough MEIs are not eligible and thus are not counted in the following analyses.

On the SAE side, the max-match score for image i is its largest activation of concept c across embedding patches, $m_c(i) = \max_s z_{s,c}(i)$. On the brain side, its score is the spatial Pearson correlation $q_c(i) = \mathrm{corr}_v(\beta_i, \Phi_{c,:})$, computed across the same `nsdgeneral` voxels using the image's response vector β_i. Ranking images by these scores gives the top-K sets $\mathcal{M}_c^{\mathrm{SAE}}$ and $\mathcal{M}_c^{\mathrm{brain}}$, respectively. These paired sets connect each dictionary feature to the stimuli preferred by its map, and form the basis of the following analyses. Quantitative MEI comparisons use $K = 8$ unless noted otherwise.

4.1 Concept interpretability is strong on both sides of the alignment

We assess whether the images retrieved by each concept map share a recognizable visual property. Paired SAE- and brain-side MEIs and cortical maps support qualitative inspection. Figure 2c pairs model-side and brain-side MEIs revealing familiar and surprising features, including horses pulling carts, montages that tile the stimuli, baseball fields, and black-and-white photography.

We quantify MEI perceptual coherence using the **odd-one-out interpretability** measure of Issa et al. (2026) as a proxy for human interpretability. This metric uses DreamSim embeddings as a perceptual distance function given its reported correspondence with human judgment (Fu et al., 2023). Let $\mathbf{e}_i$ be the unit-normalized DreamSim embedding of image i, with cosine similarity $s_{ij} = \mathbf{e}_i^\top \mathbf{e}_j$. For a set $\mathcal{M}$ of K images, we compute its mean pairwise similarity excluding self-pairs, then the fraction of other pool images whose mean similarity to the set falls below it:

$$w(\mathcal{M}) = \frac{1}{K(K-1)} \sum_{\substack{i,j \in \mathcal{M} \\ i \neq j}} s_{ij}, \quad I(\mathcal{M}) \quad = \frac{1}{N-K} \sum_{j \in \mathcal{I} \setminus \mathcal{M}} \mathbf{1}\left[\frac{1}{K} \sum_{i \in \mathcal{M}} s_{ij} < w(\mathcal{M}) \right]. \tag{2}$$

We evaluate I separately for both MEI sets of every eligible concept using $K = 8$. Higher scores indicate that the retrieved images resemble one another more than they resemble other pool images. Figure 2d compares SAE-side and brain-side scores. At $K = 8$, median coherence is **0.9294** for SAE MEIs and **0.9123** for brain MEIs; 7,921/9,974 (79.42%) eligible SAE concepts and 8,384/9,994 (83.89%) eligible brain concepts score at least 0.8; brain coherence exceeds SAE coherence for 4,537/9,974 paired concepts (45.5%). Next, we test whether concepts with high SAE- and brain-side interpretability yield MEI sets that express similar meanings.

4.2 Concept identity is mostly preserved but sometimes transformed

To further explore this issue, we quantify MEI correspondence by introducing two intuitive metrics over paired MEI sets of size K. The (unnormalized) **Top-K overlap** $O_c(K)$ is the number of images shared by the two sets, measuring strict agreement in stimulus preference; for a perceptual measure, we define $P_c(K)$ as **pairwise DreamSim similarity**:

$$O_c(K) = |\mathcal{M}_c^{\mathrm{SAE}} \cap \mathcal{M}_c^{\mathrm{brain}}|, \qquad P_c(K) = \frac{1}{K^2} \sum_{i \in \mathcal{M}_c^{\mathrm{SAE}}} \sum_{j \in \mathcal{M}_c^{\mathrm{brain}}} s_{ij}. \tag{3}$$

This allows for the same visual property to be expressed by different images. Figure 4b shows the distributions of top-K overlap $O_c(K)$ and mean pairwise DreamSim similarity $P_c(K)$ for the 9,974 eligible concepts, as well as chance-level values. Through inspection, and given the extremely low chance-level floors, we find that an overlap as low as $O_c(50) = 5$ (chance: 2.54) can signal consistent concept identity.

Figure 4c shows paired MEI sets for concepts ranked by these alignment scores, with dashed lines in Figure 4b showing the place in the distribution at which the images were sampled. Alignment is strong at high $O_c(50) \in [5, 19]$ overlap ("pizza toppings", "airplanes"), but inconsistent below that ($O_c(50) = 2$; "horses running" turns to "people running") and clearly disjoint with no overlap ($O_c(50) = 0$, food turns to "car mirrors").

Mean top-8 intersection is 0.2508 images, corresponding to mean $O_c(8) = 0.03134$ versus the independent uniform-set expectation of 0.008; 7,905 concepts share no image, and mean $P_c(8)$ is

0.2112 versus an independent-set expectation of 0.1829. The restricted visual semantics of MS COCO limit the stimuli available to express each concept; Appendix E provides exploratory results on LAION-fMRI's broader stimulus distribution (Zerbe et al., 2026).

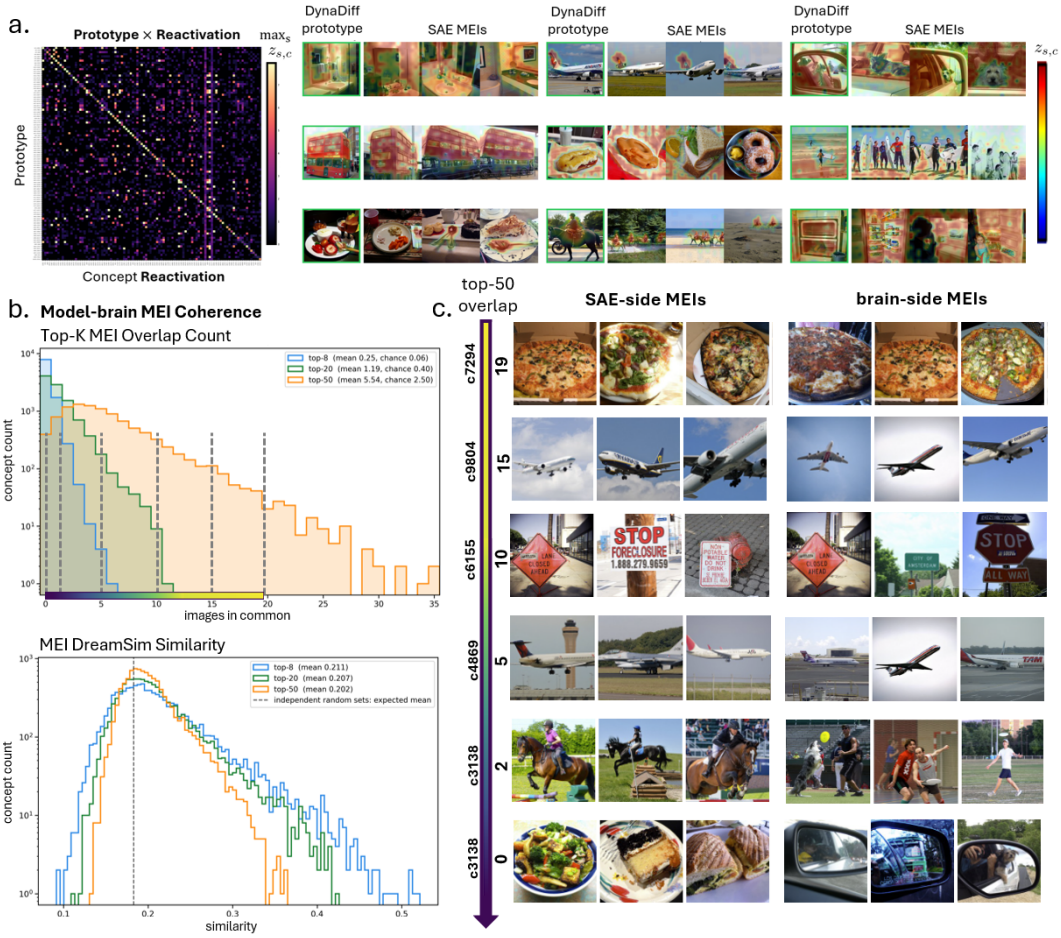

Figure 4: **Concept maps preserve the SAE selectivity**. (a) Prototype reactivation matrix, cell (i, j) measures the activation of concept j in the decoded prototype i; strong diagonal structure illustrates strong selectivity, and MEI comparisons with patch-wise SAE heatmaps make visual similarity apparent. (b) Model-brain coherence metrics. Top-K overlap counts identical images across MEI sets, while DreamSim similarity captures perceptual alignment; both reveal hundreds of visually-aligned concepts far above their respective chance-nulls. (c) Paired MEIs ranked by Top-50 overlap. Concepts with as little overlap as 5 appear coherent through the model-brain mapping.

