



LUNDS
UNIVERSITET

Department of Psychology

An automated approach using sparse dictionary learning to discover visual representations in the human visual cortex

Qinglin Qi

Master's Thesis

2026

Supervisor: Geoff Patching

Abstract

Visual features are encoded by the population activity of neurons across the human visual hierarchy, from primary visual features in early areas to semantic visual features in the higher visual cortex. Sparse dictionary learning and sparse autoencoders have recently emerged as powerful automated and unsupervised methods for discovering monosemantic representations of features in the massive data of artificial and biological neural networks. The present study set out to discover, validate, and interpret the visual features encoded by fMRI activity across the visual hierarchy by analyzing a publicly released dataset of 7T fMRI responses collected from eight participants viewing over 70,000 natural images, with additional data augmentation. The study focused particularly on expanding the scope of visual feature discovery, separating feature recovery from its interpretation, and reducing agentic bias in discovering and interpreting visual representations. Participant-wise fMRI responses were decomposed by a sparse autoencoder into localised fMRI activity patterns. Each pattern was independently evaluated under two dedicated streams. In total, 35,181 and 6,603 fMRI patterns were discovered, validated, and interpreted, representing 113,714 semantic and 6,870 primary visual features, respectively. Two limitations identified in past literature were also tested quantitatively. The findings confirmed the dissociation between the selectivity and the interpretability of primary visual features, and that interpreting primary visual features as semantic risks representational misattribution. The present work contributes a novel method for the automated discovery of visual representations in human fMRI data, emphasising the authenticity of representation discovery and interpretation.

Keywords: human vision, visual cortex, visual representation, functional magnetic resonance imaging, sparse dictionary learning

Acknowledgements

This work owes credit to my supervisor, Prof. Geoff Patching, whose kind instruction and continued support carried me through the hardship of last February, when I parted ways with my former supervisor and abandoned my first thesis plan, and had to rebuild this project from the beginning. I would not have made it through without his warm help. I also appreciate the sustained advice and guidance from my external advisors at the National University of Singapore, Prof. Dianbo Liu and Yiming Tang, who offered key instructions and assistance throughout, and provided the computational resources on which parts of this work depended. I also appreciate the feedback and the prospective opportunities I received from Prof. Xufeng Duan at the Chinese University of Hong Kong, who helped shape my career into a new adventure in the fields of machine learning and neurolinguistics. Finally, I want to thank Prof. Roger Johansson and Åsa Arvidsson, whose kind assistance and guidance have helped me greatly through my master's study and thesis work. I sincerely wish all of you the best of luck in the future.

Beyond academia, I want to say thanks to my favourite writers, Dan Abnett and Aaron Dembski-Bowden. Your beautiful works, *The End and the Death*, *Echoes of Eternity*, and *Night Lords: The Omnibus*, have been a continued inspiration and source of courage for me. I also want to express my gratitude to my favourite composer, Jeremy Soule. Your majestic soundtracks for *The Elder Scrolls* have provided a sanctuary for me, throughout ten years, to concentrate on work and ponder life. I always remember that it was your pieces that accompanied me in my darkest, most desperate days seven years ago. I would also like to say tack så mycket to my SFI teachers, Paul Ulvros and Irina Rulevitj, for teaching me Swedish. Lycka till i framtiden!

Finally, I want to express my deepest gratitude to my lifetime Guru, Friedrich Nietzsche, whose complete works have accompanied my two-year adventure in Sweden. From my first read of *Die Geburt der Tragödie aus dem Geiste der Musik* to *Der Antichrist*, your words have challenged me to partake in my year of sustained training and my lifelong crusade towards self-overcoming and self-transcendence. *Denn nur als ästhetisches Phänomen ist das Dasein und die Welt ewig gerechtfertigt*. These exact words once transformed me from a decadent pessimist to a resolute warrior three years ago, and rekindled my life ambitions. I present this thesis, my first research project in machine learning and cognitive neuroscience, and the most challenging work I have devoted myself to thus far, in honour of Friedrich Nietzsche. It is a finale to one of the many battles in my life, and a prelude to my own *Morgenröte*.

Introduction

The Human Visual System and Visual Hierarchy

The visual system is the part of the central nervous system that receives, processes, and interprets visible light in order to make sense of the surrounding environment. Its task is to extract, from the two-dimensional pattern of light falling on the retina, the information an organism needs in order to identify its environment and decide how to act on it. This task is highly complicated and challenging, since the same object produces very different retinal images depending on its position, size, pose and illumination. The human visual system nonetheless demonstrates astonishingly high performance, identifying visual inputs within roughly a tenth of a second (DiCarlo et al., 2012). This is largely achieved by the architecture of visual neurons and by the hierarchical organising principle of the human visual system. At a cellular level, the properties of visual neurons are finely tuned to the statistical regularities of natural scenes. At a system level, the visual system is adapted to process those regularities economically under principles of efficient coding (Simoncelli & Olshausen, 2001). Visual function is consequently distributed across many cortical areas, which are linked by a largely hierarchical pattern of connections (Felleman & Van Essen, 1991). Through this arrangement, visual information of different kinds is processed at specialised levels of the visual hierarchy, allowing localised and efficient processing.

The visual hierarchy generally progresses from primary visual areas to higher cortical areas. Along this progression, the receptive fields of visual neurons grow larger, and the properties that drive a response become more complex. Responses also become increasingly invariant, so that the information from visual inputs becomes robust irrespective of viewpoint, relative size, and illumination (DiCarlo et al., 2012). Visual processing is conventionally divided into three levels along this progression (Marr, 1982). Low-level processing concerns primary visual features such as edges, luminance, and colour. Mid-level processing concerns the organisation of that pattern into surfaces, contours, and shapes. High-level processing concerns semantic visual features, such as objects and scenes (Peirce, 2015). The nature of the processed visual information also shifts as the hierarchy progresses. The primary and mid-level visual features vary continuously and are continuously perceived, whereas the high-level visual features fall into discrete semantic categories (Bao et al., 2020).