We also visualize concept identity through a brain decoder by reconstructing concept maps into image prototypes with DynaDiff (Careil et al., 2025) following the approach in Section 6. Figure 4a shows various prototypes paired with SAE MEIs, and source-feature activation maps. Concept prototypes convincingly recreate the interpretable semantics captured by the SAE MEIs. Constructing a prototype-by-concept reactivation matrix for this set shows that prototype i selectively activates concept i over $j \neq i$, leading to a highly visible diagonal structure.

5 Do concept maps recover known functional organization?

If Φ captures genuine cortical selectivity, then it should recover established functional organization. We compare concept maps with NSD functional-localizer contrasts for faces, places, characters, bodies, and objects, projecting the maps onto each subject's native cortical surface and retaining only vertices fully supported by `nsdgeneral` voxels. These continuous t-contrast maps supported

the thresholding and manual delineation of the fusiform face area (FFA) and body area (FBA), parahippocampal place area (PPA), occipital place area (OPA), visual word form area (VWFA), and extrastriate body area (EBA), whose ROI boundaries provide anatomical references for the maps.

To assess functional organization across the dictionary, we assign each concept to one of ten categories using `CLIP ViT-L/14`: face, body, place, text, food, animal, vehicle, object, plant, or sport. For each concept, we average unit-normalized CLIP embeddings of up to six highest-activating SAE MEIs from the 73,000-image NSD collection. Each category is represented by the average of four unit-normalized text embeddings, using template prompts such as "a human face" and "a portrait of someone's face"; all prompts and implementation details in Appendix F. We filter out concepts with $\|\Phi_{c,:}\|_2$ below dictionary median. Figure 3b shows correspondence between these image-derived categories and established functional selectivity: each category correlates highly with its localizer.

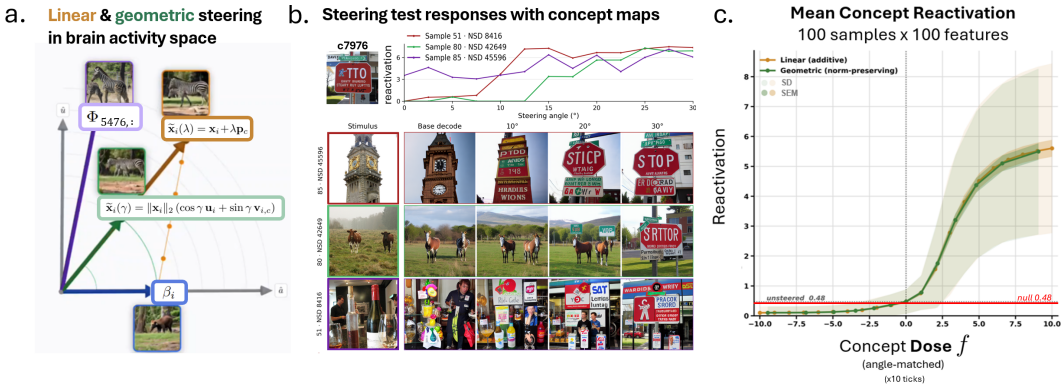

Figure 5: **Targeted semantic image editing by steering brain activity.** (a) A diagram illustrating linear and geometric steering. Linear steering adds the concept direction $\Phi_{c,:}$ to the fMRI-as-vector β_i, while geometric steering rotates β_i towards $\Phi_{c,:}$ while retaining its norm. (b) Target-feature reactivation curves when editing test responses. We add c7976 "stop signs" to different scene contexts, noting smooth visual interpolation that tracks reactivation across decoded steps. (c) We quantify this process at scale over 210,000 generations, showing consistent target reactivation.

6 Do concept maps support visualizing concepts through brain decoders? Do they support "steering" brain activity?

If Φ captures genuine cortical selectivity, then steering fMRI along a concept should amplify or suppress that concept in a reconstruction. We test this through changes in decoded appearance and source-feature activation. Our quantitative experiment uses NSD Subject 1 and DynaDiff (Careil et al., 2025). We express each map in the decoder's voxel order and scale it to the decoder's input units, obtaining a direction $\mathbf{p}_c$. For DynaDiff, the spatial map steering is repeated identically across the six input time points. Decoding the scaled map alone produces a concept prototype (as in Figure 2a); steering instead modifies a measured decoder input $\mathbf{x}_i$. We test additive steering, $\widetilde{\mathbf{x}}_i(\lambda) = \mathbf{x}_i + \lambda \mathbf{p}_c$, and spherical steering, which rotates the input toward the concept direction while preserving its norm. For additive steering, the target map is scaled so its largest absolute voxel value equals a decoder-input scale estimated from 200 fitting-pool contexts. With the decoder input and direction flattened into vectors, define $\mathbf{u}_i = \mathbf{x}_i / \|\mathbf{x}_i\|_2$ and rotate $\mathbf{u}_i$ towards $\mathbf{p}_c$ by setting

$$\mathbf{v}_{i,c} = \frac{\mathbf{p}_c - (\mathbf{p}_c^\top \mathbf{u}_i)\mathbf{u}_i}{\|\mathbf{p}_c - (\mathbf{p}_c^\top \mathbf{u}_i)\mathbf{u}_i\|_2}, \qquad \widetilde{\mathbf{x}}_i(\gamma) = \|\mathbf{x}_i\|_2 \left(\cos\gamma\, \mathbf{u}_i + \sin\gamma\, \mathbf{v}_{i,c}\right), \tag{4}$$

We vary the alignment fraction $f \in [-1, 1]$, with $\gamma = f \arccos(\mathbf{u}_i^\top \widehat{\mathbf{p}}_c)$ and $\widehat{\mathbf{p}}_c = \mathbf{p}_c / \|\mathbf{p}_c\|_2$. At $f = 0$ the input is unchanged; at $f = 1$ it points along the concept map while retaining the measured input's total norm across voxels, at each time point. We decode each perturbed response and measure target max-patch SAE activation relative to its unsteered reconstruction.

Figure 1 shows selected examples of "brain editing" test samples using spherical steering across DynaDiff and MindEye 2 (Scotti et al., 2024). The visual changes in the scene appear smoothly;

5b shows the result of steering DynaDiff by steering with a “stop sign" feature; visual onset of the concept tracks the reactivation curve across test samples. We quantify both steering methods further by using 100 concepts sampled from the 300 highest-overlap features in the development pool and 100 brain responses (50 from train set, 50 from subject-shared test), evaluated at 21 fractions from -1 to 1. This gives 210,000 SAE-scored dose observations per condition, using the same generation settings as Careil et al. (2025). A paired control steers a *different* usable dictionary map, sampled per feature-context pair, while measuring the original target feature, and produces a flat null. Note that target and control share alignment fractions but not necessarily rotation angles.

Figure 5c shows the grand mean dose response across features and contexts showing steady increase on average across. At full alignment ($f = 1$), held-out mean target activation increases by 6.613 SAE units versus -0.0016 under random-map steering, and 99.12% of paired target-minus-control differences are positive. After standardizing each candidate by its activation mean and standard deviation over the 50 unsteered test reconstructions, the target ranks in the top 10 of the frozen 100-feature panel in 94.0% of cases and first in 44.6%. Mean DreamSim similarity to the unsteered reconstruction falls to 0.190 showing the tradeoff between target reactivation and content preservation. Spherical reactivation curves for development and test contexts see similar gains at full alignment (6.629 versus 6.613) but differ at intermediate strengths (1.205 versus 2.535 at $f = 0.2$). Steering therefore increases reactivation in aggregate, but its strength and specificity vary across concepts and contexts.

7 Discussion

We deliver hundreds of interpretable, distributed concept maps by following a vision model dictionary through a fitted brain encoder. The paired MEIs show strong perceptual coherence on both sides of the mapping, while overlap and perceptual similarity reveal strong identity preservation for hundreds of concepts. Correspondence with independent localizers connects these maps to established cortical organization. Connecting inferred selectivity via prototype decoding and controlled steering provides a complementary view that is both visually compelling and measurable: the correct map-feature pairing increases target reactivation far beyond random-map perturbations. Together, these results expand the vocabulary of visual selectivity and make the preservation of feature meaning part of the mapping problem itself, a tool that can be useful for interpreting encoders & decoders.

A further open question is how to turn a signed, distributed concept map into an ROI with a reproducible boundary. NSD’s category-localizer ROIs themselves combine a liberal $t > 0$ threshold with boundaries drawn manually by inspection (Allen et al., 2021). Their boundaries provide anatomical context rather than a unique rule for converting concept maps in Φ into discrete regions.