Population Visual Encoding and the Discovery of Visual Representations

Across the visual hierarchy, a visual feature can be specifically responded to by the coordinated activity of populations of neurons. This leads to visual features being encoded by populational activities of neurons for the brain to establish visual representations, a principle

known as population coding (Averbeck et al., 2006). Since population encoding involves a characteristic distribution of neuron activity over a massive number of cells, it produces measurable signals through neuroimaging methods (Panzeri et al., 2015). When a measured population activity pattern is shown to be specifically evoked by a visual feature, a bi-directional correspondence is established, so that the represented visual feature is identified from the neural signal it evokes, a research method also known as decoding (Naselaris et al., 2011).

Decoding neuroimaging data has accordingly become a common means of discovering visual representations in the brain and has yielded a large body of findings. For example, ensemble functional magnetic resonance imaging (fMRI) signals in early visual areas carry detailed information about the orientation of a viewed grating, even though individual voxels are only weakly selective (Kamitani & Tong, 2005), and viewing different object categories produces distinct and reproducible activity patterns across the ventral temporal cortex (Haxby et al., 2001). This approach has established a detailed picture of the visual hierarchy in the human visual system, including the primary visual features encoded in early areas. In early visual areas, orientation and spatial frequency are decodable from V1 responses to natural scenes (Roth et al., 2022), chromatic information is carried across the early retinotopic cortex (Brouwer & Heeger, 2009; Goddard et al., 2010), and sensitivity to binocular disparity appears in V2 and V3 (Nasr et al., 2016). Mid-level features follow a comparable pattern: figure regions are enhanced relative to backgrounds in early visual cortex (Poltoratski et al., 2020), local edge signals are integrated into extended contours (Qiu et al., 2016), V2 responds more strongly to naturalistic texture than to spectrally matched noise (Freeman et al., 2013), symmetry sensitivity emerges in V3 (Keefe et al., 2018), skeletal descriptions of shape explain response structure in V3 and lateral occipital cortex (Ayzenberg et al., 2022), and curvature is processed across several intermediate maps (Yue et al., 2020). In higher visual areas, distinct regions respond preferentially to faces (Kanwisher et al., 1997), places (Epstein & Kanwisher, 1998), bodies (Downing et al., 2001), and written words in readers (Cohen et al., 2000), and semantic category information is distributed broadly across this cortex (Huth et al., 2012).

Discovering visual representations has major scientific value in two respects. First, it is the basis on which computational accounts of vision are tested, where a model of visual processing is evaluated by how well it predicts the population responses actually measured (Naselaris et al., 2011). By doing this, fine-grained mechanisms of visual processing can be tested and established, deepening our understanding of the human visual system. Second, it informs and benefits the design of computer vision models (Yamins & DiCarlo, 2016).

Networks optimised for object recognition have been shown to predict responses in the higher visual cortex, and the degree of correspondence has been used to select among candidate architectures (Yamins et al., 2014). Constraining a network's early layers to reproduce the response properties of primate V1 has in turn improved its robustness to image perturbations (Dapello et al., 2020). This can further extend to applications in prosthetic vision, such as restoring rudimentary vision by converting camera input into patterns of electrical stimulation of the visual cortex (Lewis et al., 2015).

Representation Discovery Approaches and Their Challenges

Visual representations have conventionally been discovered through hypothesis-driven experimental designs. Under this paradigm, the visual features of interest are specified in advance by the researchers. The study then constructs or selects stimuli that vary along such visual features while holding others constant, measures the evoked responses, and tests whether the feature can be decoded from them (Felsen & Dan, 2005). An alternative approach is to use encoding models, which relax the requirement for controlled stimulus sets by presenting natural images and predicting the response each one evokes. The prediction is nonetheless computed from a feature space that the researcher defines in advance, so that a successful model predicts measured responses consistent with the tested features (Kay et al., 2008; Naselaris et al., 2011). Experimental representation discovery dominates the field due to two significant advantages: the stimulus properties can be manipulated rather than merely observed, and the stimuli constructed with carefully selected properties can support inferences about which feature drives a population response (Rust & Movshon, 2005). However, this approach also carries two constraints. The first concerns the stimuli: experimentally constrained stimuli such as bars and gratings rarely occur in isolation in natural environments, and so the results may suffer from limited generalizability in natural settings (Felsen & Dan, 2005; Rust & Movshon, 2005). The second concerns the discovered features, because a feature must be experimentally specified before it can be tested. Representation discovery under this paradigm is therefore bounded by the hypotheses a researcher has made.

A novel approach of automated representation discovery has emerged from the mechanistic interpretability of deep neural networks, where the representations of a trained network are reverse-engineered into human-understandable concepts (Bereska & Gavves, 2024). Since a trained model can comprise millions of neurons distributed across many layers, and each neuron's function is unknown in advance, testing them individually under experimental designs or inspecting them manually is not feasible at that scale (Bereska & Gavves, 2024; Sharkey et al., 2025). Automated discovery methods address this issue by

inferring what a unit represents from the inputs that most strongly activate it, and then translating the inference to natural language for human interpretability (Bereska & Gavves, 2024). The workflow involves determining the inputs that most strongly activate a unit, which are then presented to an interpreting agent for an account of what the unit represents. However, implementations of this workflow differ in how the agent is supplied with candidate concepts. Network Dissection compared each unit's activation map against pixel-level annotations of a set of concepts, and assigned the unit the concept whose annotated regions overlapped most closely with the regions the unit responded to (Bau et al., 2017). Later methods removed the requirement for annotated data, either by scoring the similarity between a unit's most activating images and an open vocabulary of concepts using a vision-language model (Oikarinen & Weng, 2022), or by prompting a language model to generate an explanation of a unit from its activating inputs and then scoring that explanation (Bills et al., 2023).

Automated representation discovery offers three unique strengths over experimental methods. First, it enables feature discovery at a scale that no manual or experimental procedure could. Second, it requires no a priori specification of the features to be discovered. Third, it can learn meaningful activity patterns from high-dimensional, noisy neural network data, where individual units are weakly informative on their own (Bereska & Gavves, 2024; Sharkey et al., 2025). These strengths of automated representation discovery are suitable for the analysis of artificial neural networks as well as neuroimaging data, which shares its properties of high-dimensionality, large data scale, and low signal-to-noise ratio (Allen et al., 2022; Norman et al., 2006).

However, automated representation discovery also faces a major obstacle in artificial neural networks, where it was developed. Applying this approach to discovering the representations of units in a trained neural network is complicated by the fact that individual computational units, commonly referred to as neurons, rarely correspond to single interpretable features. The phenomenon is called polysemanticity, and it arises naturally in artificial neural networks, since a network may represent more features than it has neurons. The network is consequently fitted to encode features through combinations of neurons, so that individual neurons respond to several unrelated features at once (Elhage et al., 2022). Polysemanticity leaves multiple plausible features available for the attribution of a neuron's activity, which complicates representation discovery in artificial neural networks. A similar issue pertains to neuroimaging data: without experimental control, natural stimuli contain many features at once, which are commonly populationally encoded. As a result, individual voxels in neuroimaging data are typically found to participate in the responses of multiple functionally heterogeneous

neuronal populations. Voxel-wise representations are therefore typically distributed across a continuous semantic space (Huth et al., 2012) with mixed feature selectivity (Taylor & Xu, 2024). Automated representation discovery therefore requires a method to identify monosemantic representations.

Sparse Dictionary Learning Approach of Automated Representation Discovery

Sparse dictionary learning (SDL) addresses polysemanticity by introducing the idea of sparse coding. This framework assumes that representations are encoded as a linear combination of a small number of units drawn from a learnt dictionary (Olshausen & Field, 1997). Under this framework, a unit learns a fixed pattern that does not vary across inputs, and a particular input only evokes the response of a sparse number of units (Cunningham et al., 2023). Units learnt under SDL are driven toward responding to single features rather than to several at once, and are more monosemantic and more interpretable than the neurons of the source network (Bricken et al., 2023; Cunningham et al., 2023).

Specifically, a sparse autoencoder (SAE) is a particular SDL approach in which a dictionary is learnt by being trained to reconstruct the input under a sparsity penalty (Cunningham et al., 2023). Because the reconstruction is linear, each dictionary unit's contribution to an input sample can be isolated as its activation specific to that sample. A unit's representation can be further discovered by examining its activation across the entire input as a particular feature that specifically evokes the unit's response (Bricken et al., 2023; Cunningham et al., 2023). Once trained, the encoder can calculate the response of every unit for any input sample in a single pass. This feature makes it particularly feasible for obtaining every unit's response across every sample in a massive input, which is beneficial for the large scale of inputs in automated representation discovery. The learning is also unsupervised throughout, requiring only a reconstruction objective and no labels, annotations, or specification of which features to look for (Bricken et al., 2023; Cunningham et al., 2023). Resolving polysemanticity, nonetheless, only pushes the learnt units toward being more interpretable. Because the objective is to reconstruct the input, the features driving the input are never directly observed during the decomposition. Representations are therefore discovered post hoc from the learnt dictionary units. Individual units are not guaranteed to represent meaningful features, and which features can be discovered depends on the feature recovery method applied (Cui et al., 2025).

SAEs have been extensively used to identify fine-grained representations across a range of artificial neural networks. In a one-layer language model, learnt units corresponded to specific and recognisable contexts such as DNA sequences, HTTP requests, and text in particular scripts, where the underlying units had been uninterpretable (Bricken et al., 2023).

At the scale of large language models, SAE discovered units encoding concrete entities, abstract properties, and nuanced behaviours such as sycophantic praise (Templeton et al., 2024). The same approach applied to a protein language model also recovered interpretable units corresponding to identifiable protein families and molecular functions (Gujral et al., 2025). In vision models, SAE recovered units corresponding to visual concepts that could be localised to the image regions in which they were expressed (Stevens et al., 2025).

Direct applications of SAEs to representation discovery in neuroimaging data, however, remain far scarcer. Existing work is mostly confined to BrainExplore (Wasserman et al., 2025), which applied SAE to decompose fMRI responses evoked across viewing different images into individual units. The top-activating images of each unit were then retrieved, and a vision-language model (VLM) was used to draw a broad range of hypotheses of potential visual representations from these images. The units were then individually evaluated by another VLM to resolve the best-matching hypothesis of what each represents, yielding interpretable visual representations en masse across the visual hierarchy. Overall, while the potential of SDL and SAE in automated representation discovery has been extensively demonstrated, these methods remain to be further established and validated in neuroimaging data.

The only existing study, nonetheless, has a few significant limitations (Wasserman et al., 2025). The first concerns the true objective of representation discovery. In BrainExplore, a visual representation is treated as discovered from a decomposed fMRI pattern when it is the best-matching feature for the top-activating sample based on a VLM's judgement (Wasserman et al., 2025). This criterion is also commonly seen in general automated interpretability methods, where a representation is discovered when its interpretation successfully describes its activation on samples (Paulo et al., 2025). This criterion, however, mixes two objectives of representation discovery: one being to reveal the latent visual feature represented by a neural activity pattern, and another being to make sense of it. On the one hand, a latent feature has to be represented for an activity pattern to be meaningfully interpreted; on the other, a present, established representation does not have to be directly interpretable (Jacovi & Goldberg, 2020), and the presence and interpretability of certain representations can be dissociated. Under this theoretical view, a feature's presence should first be evaluated independently from its interpretability, and the interpretation should happen post-hoc, constrained by the findings of the former. This separated, sequential representation discovery method could potentially discern features that are less interpretable yet genuinely represented, as well as benefit the authenticity of interpretation (Jacovi & Goldberg, 2020), but is yet underdeveloped in the present literature.

Second, the discovery results reported by BrainExplore were uneven across the hierarchy, between the primary and high-level visual areas. The proportion of dictionary units from which a visual representation was successfully discovered was 0.6% in V1, 0.9% in V2, and 1.3% in V3, against 14.5% in the extrastriate body area. While this result was attributed to the fact that representations in higher-level visual areas are easier to decompose and interpret (Wasserman et al., 2025), a deeper underlying issue concerns whether the representations in primary areas are fundamentally discoverable by its methodology at all. Under its pipeline, the visual representation of a decomposed fMRI pattern can be discovered only if it can be noticed, understood, and selected above all other alternatives by the VLM. This is problematic, since VLMs tend to mention semantically salient objects preferentially in their description of images (Tavakoli et al., 2017), while failing to translate detailed low-level visual features into words (Dahou et al., 2025; Rahmanzadehgervi et al., 2024). This makes the results of representation discovery dependent on the interpretability of different visual representations, thereby risking structurally omitting non-semantic, less interpretable, and subtle visual features.