Interpretation also depends on the images available to express a concept and the measures used to assess it. DreamSim coherence captures perceptual consistency, but numerical “sufficiency" thresholds remain prescriptive and lack systematic psychophysics validation in this context. Section 4.2 shows that a source feature’s semantic label cannot be assumed to transfer, motivating care when interpreting model-side features through neural mappings. NSD’s natural scene distribution also limits which distinctions can be tested. Human judgments and broader, controlled stimulus sets would help distinguish feature identity from correlated visual properties.

The present analyses use one frozen SAE dictionary and two subjects and restricted coverage. Albeit we report summary results for generalization across feature spaces like DINOv3 and CLIP (Appendix C), NSD Subject 2 (Appendix D), and experiments like LAION-fMRI (Appendix E), our thorough psychophysics inspection and quantitative experiments were limited to DINOv2 and Subject 1 of NSD. As with all *in silico* discovery, future work must pursue experiments with human subjects in order to decidedly characterize reproducibility.

The resulting maps connect model concepts, measured cortical responses, established functional organization, and decoded visual effects in a single framework. Through these concept maps and framework, we provide a starting point for functional localizers with a much larger vocabulary, and for testing which visual distinctions are preserved or transformed across models and brains. More broadly, Φ applies whenever a brain encoder or decoder can be composed with an embedding dictionary, providing an interpretable handle that lets us “read" and “write" the visual cortex.

AI USE STATEMENT

In this work, we used generative AI tools to assist in writing software like training loops of visualizations. We have not used generative AI tools to Generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. Additionally, we used generative AI tools to catch spelling or grammar errors. We have reviewed all AI-assisted work. LLM-generated code was verified by the authors and we take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

In order to ensure that this work is reproducible, we have taken the following steps. In Appendix A and B we provide a complete description of the datasets, model architectures, training procedures, hyperparameters, and evaluation to accompany and clarify our descriptions of methods.

REFERENCES

Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, December 2021. ISSN 1546-1726. doi: 10.1038/s41593-021-00962-x. URL http://dx.doi.org/10.1038/s41593-021-00962-x.

Usha Bhalla, Thomas Fel, Can Rager, Sheridan Feucht, Tal Haklay, Daniel Wurgaft, Siddharth Boppana, Matthew Kowal, Jack Merullo, Atticus Geiger, and Ekdeep Singh Lubana. Do sparse autoencoders capture concept manifolds? In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=euoAoOflaN.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Marlène Careil, Yohann Benchetrit, and Jean-Rémi King. Dynadiff: Single-stage decoding of images from continuously evolving fmri, 2025. URL https://arxiv.org/abs/2505.14556.

Laurent Cohen, Stanislas Dehaene, Lionel Naccache, Stéphane Lehéricy, Ghislaine Dehaene-Lambertz, Marie-Anne Hénaff, and François Michel. The visual word form area. *Brain*, 123 (2):291–307, February 2000. ISSN 0006-8950. doi: 10.1093/brain/123.2.291. URL http://dx.doi.org/10.1093/brain/123.2.291.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009. doi: 10.1109/cvpr.2009.5206848. URL http://dx.doi.org/10.1109/cvpr.2009.5206848.

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. ISSN 1095-9203. doi: 10.1126/science.1063414. URL http://dx.doi.org/10.1126/science.1063414.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse,

Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/toy$_m odel/index.html$.

Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, April 1998. ISSN 1476-4687. doi: 10.1038/33402. URL http://dx.doi.org/10.1038/33402.

Thomas Fel, Ekdeep Singh Lubana, Jacob S. Prince, Matthew Kowal, Victor Boutin, Isabel Papadimitriou, Binxu Wang, Martin Wattenberg, Demba E. Ba, and Talia Konkle. Archetypal SAE: Adaptive and stable dictionary learning for concept extraction in large vision models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16543–16572. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/fel25a.html.

Thomas Fel, Binxu Wang, Michael Lepori, Matthew Kowal, Andrew Lee, Randall Balestriero, Sonia Joseph, Ekdeep Singh Lubana, Talia Konkle, Demba Ba, and Martin Wattenberg. Into the rabbit hull: From task-relevant concepts in dino to minkowski geometry. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 154538–154584, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/fa68d2dfd8eb6202e7902586197f7629-Paper-Conference.pdf.

Matteo Ferrante, Tommaso Boccato, Nicola Toschi, and Rufin VanRullen. Evidence for compositionality in fmri visual representations via brain algebra. *Communications Biology*, 8(1), August 2025. ISSN 2399-3642. doi: 10.1038/s42003-025-08706-4. URL http://dx.doi.org/10.1038/s42003-025-08706-4.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 50742–50768. Curran Associates, Inc., 2023. doi: 10.52202/075280-2208. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9f09f316a3eaf59d9ced5ffaefe97e0f-Paper-Conference.pdf.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders, 2024. URL https://arxiv.org/abs/2406.04093.

Alessandro Thomas Gifford, Domenic Bersch, Daniel Janini, Gemma Roig, and Radoslaw Martin Cichy. A resource of encoding models for in silico neuroscience. In *9th Annual Conference on Cognitive Computational Neuroscience*, 2026. URL https://openreview.net/forum?id=KJXoaBcNgz. Extended abstract.

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, February 2022. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2021.118812. URL http://dx.doi.org/10.1016/j.neuroimage.2021.118812.

D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154, January 1962. ISSN 1469-7793. doi: 10.1113/jphysiol.1962.sp006837. URL http://dx.doi.org/10.1113/jphysiol.1962.sp006837.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 7827–7845, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/1fa1ab11f4bd5f94b2ec20e794dbfa3b-Paper-Conference.pdf.

Habon Issa, Sunny Liu, Jona Ballé, and David Klindt. High-dimensional population codes reveal interpretable and diverse features underlying visual perception. 2026. doi: 10.64898/2026.09.05.748439. URL http://dx.doi.org/10.64898/2026.09.05.748439.

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), February 2023. ISSN 2399-3642. doi: 10.1038/s42003-023-04546-2. URL http://dx.doi.org/10.1038/s42003-023-04546-2.

Nancy Kanwisher, Josh McDermott, and Marvin M. Chun. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, 17(11): 4302–4311, 1997. ISSN 1529-2401. doi: 10.1523/jneurosci.17-11-04302.1997. URL http://dx.doi.org/10.1523/jneurosci.17-11-04302.1997.

Kendrick N. Kay, Thomas Naselaris, Ryan J. Prenger, and Jack L. Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, March 2008. ISSN 1476-4687. doi: 10.1038/nature06713. URL http://dx.doi.org/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.e9, October 2022. ISSN 0960-9822. doi: 10.1016/j.cub.2022.08.009. URL http://dx.doi.org/10.1016/j.cub.2022.08.009.

David Klindt, Charles O'Neill, Patrik Reizinger, Harald Maurer, and Nina Miolane. A unifying framework from neural superposition to sparse interpretable codes. *Nature Machine Intelligence*, 8(7):1025–1037, 2026. ISSN 2522-5839. doi: 10.1038/s42256-026-01259-z. URL http://dx.doi.org/10.1038/s42256-026-01259-z.

Michael A. Lepori, Kendrick Kay, and Greta Tuckute. Interpreting brain responses to language with sparse features from language models, 2026. URL https://arxiv.org/abs/2606.06857.

André Longon, David Klindt, and Meenakshi Khosla. Superposition disentanglement of neural representations reveals hidden alignment. In Marco Fumero, Clementine Domine, Zorah L"ahner, Irene Cannistraci, Bo Zhao, and Alex Williams (eds.), *Proceedings of UniReps: the Third Edition of the Workshop on Unifying Representations in Neural Models*, volume 322 of *Proceedings of Machine Learning Research*, pp. 342–359. PMLR, 06 Dec 2026. URL https://proceedings.mlr.press/v322/longon26a.html.

Andrew Luo, Margaret Marie Henderson, Leila Wehbe, and Michael J. Tarr. Brain diffusion for visual exploration: Cortical discovery using large scale generative models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=9VqMaSjf7U.

Andrew Luo, Margaret Marie Henderson, Michael J. Tarr, and Leila Wehbe. BrainSCUBA: Fine-grained natural language captions of visual cortex selectivity. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mQYHXUUTkU.

Andrew Luo, Jacob Yeung, Rushikesh Zawar, Shaurya Rajat Dewan, Margaret Marie Henderson, Leila Wehbe, and Michael J. Tarr. Brain mapping with dense features: Grounding cortical semantic selectivity in natural images with vision transformers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=yJ9QNbpMi2.

Ziming Mao, Jia Xu, Zeqi Zheng, Haofang Zheng, Dabing Sheng, Yaochu Jin, and Guoyuan Yang. Sparse autoencoders bridge the deep learning model and the brain, 2025. URL https://arxiv.org/abs/2506.11123.

Takuya Matsuyama, Shinji Nishimoto, and Yu Takagi. Lavca: Llm-assisted visual cortex captioning, 2026. URL https://arxiv.org/abs/2502.13606.

Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fmri. *NeuroImage*, 56(2):400–410, May 2011. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2010.07.073. URL http://dx.doi.org/10.1016/j.neuroimage.2010.07.073.

Shinji Nishimoto, An T. Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646, October 2011. ISSN 0960-9822. doi: 10.1016/j.cub.2011.08.031. URL http://dx.doi.org/10.1016/j.cub.2011.08.031.

Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996. ISSN 1476-4687. doi: 10.1038/381607a0. URL http://dx.doi.org/10.1038/381607a0.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features, 2025. URL https://arxiv.org/abs/2501.16615.

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

Paul Steven Scotti, Mihir Tripathy, Cesar Torrico, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq Mathew Abraham. MindEye2: Shared-subject models enable fMRI-to-image with 1 hour of data. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44038–44059. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/scotti24a.html.

Ghislain St-Yves, Emily J. Allen, Yihan Wu, Kendrick Kay, and Thomas Naselaris. Brain-optimized deep neural network models of human visual areas learn non-hierarchical representations. *Nature Communications*, 14(1), 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-38674-4. URL http://dx.doi.org/10.1038/s41467-023-38674-4.

Katherine R. Storrs, Tim C. Kietzmann, Alexander Walther, Johannes Mehrer, and Nikolaus Kriegeskorte. Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of Cognitive Neuroscience*, pp. 1–21, August 2021. ISSN 1530-8898. doi: 10.1162/jocn_a_01755. URL http://dx.doi.org/10.1162/jocn_a_01755.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Harrish Thasarathan, Julian Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos G. Derpanis. Universal sparse autoencoders: Interpretable cross-model concept alignment. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59304–59325. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/thasarathan25a.html.

Navve Wasserman, Matias Cosarinsky, Yuval Golbari, Aude Oliva, Antonio Torralba, Tamar Rott Shaham, and Michal Irani. Brainexplore: Large-scale discovery of interpretable visual representations in the human brain, 2025. URL https://arxiv.org/abs/2512.08560.

Daniel Wurgaft, Can Rager, Matthew Kowal, Sheridan Feucht, Usha Bhalla, Tal Haklay, Eric Bigelow, Raphaël Sarfati, Jack Merullo, Noah Goodman, Thomas Fel, Atticus Geiger, and Ekdeep Singh

Lubana. Manifold steering reveals the shared geometry of neural network representation and behavior. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=pf3HKAGT2M.

Huzheng Yang, James Gee, and Jianbo Shi. Brain decodes deep nets, 2024. URL https://arxiv.org/abs/2312.01280.

Alicia Zeng and Jack Gallant. Disentangling superpositions: Interpretable brain encoding model with sparse concept atoms. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 95964–95994. Curran Associates, Inc., 2025. doi: 10.52202/085713-3208. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/8a1bc250c372e0310d3d50892002b69b-Paper-Conference.pdf.

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.

A Brain Encoder Training For Voxels & Vertices

A.1 Pre-processing of NSD responses & DINOv2 embeddings

Our NSD Subject 1 encoder predicts responses over $V = 15,724$ native 1.8-mm `nsdgeneral` voxels. Images are resized to 224×224 via bicubic interpolation and normalized with ImageNet channel statistics. Frozen DINOv2 ViT-B/14 supplies the last post-LayerNorm patch embeddings: we concatenate $256 \times 768 = 196,608$ features, retaining patch position while omitting the class or register tokens. No additional feature centering, standardization, spatial pooling, or dimensionality reduction is applied and we fit our encoders to dense embeddings.

Responses are NSD `betas_fithrf_GLMdenoise_RR` estimates standardized per-voxel using statistics from each acquisition session, and then averaged over the three image presentations. Responses undergo a second normalization for ridge fitting using only the fitting images' voxelwise means and sample standard deviations.

A.2 Fitting and model selection.

The 9,000 non-shared images are sorted by NSD image ID, and the first 8000 determine the weights while the last are reserved for α fitting. All 1,000 shared images are withheld from both steps. With fitting features X and standardized responses Y_z, we solve

$$W_z = X^\top (XX^\top + \alpha I)^{-1} Y_z. \tag{5}$$

A single penalty is selected across voxels by maximizing the mean validation Pearson correlation over 20 log-spaced values from 10^{-2} to 10^{8}. Undefined correlations are excluded with equal weight given to each remaining voxel. For subject 1, the optimal alpha is $\alpha = 2.637 \times 10^{6}$. The same fitting and penalty-selection procedure was repeated independently for Subject 2. On the shared 1,000-image test pool, mean voxelwise Pearson correlations were 0.372 for Subject 1 (15,724 voxels) and 0.377 for Subject 2 (14,278 voxels).

Evaluation against an established encoder. The voxel encoder was primarily chosen due to its popularity in the NSD encoding and decoding scene, but surface encoders enable comparisons with performant non-linear encoders. We compare our linear training recipe against the HUZE architecture (Yang et al., 2024), winner of the Algonauts 2023 Challenge and implemented in the Brain Encoding Response Generator (BERG) project (Gifford et al., 2026). This model uses a strictly bigger feature pool by sampling from multiple DINOv2 layers and employing deep mapping networks. Our surface encoder predicts 37,984 `nsdgeneral` fsaverage vertices. Surface ridge and HUZE are evaluated on the same 515 test images and surface response targets, yielding a final mean vertexwise correlation of 0.482 and 0.520 respectively. Figure 6 also includes the voxel model rescored on those

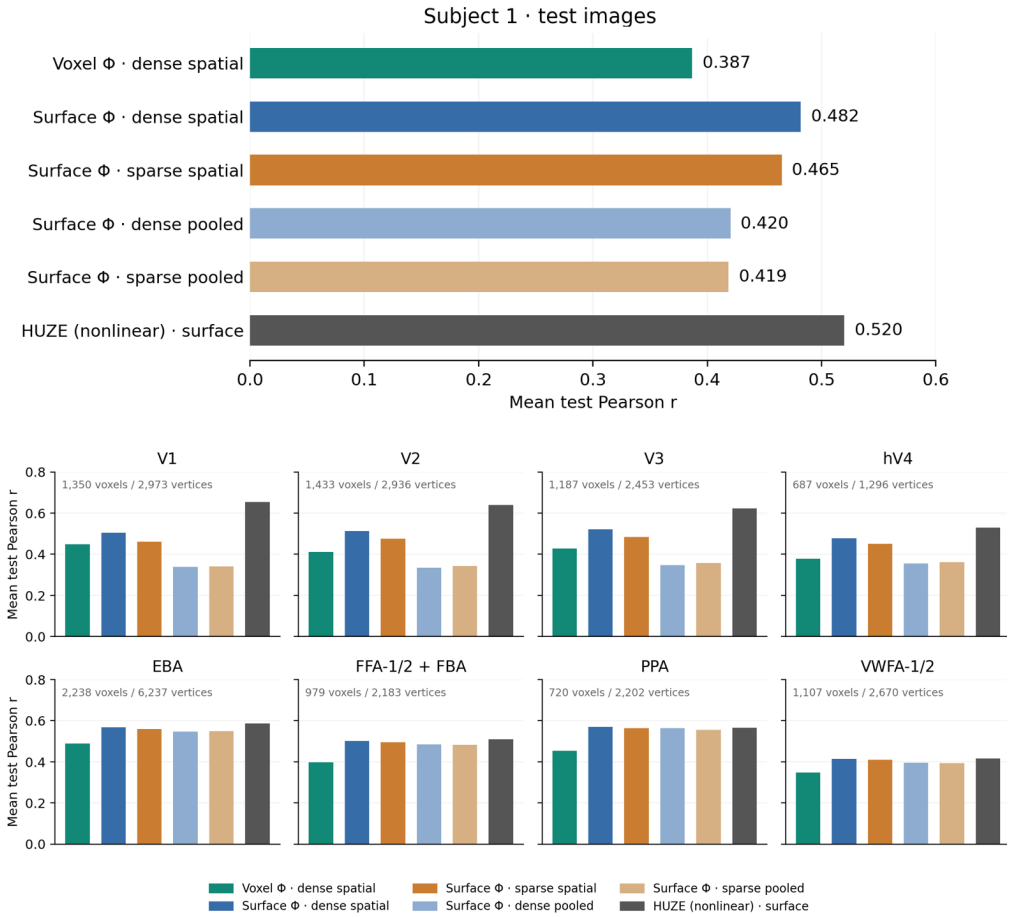

Figure 6: Encoder recipe tests with dense embeddings, sparse concepts, as well as pooling. Voxel encoder provided for reference, but should not be compared with surface encoders like HUZE even under the same mask.