Third, where a primary visual feature is represented, the VLM may still draw a semantic hypothesis for the fMRI pattern, based on a semantic feature that merely co-occurs with that primary feature. This might result in an fMRI pattern being misattributed to an incorrect visual feature while still showing considerable interpretability, so that the interpretation merely demonstrates plausibility instead of authenticity.

The present study

The present study primarily aims at discovering visual representations in human fMRI data by automatically discovering, validating, and interpreting the visual features encoded by fMRI activity across the entire human visual hierarchy using SAE.

Two further aims are proposed to address the aforementioned literature gaps. The first is to develop and validate a novel evaluation method for SAE-based automated representation discovery, which separates the evaluation of the presence of a visual representation from its interpretability. The present work establishes the presence of visual representations primarily by applying statistical analysis to recover visual features from the image-wise activation profiles of the decomposed fMRI patterns. Interpretation is implemented post hoc to feature recovery and is constrained by the recovered visual features. This emphasises feature recovery as the fundamental objective of SDL, which could improve the authenticity of interpretation.

The second aim is to extend the scope of automated representation discovery to primary visual features. The present work develops a dedicated feature recovery, validation, and interpretation method to recover primary visual features represented by the fMRI activity

patterns in primary visual areas. By systematically addressing the differences in properties and interpretability between primary and semantic visual features, the present work enables a more holistic discovery of visual representations from fMRI activity in the human visual system, guided by the hierarchical organisation of visual processing and the distinction between primary and semantic visual features.

RQs

RQ1. Can we discover and interpret fMRI activity patterns that encode semantic visual features through SDL?

RQ2. Can we discover and interpret fMRI activity patterns that encode primary visual features through SDL?

Hypotheses

The present work garners two hypotheses, which test the premises on which the above design rests. The first concerns whether the presence of a primary visual feature is genuinely dissociated from whether it can be interpreted (H1), and the second concerns whether interpreting a primary visual feature as semantic truly risks misattribution, while the result might still be masqueraded by high apparent interpretability and selectivity (H2).

H1. The selectivity and interpretability of the visual features represented by fMRI patterns learnt from primary visual areas are doubly dissociated. Some fMRI activity patterns may represent an uninterpretable yet highly selective primary visual feature; some may represent a highly interpretable yet less selective primary visual feature.

H2. If an fMRI activity pattern representing a primary visual feature is mischaracterized as representing semantic features, it may still appear highly interpretable and selective.

Method

Design

This study was a secondary, computational analysis of an existing functional neuroimaging dataset (Natural Scenes Dataset, NSD; Allen et al., 2022; Gifford et al., 2023). No new data were collected, no participants were recruited, and no variable was experimentally manipulated. The research design was observational and exploratory.

As the NSD comprises population-level fMRI activation patterns evoked by viewing images across the visual cortex, discovering individual fMRI patterns encoding a single visual concept can be challenging. The present work therefore decomposes the fMRI activation pattern across viewing all images by unsupervised learning into participant- and ROI-specific individual fMRI dictionary units as the unit of analysis. Each dictionary unit has a set of

activation strengths computed across all images viewed by its source participant. Each activation strength denotes how strongly the unit corresponds to the fMRI activity evoked by viewing each image. The decomposition is described in the Analysis section. Each dictionary unit was then analysed in three steps: determining whether it encodes any visual feature and recovering the feature(s) if present; evaluating the selectivity of the recovered feature(s); and interpreting the recovered feature(s). Two parallel data analysis streams applied these steps, one dedicated to primary visual features and one to semantic visual features.

Datasets and participants

The present study is based on two datasets; namely, the Natural Scenes Dataset (NSD; Allen et al., 2022) and the Brain Encoder Generated Dataset (BEGD). NSD is a public high-resolution 7T fMRI dataset collected while eight human participants each viewed 9,000 to 10,000 colour natural scene images, amounting to responses to 70,566 distinct images aggregated across participants. Specifically, I used the Algonauts 2023 preprocessed version of NSD (Gifford et al., 2023), which contains 75,086 image–fMRI pairs across eight participants. This dataset has preprocessed NSD into BOLD response amplitudes, projected the amplitudes onto the FreeSurfer fsaverage cortical surface, and restricted them to visual-cortex vertices maximally responsive to visual stimulation, in a normalized format that can be directly used for unsupervised learning. The vertices are labelled with a subset of visual regions of interest (ROIs) ranging from early retinotopic areas (V1–hV4) to higher category-selective areas, defined within each participant from the population receptive field and category functional localizer experiments reported by Allen et al. (2022). Because these ROIs are functionally rather than anatomically defined, a region may be absent in a participant for whom the localizer did not identify it. 20 ROIs whose activation was present in the majority of participants were retained, as listed in Table 1, yielding 158 participant-by-ROI response matrices per dataset, as one ROI (FBA-1) was absent in each of two participants.

BEGD is an original fMRI-image dataset generated based on the MS-COCO 2017 unlabeled dataset (Lin et al., 2014). BEGD is generated by deploying a Transformer Brain Encoder (TBE; Adeli et al., 2025) and predicting the fMRI activations evoked by viewing 123,403 images from MS COCO 2017 (Lin et al., 2014) for each participant. TBE was first trained on the image–fMRI pairs from the Algonauts 2023 Challenge (Gifford et al., 2023). For each participant, TBE learned a mapping from image input to the participant's fMRI activation pattern. After training, each participant's fMRI activations for each image were inferred in the MS COCO 2017 dataset, resulting in 123k inferred image-fMRI pairs per participant. BEGD was generated to support subsequent unsupervised learning by increasing the number of image–

response pairs and broadening the coverage of visual content available for dictionary learning. BEGD data and all derived results have been explicitly labeled and distinguished from NSD. For prospective research involving the reconstruction of seen images based on fMRI data, 100 image-fMRI pairs per participant were randomly selected from both datasets as held-out data.

Table 1

Regions of Interest Retained for Analysis

Abbreviation	Region
<i>Early retinotopic areas</i>	
V1v	Primary visual cortex, ventral
V1d	Primary visual cortex, dorsal
V2v	Secondary visual cortex, ventral
V2d	Secondary visual cortex, dorsal
V3v	Third visual area, ventral
V3d	Third visual area, dorsal
hV4	Human visual area 4
<i>Body-selective areas</i>	
EBA	Extrastriate body area
FBA-1	Fusiform body area 1
FBA-2	Fusiform body area 2
<i>Face-selective areas</i>	
OFA	Occipital face area
FFA-1	Fusiform face area 1
FFA-2	Fusiform face area 2
<i>Place-selective areas</i>	
OPA	Occipital place area
PPA	Parahippocampal place area
RSC	Retrosplenial cortex
<i>Word-selective areas</i>	
OWFA	Occipital word form area
VWFA-1	Visual word form area 1
VWFA-2	Visual word form area 2
mfs-words	Mid-fusiform sulcus word-selective region

Note. Region labels and definitions follow the Natural Scenes Dataset (Allen et al., 2022), in which retinotopic areas were defined from a population receptive field experiment and category-selective areas from a category functional localizer experiment. Ventral and dorsal subdivisions are denoted by the suffixes v and d.

Procedure

The experimental procedure of the source dataset is reported in full by Allen et al. (2022). In brief, fourteen right-handed candidates aged 18 to 65 with no known cognitive deficits or colour blindness and with normal or corrected-to-normal vision were invited to an initial 7T screening session comprising population receptive field and category functional localizer experiments. Eight participants were selected for the full experiment on the basis of data quality, behavioural performance, and head motion. The eight participants (two male, age range 19 to 32 years) each completed 30 to 40 core scanning sessions distributed over the course of one year. Informed written consent was obtained from all participants. The study was approved by the University of Minnesota Institutional Review Board.

Functional images were acquired on a 7T Siemens Magnetom passively shielded scanner with a single-channel-transmit, 32-channel-receive RF head coil, using whole-brain gradient-echo echo-planar imaging at 1.8 mm isotropic resolution (field of view 216×216 mm, matrix 120×120 , TR 1.6 s, TE 22.0 ms, flip angle 62°). Stimuli were colour natural scenes drawn from the MS-COCO database (Lin et al., 2014), presented at a visual angle of $8.4^\circ \times 8.4^\circ$ for 3 s with 1 s gaps. Each run lasted 5 min and contained 62 or 63 stimulus trials with occasional interspersed blank trials, and each session comprised 12 runs totalling 750 stimulus trials. While maintaining central fixation, participants performed a continuous recognition task, judging for each image whether it had been shown at any previous point in the experiment. Each participant viewed 9,000 to 10,000 distinct images over 22,000 to 30,000 trials, with each image presented three times and a set of 1,000 images shared across participants, yielding responses to 70,566 distinct images aggregated across the eight participants.

Analysis

Sparse dictionary learning

The input to SDL was a set of image–fMRI pairs from both NSD and BEGD, in which each stimulus image is paired with a participant-specific fMRI response vector across both hemispheres. Because the functions of visual ROIs are highly specialized and localized, learning from full hemispheric fMRI responses risks mixing fMRI patterns with diverse representations that require distinct proxies for feature probing. The present work therefore used SAE to learn from an ROI-wise response matrix, where rows index images and columns index voxels or vertices. For each participant, a Matryoshka sparse autoencoder (Matryoshka SAE;

Bussmann et al., 2025) was trained on the image–fMRI activation matrix for each non-empty ROI of each participant. The autoencoder maps each image-evoked ROI response into a sparse latent code and reconstructs the original voxel response from a learnt dictionary of a predefined size. Sparse autoencoders are widely used for discovering monosemantic patterns from large-scale activation data. This is because the decoding is linear and the encoding is sparse, so that each unit contributes a fixed, additive pattern to the reconstruction and can be examined in isolation (Cunningham et al., 2023).

Each learned unit denotes one fMRI pattern, whose activation strength across all images its source participant viewed can be computed. Each activation strength denotes how much the unit contributes to the reconstruction of the fMRI activity evoked by viewing that image. The activation strengths across all images constitute a unit's activation profile, which denotes how strongly each stimulus evokes the specific fMRI activity pattern. The sparsity constraint restricts how many units may be active for any single image, such that each unit is driven by a restricted set of images and fine-grained representations of visual concepts might be discovered. SDL was unsupervised, being trained to reconstruct the input fMRI activity alone; the visual contents of the images did not enter the training stage. This carries a cost as well as a benefit. Because no property of the images supervised the decomposition, no unit is guaranteed to represent any visual feature. The dictionary was also overlearnt, meaning the dictionary size largely exceeds the number of potential fMRI patterns representing meaningful visual features. These require subsequent screening and validation of the learnt units to determine whether each unit represents any meaningful visual feature. However, the nature of unsupervised learning also makes any correspondence subsequently found between a unit and a visual feature genuinely discovered.

As a result, SDL yielded a dictionary of 2,048 units per ROI per participant. Across the 158 participant-by-ROI response matrices retained per dataset, this produced 323,584 units per dataset. The decomposition of NSD was replicated twice under different random initialisations, so that including BEGD, a total of 970,752 units entered evaluation, each of which was evaluated independently.

Primary visual feature recovery, evaluation, and interpretation

The low-level visual feature (LVF) recovery stream was dedicated to evaluating the fMRI units learned from ROIs across primary and middle visual areas: V1v, V1d, V2v, V2d, V3v, V3d, and hV4. These areas show comparatively sparse semantic representations relative to higher visual cortex (Kriegeskorte et al., 2008) and are commonly modeled as representing

visual input through LVF channels tuned to natural statistical properties (Olshausen & Field, 1997).

Expert proxy probing. Seven LVF channels were selected, each implemented through an academically validated image-decomposition tool acting as an expert proxy for an early visual processing modality, decomposing every image into channel-specific tensor maps. Table 2 lists each channel, its supporting fMRI evidence, the corresponding tool and review, and the output tensor shapes. To avoid testing every unit under every channel, units were probed only under the biologically plausible channel $\times$ ROI combinations defined in Table 3, where the channels, their tested ROIs, and the supporting evidence are summarised.