515 images, though voxel and surface mean describe different response spaces and are **not** a controlled comparison of encoder accuracy. Correlations are averaged without clipping or noise-ceiling normalization.

Using latents ($z_{s,c}$) instead retains around 95% of this performance, indicating that concepts themselves support strong cortical alignment. Mean-pooling decreases this for both dense embeddings as well as SAE latents, with steep drops in V1 through V4. The HUZE encoder has a strong advantage in early visual areas, likely due to its built-in retinotopic priors and its sampling from additional, earlier DINOv2 layers.

B Sparse Autoencoder Architecture, Training, and Quality

This appendix provides additional details on the sparse autoencoder (SAE) dictionary used in our analyses. We describe its architecture, preprocessing, training configuration, and qualitative and quantitative evaluation.

B.1 Image preprocessing and backbone

The SAE is trained on 768-dimensional patch embeddings extracted from a frozen DINOv2 ViT-B/14 image backbone using the ImageNet-1k training set. Input images are resized to 224×224 pixels using bicubic interpolation in RGB and processed according to the DINOv2 input speci-

fications. For each image, we use all 256 normalized spatial patch embeddings and exclude the classification token. The SAE is trained on spatial patches independently of patch location. Unless otherwise specified, evaluation is performed on the ImageNet-1k validation set.

Each patch embedding is further standardized channelwise using embedding statistics computed from 1,281,167 images in the training set:

$$\widetilde{x}_s = \frac{x_s - \mu}{\sigma + 10^{-8}}.$$

B.2 SAE ARCHITECTURE

We use a top-k SAE consisting of a linear encoder followed by a ReLU nonlinearity and top-k sparsification. The latent representation retains only the 100 largest positive activations, so at most 1% of the 10,000 dictionary features are active for any patch. The decoder is an untied linear projection to the standardized $\mathbb{R}^{768}$ embedding space. Decoder directions are constrained to unit norm to facilitate comparisons of feature magnitudes and reduce shrinking in the decoder weights.

$$z_s = \mathrm{TopK}_{100}\left[\mathrm{ReLU}\left(W_E \widetilde{x}_s + b_E\right)\right],$$
$$\widehat{\widetilde{x}}_s = W_D z_s.$$

The latent representation satisfies $z_s \in \mathbb{R}^{10,000}$, corresponding to a dictionary of 10,000 atoms. This is approximately 13 times the dimensionality of the original embedding space, providing an overcomplete representation in which a large number of candidate visual concepts can be represented. The model contains 15.37 million parameters in total.

B.3 SAE TRAINING

The SAE was trained prior to the analyses in this work. It uses a standard reconstruction objective with a modified initialization intended to encourage distinct and stable feature directions.

B.3.1 INITIALIZATION

Inspired by Fel et al. (2025), we use spherical k-means to identify distinct directions in the embedding space. We first fit spherical k-means with $k = 10{,}000$ using $8{,}192{,}000$ patches for 20 iterations and retain the best solution across three runs. These centroids are used to initialize both the encoder and decoder weights with 20% added noise. Let C denote the centroid matrix and let ϵ denote independent standard-normal noise. We first form

$$\widetilde{C} = (1-p)C + p\ \mathrm{std}(C)\epsilon$$

The decoder directions are initialized from $\widetilde{C}$ and normalized to unit norm. The encoder directions are initialized from the same directions and scaled to have norm $1/\sqrt{k}$.

B.3.2 TRAINING OBJECTIVE

The SAE is trained using a standard mean-squared reconstruction objective:

$$\mathcal{L}_{\mathrm{recon}} = \mathrm{MSE}\left(\widehat{\widetilde{x}}, \widetilde{x}\right).$$

We additionally use a weak auxiliary loss for features that do not activate for any patch in a batch, encouraging otherwise inactive features to become active during training.

The model is trained for 50 epochs using Adam with a learning rate of 10^{-3}, linear warm-up, and cosine decay. The batch size is 32,768 patches. During the first 10 epochs, the decoder is frozen to encourage alignment of the encoder with the initialized feature directions.

B.4 SAE QUALITY EVALUATION

We evaluate SAE quality using both qualitative visualizations and quantitative measures.

B.5 MAXIMALLY ELICITING IMAGES

A maximally eliciting image (MEI) is an image that produces a high activation for a given SAE feature. For each feature, we identify the K images whose patches produce the largest feature activations.

For feature c, candidate images are ranked according to

$$m_c(i) = \max_s z_{s,c}(i).$$

Quantitative analyses use $K = 8$, while figures generally display $k = 3$ images for compactness. MEIs are selected among an image distribution and can be misleading when computed over a small or mismatched dataset. Depending on the analysis, we compute MEIs using either the ImageNet-1k dataset or the NSD stimulus set.

B.6 ACTIVATION OVERLAYS

To visualize the spatial regions associated with each SAE feature, we display patchwise feature activations as heatmaps over the input image. The 16×16 activation map is bilinearly upsampled to the image resolution, normalized for visualization, mapped using the `jet` colormap, and overlaid on the original image with opacity 0.65.

B.7 SAE CONCEPT EXAMPLES

Figure 7 shows 24 examples illustrating the diversity of visual concepts represented by the learned SAE dictionary.

B.8 QUANTITATIVE SAE EVALUATION

The SAE achieves R^2 = 0.8440, and explained-variance=0.8441 on the ImageNet-1k validation set. On the shared $1{,}000$ NSD test pool the corresponding scores where 0.8436 and 0.8439. Following Section 4.1, the top-eight SAE-side MEI sets have median DreamSim odd-one-out coherence scores of 0.9960 on ImageNet validation and 0.9224 on NSD.

C GENERALITY ACROSS IMAGE EMBEDDING SPACES

We repeated the main analysis using DINOv3 ViT-L/16 and CLIP ViT-L/14 image embeddings. For each embedding space, we trained a separate SAE dictionary, which was then frozen. Image processing followed each model's standard preprocessing procedure. Each SAE used the architecture and training procedure described in Appendix B, with 32,000 atoms and top-k selection with $k = 160$. For each embedding space, we used the same brain-encoder training procedure described in Appendix A. Recruitment maps and preferred-image metrics were computed as in Sections 3–4, using eight MEIs per feature.

Table 1: **Generality across image embedding spaces.** SAE coherence, brain coherence, and cross-set DreamSim scores for features derived from DINOv2, DINOv3, and CLIP image embeddings.

Image embedding	SAE coherence	Brain coherence	Cross-set DreamSim
DINOv2 ViT-B/14	0.922	0.906	0.211
DINOv3 ViT-L/16	0.928	0.904	0.213
CLIP ViT-L/14	0.849	0.913	0.208

These findings largely replicate the main results, demonstrating high brain-side coherence and partial SAE-brain correspondence. We show three similar concepts across image embeddings in Figure 8.

We compared the associated brain maps for each of these three concepts and found generally high correspondence.