Table 2

Low-Level Visual Feature Channels, Supporting fMRI Evidence, Computational Tools, Tool Reviews, and Output Tensor Shapes

LVF channel	fMRI evidence	Tool	Tool review	Tensor shape
Oriented energy / spatial frequency	Human V1 shows reliable orientation tuning under natural-scene sampling, and spatial-frequency preferences can be mapped at millimetre resolution across human primary visual cortex (Roth et al., 2022).	pyrtools (Simoncelli et al., 2025)	Pyrtools is a multi-scale image-processing package that implements steerable-pyramid decompositions into scale- and orientation-specific subband coefficients. Recent computational-vision work uses pyrtools steerable-pyramid bands for stimulus representation and orientation/spatial-frequency analysis (Zhang et al., 2025).	[N, 16, 32, 32]
Structured edges	V1 and V2 responses are sensitive to local contour structure, and fMRI studies of contour integration show that early visual cortical activity is modulated by contour alignment and contextual grouping (Qiu et al., 2016).	OpenCV SED (Dollár & Zitnick, 2013)	OpenCV StructuredEdgeDetection maps images into dense edge-probability maps using a structured-forest edge detector. Structured forests were shown to learn local edge-mask structure from natural image patches and to achieve strong edge-detection performance on BSDS500 and NYU Depth (Dollár & Zitnick, 2015).	[N, 1, 128, 128]

LVF channel	fMRI evidence	Tool	Tool review	Tensor shape
Opponent colour	Colour can be decoded and reconstructed from multivoxel responses across early retinotopic cortex and hV4, and colour encoding models show that V1 BOLD responses are systematically predicted by chromatic direction and contrast (Barnett et al., 2021; Brouwer & Heeger, 2009).	LuxPy (Smet, 2020)	LuxPy is a colour-science toolbox implementing standardised colour-appearance transformations, which map image colours into perceptual lightness and opponent-like chromatic axes, supporting its use over raw RGB for extracting perceptually organised chromatic-contrast maps (Smet, 2020).	[N, 3, 32, 32]
Relative depth	High-resolution human fMRI identifies disparity-selective structures in V2 and V3 and shows near-far response biases within these extrastriate maps (Nasr et al., 2016).	DPT (Ranftl et al., 2021)	DPT is a dense-prediction transformer that converts images into full-resolution dense prediction maps, including monocular depth estimates. Ranftl et al. (2021) use DPT for dense prediction and report strong monocular-depth performance, making it suitable for extracting relative-depth maps from natural images.	[N, 1, 64, 64]
Saliency	Human early visual cortex shows spatially resolved figure enhancement, indicating that primary visual responses contain foreground-background organisation (Poltoratski et al., 2020).	U ² -Net (Qin et al., 2020)	U ² -Net is a salient-object-detection network that produces dense saliency maps from images. Qin et al. (2020) use a two-level nested U-structure to capture multi-scale local and global context for salient-object detection, supporting its use as a foreground-likelihood proxy.	[N, 1, 256, 256]

LVF channel	fMRI evidence	Tool	Tool review	Tensor shape
Symmetry	Symmetry-selective BOLD responses first emerge in V3 and continue into downstream extrastriate regions, supporting V3 and hV4 sensitivity to visual regularity (Keefe et al., 2018).	EquiSym (Seo et al., 2022)	EquiSym is a group-equivariant convolutional network that outputs spatial maps for reflection axes and rotation centres. Seo et al. (2022) show that dihedral-equivariant feature maps improve reflection and rotation symmetry detection and achieve state-of-the-art performance on the LDRS and DENDI benchmarks.	[N, 1, 64, 64]
Skeleton	Skeletal-similarity models explain unique variance in human V3 response profiles, supporting V3 sensitivity to shape skeleton structure beyond simpler visual similarity (Ayzenberg et al., 2022).	scikit-image (van der Walt et al., 2014)	Scikit-image skeletonisation converts binary foreground masks into one-pixel-wide skeletons and medial-axis representations. The underlying algorithms preserve object connectivity and topology while reducing shape regions to skeletal structure, supporting their use as medial-axis topology maps.	[N, 1, 256, 256]

Note. Tensor output is reported in [N, C, H, W] format, where N denotes the number of images, C the number of signal channels, and H and W the spatial height and width of each tensor map. LVF = low-level visual feature; SED = structured edge detection; DPT = dense prediction transformer.

Table 3

Low-Level Visual Feature Channels, Tested Regions of Interest, and Supporting Literature Evidence

LVF channel	ROIs tested	Literature evidence
Oriented energy / spatial frequency	V1v, V1d, V3v, V3d, hV4	Multi-scale oriented energy was tested in V1 and downstream shape-related ROIs because human fMRI studies show reliable orientation tuning in V1 under natural-scene sampling, while later visual areas retain orientation and scale information as components of contour, shape, and boundary representations (Roth et al., 2022).
Structured edges	V1v, V1d, V2v, V2d, V3v, V3d, hV4	Human fMRI studies show that contour alignment and contextual grouping modulate responses in early visual cortex, including V1, V2, and V3, while downstream shape areas depend on boundary structure for object-form representation (Qiu et al., 2016).
Opponent colour	V1v, V1d, V2v, V2d, V3v, V3d, hV4	Human fMRI studies show that colour is decodable from multivoxel responses in V1, V2, V3, and V4 (Brouwer & Heeger, 2009), and cortical colour-processing work indicates that chromatic signals are represented through cone-opponent and recombined opponent-like channels in early visual cortex (Goddard et al., 2010).
Relative depth	V1v, V1d, V2v, V2d, V3v, V3d, hV4	Relative depth was tested across V1 to hV4 because disparity and depth signals are measurable in early visual cortex and become systematically organised in V2 and V3, where high-resolution human fMRI identifies near–far disparity-selective structures (Nasr & Tootell, 2018).
Saliency	V2v, V2d, V3v, V3d, hV4	Human fMRI shows foreground–background segmentation during natural-image viewing, and intermediate visual cortex is implicated in border ownership and proto-object organisation. Foreground segmentation is beyond the functional scope of V1 (Papale et al., 2018; Zhou et al., 2000).
Symmetry	V3v, V3d, hV4	Human fMRI shows symmetry-selective responses emerging from V3 and continuing through downstream extrastriate visual cortex (Keefe et al., 2018).
Skeleton	V3v, V3d, hV4	Human fMRI shows skeletal-similarity models explaining response structure in V3 and lateral occipital cortex (Ayzenberg et al., 2022).

Note. LVF = low-level visual feature; ROI = region of interest. Ventral and dorsal subdivisions are denoted by the suffixes v and d.

Feature recovery and evaluation. For each unit under each candidate channel, a candidate feature was recovered using a response-weighted average (RWA), following the logic of spike-triggered averaging (STA), which measures the stimulus structure driving neural

responses by averaging the stimuli that precede spikes (Sharpee, 2013). Replacing spike counts with unit-wise activation strength over all images, the RWA tensor was contrasted against an equally weighted average over all images, isolating the visual feature that drives the unit's activation. An expression score (ES) was then defined as the cosine similarity between each image's prior-centered LVF tensor and the tested feature, measuring the trial-wise presence of the feature in the stimulus. Each candidate feature passed three gates. First, its standardized signal strength had to exceed one standard deviation above its reference distribution. Second, to avoid attributing units to features merely prevalent in the image set, the feature's image-wise expression had to be predictable from the ROI-wise fMRI response, following the principle of multi-voxel pattern analysis (MVPA; Naselaris et al., 2011). A ridge regression predicting ES from the ROI activity matrix was fitted with 3-fold cross-validation, its cross-validated score taken as the feature's predictability score (PS). PS was residualized against signal strength to remove the confound that stronger features yield more reliable estimates, and the residualized PS was required to exceed one standard deviation. The response-weighted average was computed from all images first. The ridge was then cross-validated on scores of how closely each image matched that already-computed average, so a test image had already entered the average. Third, the feature had to be enriched among the images that most strongly activate its unit. Following common practice in sparse dictionary learning, each unit's top-20 images by activation strength were taken as its top-activation samples (Cunningham et al., 2023), and a feature with ES above one standard deviation was treated as present in an image. Enrichment was tested against the feature's global prevalence with a hypergeometric test, Benjamini–Hochberg corrected across the features surviving the preceding gates, requiring $q \leq .05$. A feature passing all three gates was recovered, and a unit yielding at least one recovered feature was retained. Because each unit is probed under multiple channels, a unit may carry several recovered features that co-vary in images rather than reflecting genuine polysemanticity. For each such unit, whether each feature's ES remained predictable after removing the variance explained by the others was therefore re-tested, retaining only features with unique predictability and resolving a unit as polysemantic when more than one survived. For each recovered feature, its enrichment rate over global prevalence within the top-activation sample is reported, together with its Benjamini–Hochberg corrected significance. Recall-related metrics were omitted as structurally uninformative in a small sample. At the all-image level, receiver operating characteristic curve (AUROC) was defined with top-20 membership as the binary target and ES as the continuous score, which is used as a prevalence-independent diagnostic of feature selectivity in the analysis of the dissociation between selectivity and

interpretability. This index denotes the probability that a randomly chosen image from the top-20 sample expresses the feature more strongly than a randomly chosen image from outside it.

Post-hoc interpretation. Following common automatic-interpretability practice in sparse dictionary learning, each unit was interpreted within its top-20 activation sample. To improve agent-side interpretability, each recovered feature's spatial expression was first visualized as a heatmap over the sample, marking local positive and negative contributions to the feature. The standard two-stage procedure was then followed (Cunningham et al., 2023; Wasserman et al., 2025), generating a natural-language explanation from the sample and validating it against the same sample, but interpretation was constrained under expert-proxy and feature-recovery constraints to reduce VLM agent bias (Tavakoli et al., 2017; Dahou et al., 2025). First, a VLM (GPT-5.4-mini) was shown the raw top-20 sample, their feature heatmaps, and their overlay, and explained the recurring signal pattern of each feature in one sentence, constrained to the feature's channel and method while avoiding semantic content. Second, a separate evaluator (Gemini-3.1-flash-lite-preview) was shown the same inputs and the explanation, and assigned each of the 20 images a binary label for whether it supported the explanation. The proportion of positive labels, which reflects a feature's interpretability under agent inspection, is termed the apparent interpretability score (AIS) to distinguish it from the feature recovery metrics above. Both explainer and evaluator prompts are provided in the Appendices.

Semantic visual feature recovery, evaluation, and interpretation

The high-level visual feature (HVF) stream evaluated the fMRI units learned across all ROIs spanning the visual hierarchy. Unlike the LVF stream, the HVF stream discovers and interprets units on the basis of semantic visual features. It was nonetheless applied to evaluating the units learned from lower areas (V1–V3, hV4) to contrast the feature recovery and interpretation outcomes as a result of using different proxies.

Image captioning. To obtain a holistic semantic representation of each image before HVF recovery, a VLM (Gemini-3.1-flash-lite-preview) was first prompted to convert every image into a structured caption. The schema included a global caption alongside subfields covering the scene's primary focus, salient objects, setting, colours, activity, and any visible text. This step aims at creating a basis for feature recovery while reducing potential agent bias in attending to semantically salient or accessible features. The full prompt is provided in the Appendices.

Canonicalization. Each caption was vectorized to enable statistical downstream feature recovery. First, an LLM (GPT-4.1-mini) was prompted to convert all captions into

atomized standalone visual features. The outputs were then aggregated, normalized, and deduplicated into a standardized dictionary of 3,006 canonical visual features spanning the 11 caption fields. The full dictionary is available at <https://drive.google.com/file/d/1EX10NA19yfH40DYW7QD1OhFieCzHQT6Y/view>. For each caption, a set of candidate features from the dictionary was then shortlisted with a hybrid strategy of lexical match and high CLIP embedding cosine similarity. An LLM (Gemini-3.1-flash-lite) was prompted to verify whether each shortlisted feature was present in the caption, labeling 1 (present) or 0 (absent). The verification prompt is provided in the Appendices. Letting each canonical feature define a dimension, each caption was vectorized into a sparse binary feature vector, in which an entry is one only when its feature was both shortlisted for that image and verified present, and zero otherwise.