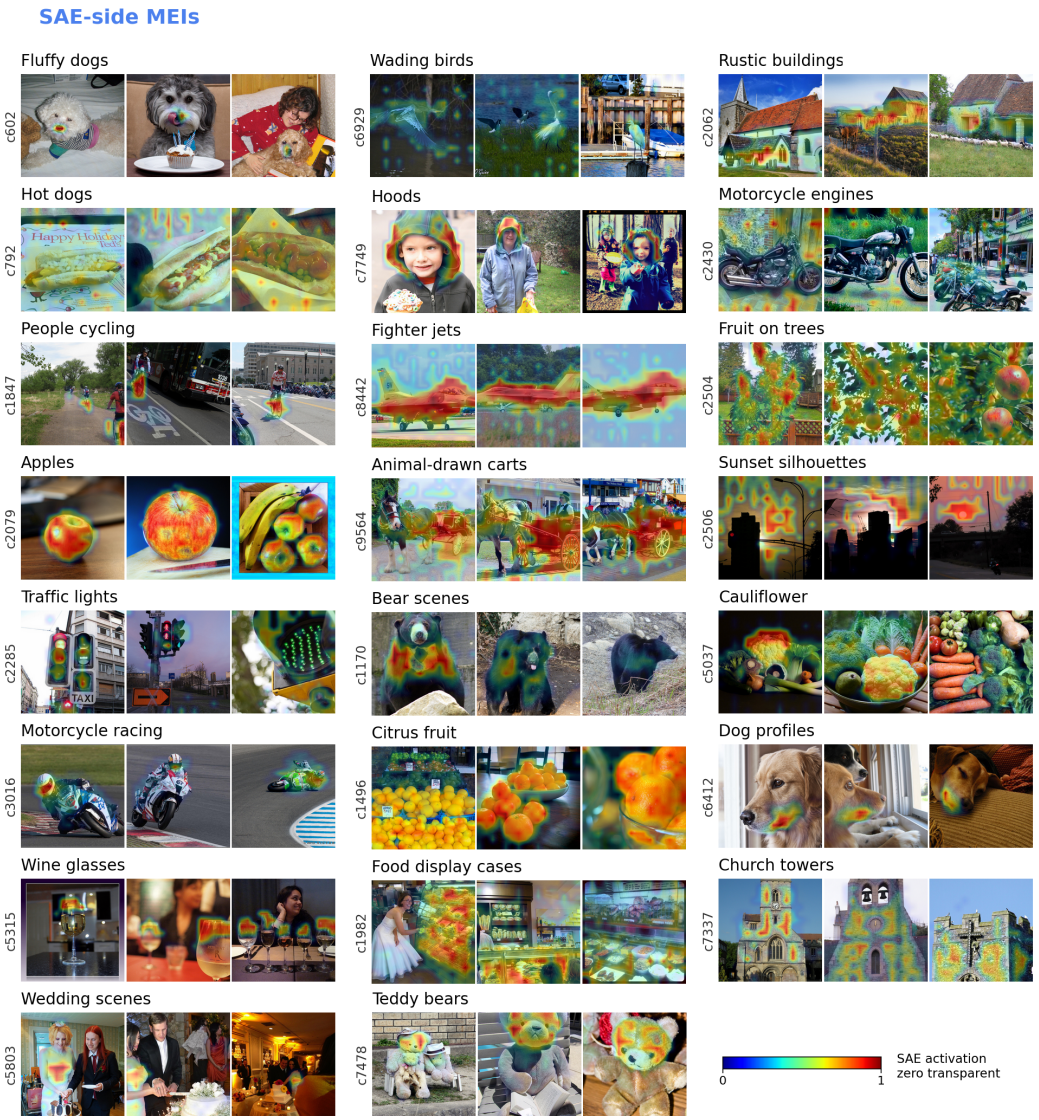

Figure 7: Examples of MEIs for 24 SAE features, illustrating the diversity of visual properties captured.

Table 2: **Cross-embedding brain-map correspondence.** Pairwise correlations between brain maps associated with matched concepts across DINOv2, DINOv3, and CLIP embedding spaces.

Concept	DINOv2–DINOv3	DINOv2–CLIP	DINOv3–CLIP
Plated food	0.569	0.491	0.799
Airplanes	0.418	0.545	0.441
Outdoor sports	0.840	0.764	0.731

D SUBJECT 2: CROSS-SUBJECT COMPARISON

D.1 MEI INTERPRETABILITY

We summarize concept map MEI analysis results for NSD Subject 1 and 2 using all eligible concepts, and the same shared 1,000 image test pool which was held out from encoder fitting and penalty selection. For each subject, brain-side MEIs are again ranked by spatial Pearson correlation

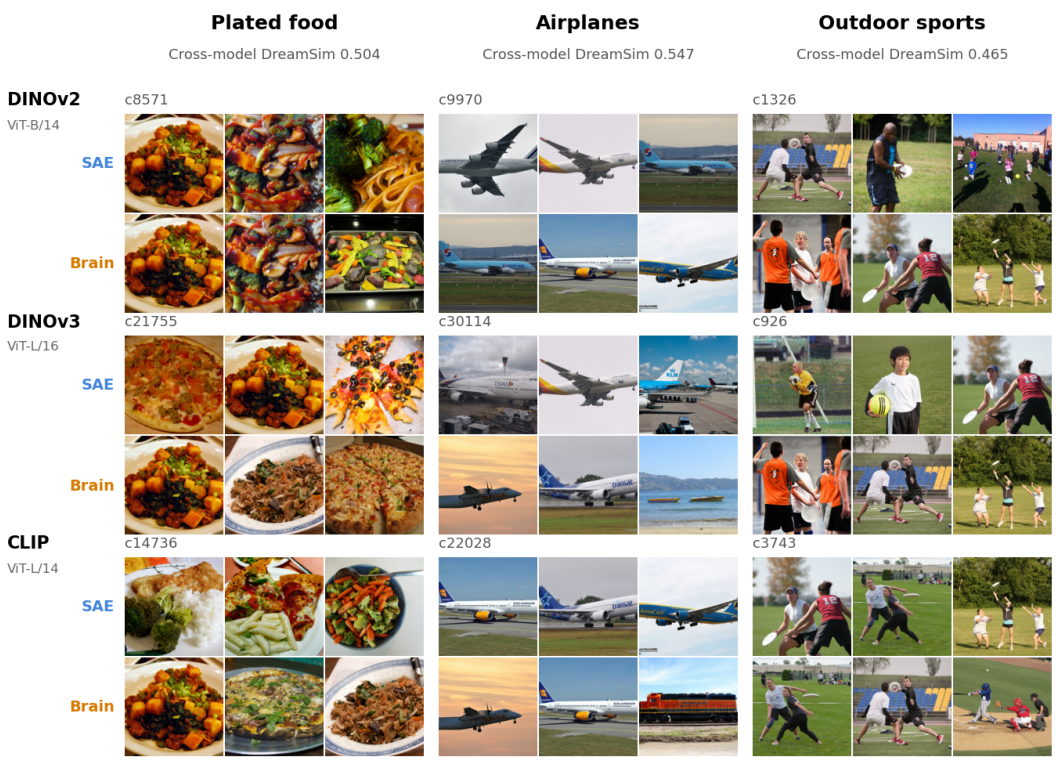

Figure 8: Examples of three MEIs of features that seem to share correspondence across image embeddings and brain regions.

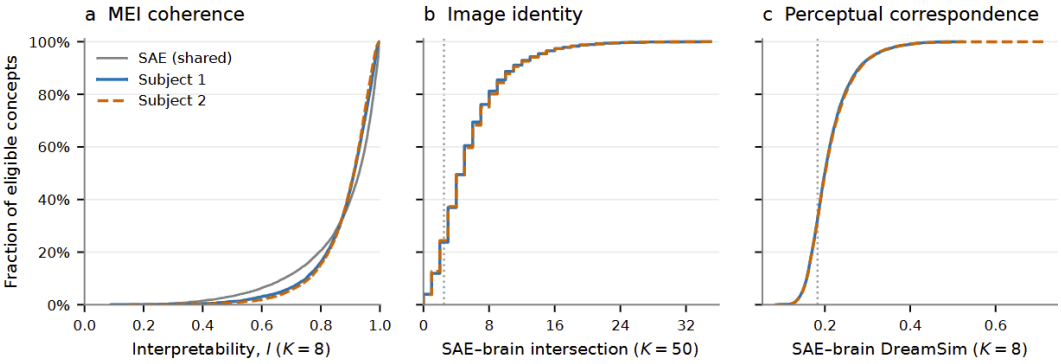

Figure 9: Empirical cumulative distributions of (a) brain-MEI interpretability I; (b) SAE-brain top-50 intersection divided by the pool-specific uniform-set expectation; and (c) top-eight SAE-brain DreamSim similarity minus the pool-specific reference.

between measured test responses and the corresponding $\Phi_{c,:}$ row. SAE-side MEIs use maximum patch activation, so their rankings and interpretability scores are identical across subjects because the dictionary and image pool are shared.

Both subjects show very similar results (Table 3; Figure 12). Median interpretability score is 0.9123 in Subject 1 and 0.9113 in Subject 2, compared with 0.9294 for the common SAE baseline.

Exact image agreement and perceptual correspondence is also very similar. Mean top-eight SAE-brain intersections are 0.2508 and 0.2438 images (equivalently, overlap fractions of just 0.03134 and 0.03048!) versus an independent uniform-set expectation of $K^2/1000 = 0.064$ images. Nonetheless, $8,000$ of the paired sets have no exact top-eight match. DreamSim similarity tells a similar

Table 3: **Concept-map retrieval on the shared test pool.** Interpretability uses $K = 8$. Brain-side summaries use each subject's eligible concepts; SAE-brain comparisons use the same paired concepts in both subjects. Intersection sizes below are image counts and not fractions.

Metric	Subject 1	Subject 2
Median brain-side interpretability	0.9123	0.9113
Brain-side concepts with $I \geq 0.8$	83.89%	84.66%
Brain-side concepts with $I \geq 0.9$	56.17%	55.67%
Mean SAE–brain intersection, $K = 8$	0.2508	0.2438
Mean SAE–brain intersection, $K = 50$	5.5412	5.5927
Mean SAE–brain DreamSim, $K = 8$	0.2112	0.2122
Mean DreamSim excluding identical pairs, $K = 8$	0.2083	0.2094

Table 4: **Development-pool summary.** Interpretability uses $K = 8$. Intersection sizes are image counts.

Metric	NSD S1	LAION S1
Development images	9,000	4,666
Eligible brain-side concepts	9,994	9,989
Median SAE-side interpretability	0.9690	0.9807
Median brain-side interpretability	0.9443	0.9264
Brain-side concepts with $I \geq 0.9$	73.46%	62.44%
Mean SAE–brain intersection, $K = 8$	0.1313	0.1672
Uniform-set expected intersection, $K = 8$	0.00711	0.01372
Mean SAE–brain intersection, $K = 50$	1.7757	2.2415
Mean SAE–brain DreamSim, $K = 8$	0.2186	0.1974
Mean DreamSim excluding identical pairs, $K = 8$	0.2171	0.1954
Uniform-set DreamSim reference	0.1859	0.1790

story, with paired set similarity scores of 0.2112 and 0.2122, versus the 0.1829 from uniformly drawn sets.

We stress that similar population distributions do not imply that every concept retains its identity, or that the two subject Φ matrices "prefer" the same images for every concept!

D.2 Qualitative MEI & prototype comparison

Figure 12 shows a spread of concepts sampled from the 100-concept set in Section 6. We overlay SAE patch activation, and count cross-subject Top-50 overlap to the side.

E LAION: Early exploration

E.1 LAION: MEI interpretability

We present summary results for LAION-fMRI (Zerbe et al., 2026), a recent fMRI scene-viewing dataset complementary to NSD. We compare NSD Subject 1 with LAION-fMRI Subject 1 using each subject's development pool: 9,000 NSD images and 4,666 images in LAION-fMRI's canonical training split. We note that while LAION-fMRI includes NSD samples, these amount to less than the shared 1000 set and did not appear evenly distributed among subjects, so we did not filter for shared images like in Appendix D. We sample concepts from a filtered set as in Section 6.

Both development pools support coherent brain-side retrieval. Median brain-MEI interpretability is 0.9443 for NSD and 0.9264 for LAION-fMRI. Both development pools support coherent brain-side retrieval. Median brain-MEI interpretability is 0.9443 for NSD and 0.9264 for LAION-fMRI. Anecdotally, we found that LAION MEI sets are indeed less interpretable in our naive application of the Φ. We attribute part this to the smaller unique stimuli in the set, as well as the broader image distribution providing less "crispy" or less numerous scenes eliciting the concepts. We stress that these are exploratory results given LAION-fMRI's recency.

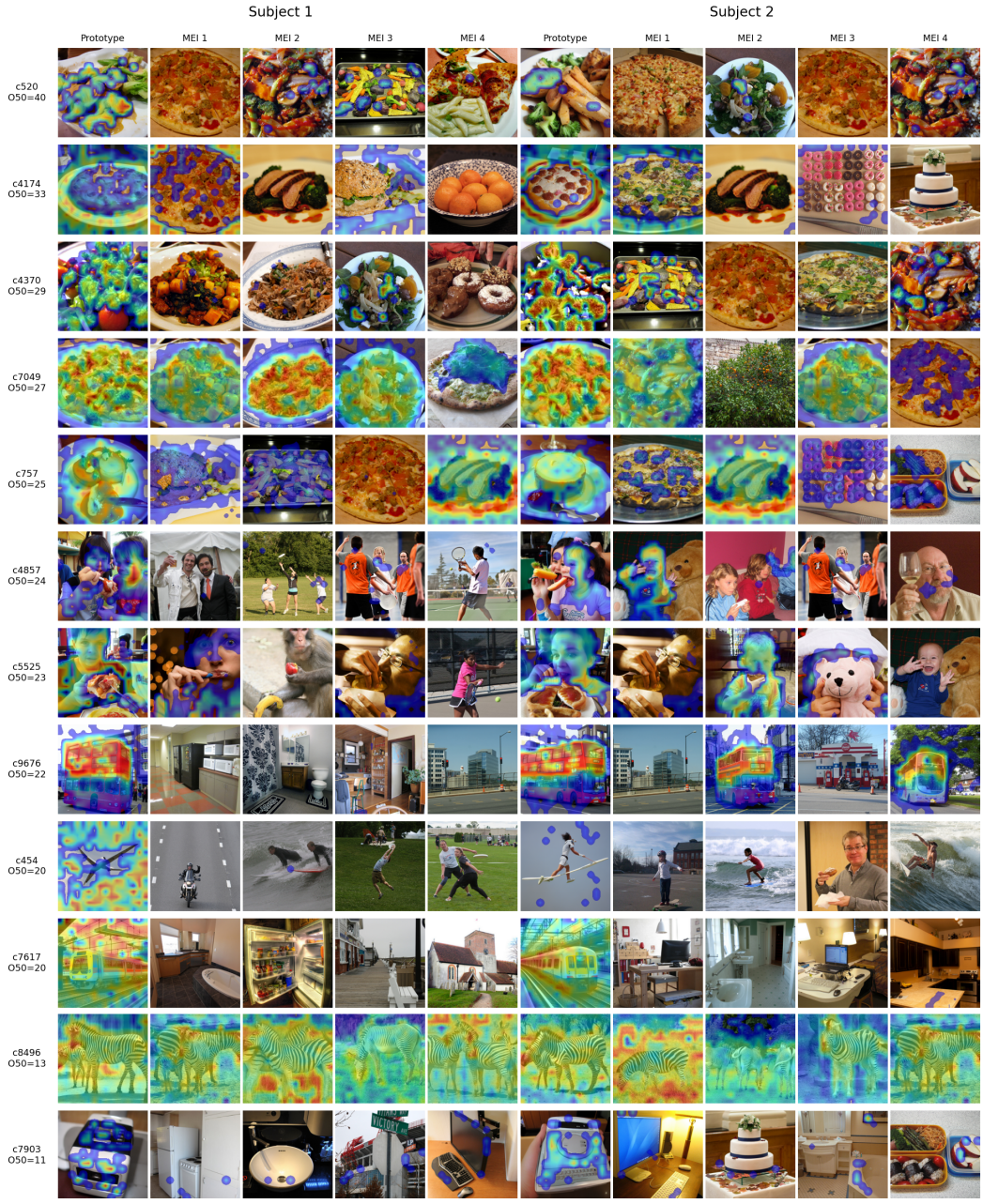

Figure 10: Prototypes and MEI set comparison for Subject 1 and Subject 2.

E.2 LAION: Qualitative MEI & prototype comparison

The following sheets show 20 concepts sampled without replacement from the 300 highest-overlap LAION concepts. We rank eligible concepts by the intersection of their LAION development-pool SAE and brain top-50 MEI sets, break ties by concept index, and display the sampled concepts in descending rank order. Note that this is a selection for high LAION correspondence, while the population statistics above continue to use all eligible concepts. Given LAION-fMRI's recency, we have not found an established and tested decoder, so we do not provide LAION-fMRI prototypes.

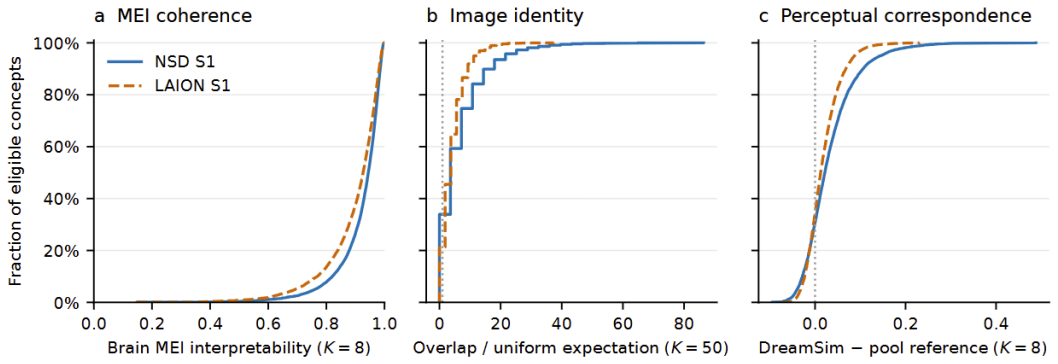

Figure 11: Empirical cumulative distributions of (a) brain-MEI interpretability I; (b) SAE-brain top-50 intersection divided by the pool-specific uniform-set expectation; and (c) top-eight SAE-brain DreamSim similarity minus the pool-specific reference.

F Construction of CLIP concept categories

Image-based category assignment. For the Figure 3b category analysis, we assign each of the 10,000 SAE concepts to one of ten manually specified categories: face, body, place, text, food, animal, vehicle, object, plant, and sport. This vocabulary includes familiar functional-localizer categories alongside broader natural-image content that "smear" across these groups. Labels are computed from SAE-side maximally eliciting images (MEIs), without using measured brain responses, recruitment maps, ROI masks, or localizer contrasts so they only represent model-side semantics. CLIP thus provides post hoc description of the learned dictionary.

For each concept, we take the first $K = 6$ images in its cached SAE-MEI ranking. These subject-independent rankings use maximum SAE activation across spatial patches over the full 73,000-image NSD stimulus collection. **Thus, this labeling pool is broader than the subject-specific fitting and held-out retrieval pools.** Each original RGB image is processed with the preprocessing transform supplied by OpenAI's `clip.load("ViT-L/14")` and encoded with the frozen CLIP image encoder. The model receives the image itself without heatmap overlays or feature-specific crops.

Prompt ensembles and scoring. Each category is represented by four text prompts (Table 5). Let $\widehat{f}(x)$ and $\widehat{g}(p)$ denote individually unit-normalized CLIP image and text embeddings. For concept c with $K_c \le 6$ selected images and category k with prompts $p_{k1}, \dots, p_{k4}$, we compute

$$u_c = \frac{\sum_{i=1}^{K_c} \widehat{f}(x_{ci})}{\left\| \sum_{i=1}^{K_c} \widehat{f}(x_{ci}) \right\|_2}, \qquad t_k = \frac{\sum_{j=1}^{4} \widehat{g}(p_{kj})}{\left\| \sum_{j=1}^{4} \widehat{g}(p_{kj}) \right\|_2}. \tag{6}$$

Images and prompts receive equal weight within their respective ensembles. The category score is cosine similarity $s_{ck} = u_c^\top t_k$, and the assigned label is $\widehat{k}_c = \arg\max_k s_{ck}$. Every concept receives one label; there is no abstention or multiple-label assignment. The saved confidence is $\max_k \exp(100 s_{ck}) / \sum_{\ell=1}^{10} \exp(100 s_{c\ell})$, using a fixed logit multiplier of 100. It is a relative score within this vocabulary and not a calibrated probability of semantic correctness.

The Figure 3 category-by-localizer matrix groups concepts by these fixed labels and averages their spatial Pearson correlations with each localizer t-map. The corresponding ROI-by-category summary averages recruitment selectivity within each category. Both summaries retain concepts with recruitment magnitude $\phi_c = \|\Phi_{c,:}\|_2$ at or above the median across concepts, and report a category mean only when at least five concepts remain. Neither summary applies a CLIP-confidence threshold. The label assignment is independent of brain measurements while this subsequent magnitude filter and the summarized statistics use Φ.

The categories are coarse forced-choice descriptions of whole-image MEI content. Agreement with localizer maps therefore assesses organization with respect to this image-derived vocabulary, but it cannot establish an exhaustive semantic interpretation of each concept.

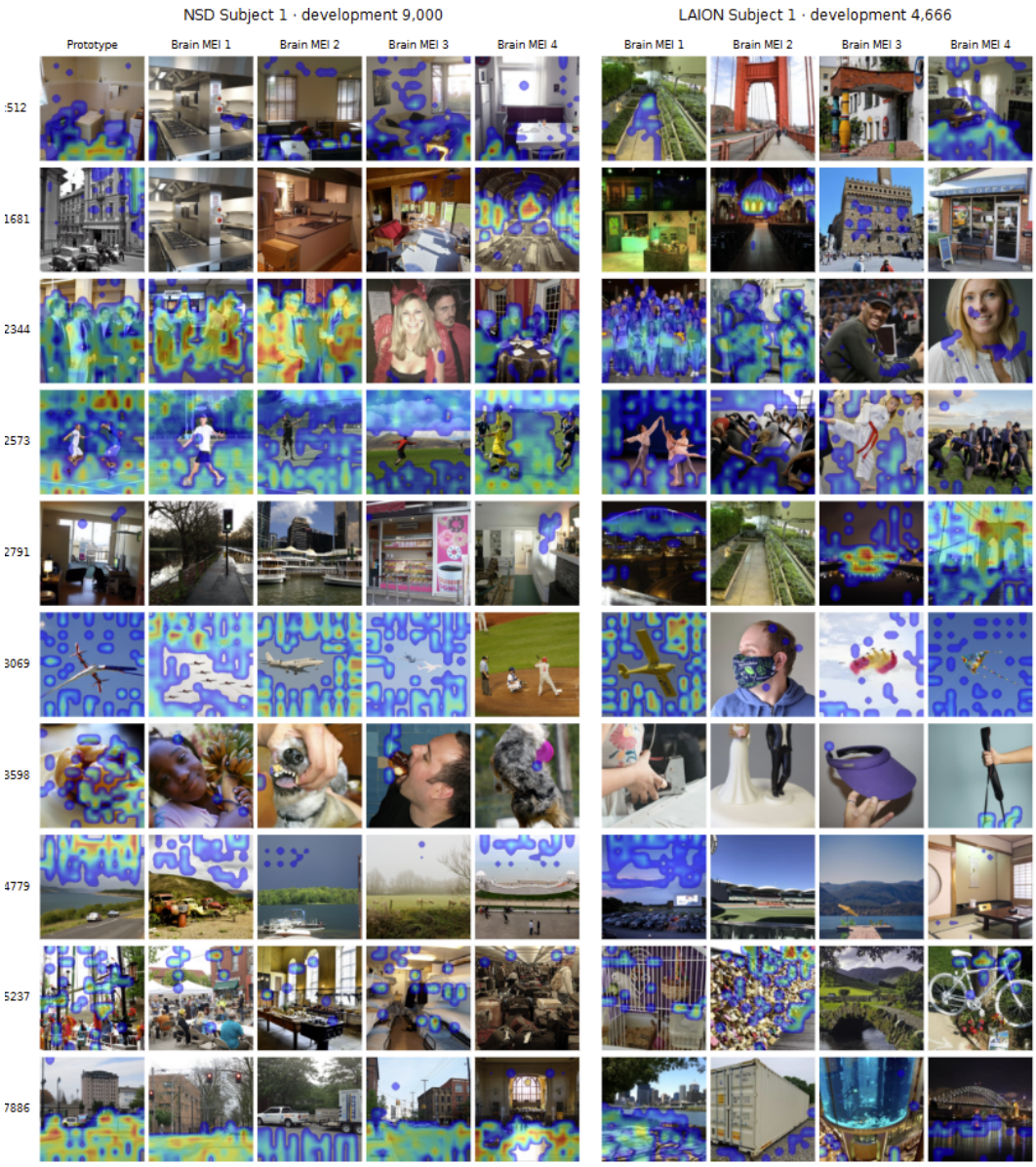

Figure 12: Prototype and MEI set comparison for NSD Subject 1 and LAION Subject 1.

Table 5: **CLIP category vocabulary and exact text prompts.** Counts describe all 10,000 concepts before the recruitment-magnitude filter used for the Figure 2 population summaries.

Category	Count	Prompts
face	56	a close-up photo of a person's face; a human face; a portrait of someone's face; a close-up of a face
body	996	a photo of people; a person's body with arms and legs; people doing an activity; a group of people
place	1,138	a photo of a place or scene; an indoor room; an outdoor landscape; a building or a street scene
text	110	a photo of text or writing; a sign with words on it; printed letters and numbers; a page of written text
food	980	a photo of food; a meal on a plate; fruit or vegetables; a dish of prepared food
animal	2,238	a photo of an animal; a wild animal; a pet animal; an animal in nature
vehicle	501	a photo of a vehicle; a car or a truck; a train or a bus; an airplane or a boat
object	2,579	a photo of an object; a household object; a tool or a device; a piece of furniture
plant	153	a photo of plants; flowers; trees and foliage; a garden
sport	1,249	a photo of a sport; people playing sports; an athlete in action; a sports field or court