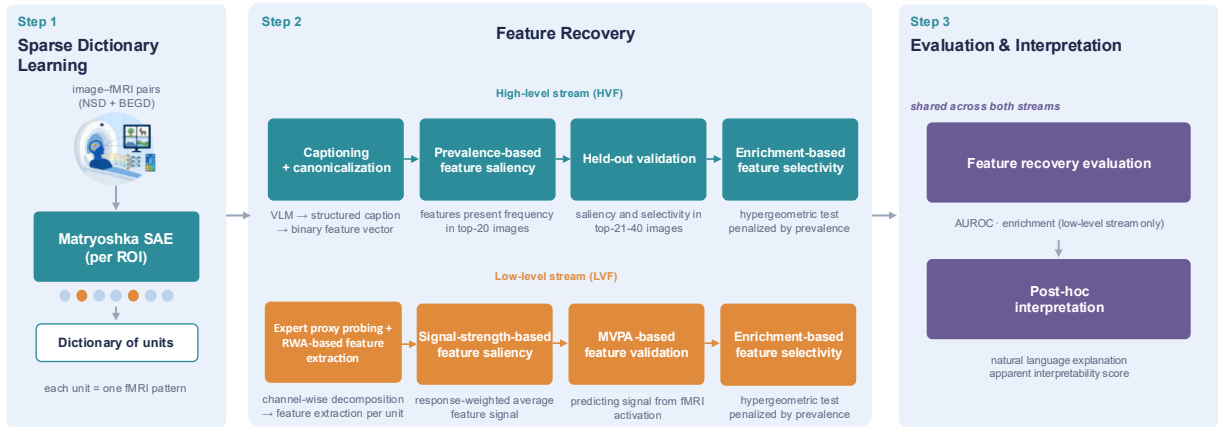
Feature recovery and evaluation. For each SAE unit, the visual features it encodes were identified from the images that most strongly activate it. First, each unit's top-20 most strongly activating images were retrieved, following the top-k sample conventions in SAE interpretability (Wasserman et al., 2025). Within this sample, only salient features appearing in at least ten images were retained, following the top-k recovery criterion in BrainExplore (Wasserman et al., 2025). For each salient feature, whether it is enriched in the activation sample relative to its prevalence across all images in the dataset was then tested, using a hypergeometric test with Benjamini–Hochberg correction across all dictionary features. Each feature was finally validated under the same saliency and enrichment criteria in its held-out top-21–40 sample. A feature is recovered when it is both salient and enriched at a corrected significance of $q \leq .05$ in both its top-20 and 21–40 activating samples. Because feature encoding may occur in both directions (Franke et al., 2025), depleted features were also recovered under the mirrored criteria. A feature is recovered as depleted when it is absent from at least ten of the unit's top-20 images and is significantly under-represented under the same hypergeometric test (lower tail), and when it re-passes both criteria in the held-out top-21–40 sample. Enriched and depleted features are never pooled. A unit with at least one recovered feature, in either direction, is retained. The feature recovery performance in each retained unit was then evaluated with the same enrichment metrics as the LVF stream. Unlike LVF, HVF has no graded per-feature expression score, since a semantic feature is either present or absent in an image. The AUROC diagnostic is therefore not constructed for HVF.

Post-hoc interpretation. The same two-stage explainer–evaluator procedure as the LVF stream was followed, with two differences. First, since the recovered features in HVF are solely semantic, all recovered features of a unit were presented to the VLM without visualizing

them. Second, units with only enriched features and units with at least one depleted feature were separately processed under different schemas. Both interpretation prompts and the evaluator prompt are provided in the Appendices. The full data analysis workflow is summarised in Figure 1.

Figure 1

Overview of the Visual Representation Discovery Procedure



Note. Step 1: a Matryoshka sparse autoencoder is trained per region of interest on image-fMRI pairs from NSD and encoder-augmented BEGD, decomposing each ROI response into a dictionary of sparse units, each of which defines an ROI-specific fMRI activity pattern. Step 2: the visual features encoded by each unit are recovered in two parallel streams under a shared structure. The high-level stream captions and canonicalizes every image into a binary semantic feature vector; the low-level stream decomposes every image through expert proxy channels and extracts a response-weighted feature signal per unit. Each recovered feature then passes through conjunctive gates to be salient, decodable, and selectively enriched, with the semantic stream additionally revalidated on a held-out sample. Step 3: both streams share the same evaluation and interpretation, in which recovered features are evaluated by enrichment and then explained in natural language and assessed by an apparent interpretability score. HVF = high-level visual feature; LVF = low-level visual feature; SAE = sparse autoencoder; MVPA = multi-voxel pattern analysis; RWA = response-weighted average; AIS = apparent interpretability score.

Ethical considerations

This work does not involve the collection or processing of any personal data. The work only deploys the data that has been developed for unrestricted academic access and research work. The thesis utilises NSD, an open dataset permitted for unrestricted academic access and usage. Informed written consent was obtained from all NSD participants, who retained the right to discontinue participation, and the original data collection was approved by the University of Minnesota Institutional Review Board (Allen et al., 2022). BEGD consists of responses predicted by a brain encoder trained on the measured responses of the same eight NSD participants, so no further data was collected.

The public NSD has been fully anonymized with no identifiable information that researchers can use to identify individuals. The Algonauts 2023 Challenge dataset, directly used in this thesis, further helps to anonymize the data by preprocessing the raw NSD. This makes the data fully anonymized, and no individual is or no longer identifiable by any means that are reasonably likely to be used. The data does not meet the legal definition of personal data according to GDPR, and does not require any ethical review. Similar studies using NSD within the European Union, for example Bersch et al. (2025) and Ozcelik and VanRullen (2023), likewise declared no requirements for ethical approval. The data collection, processing, and analysis contain no foreseeable risks for any person. All data management complies fully with Lund University's rules and regulations.

Result

Discovery, Validation, and Interpretation of Visual Representations

Sparse dictionary learning yielded a dictionary of 2,048 units per ROI per participant and 323,584 units per dataset. With NSD replicated twice under different seeds during SDL, 970,752 units were evaluated in total. Each unit defines an ROI-specific fMRI activity pattern and was independently evaluated. Following the procedures described in the Method section, the HVF stream recovered 104,111 enriched and 9,603 depleted semantic visual features, and the LVF stream recovered 6,870 primary visual features. Since more than one feature may be recovered from a unit in both streams, these features correspond to 35,181 unique units in the HVF stream and 6,603 in the LVF stream. By dataset, the HVF stream recovered 2,356 and 2,384 units from the two NSD replications, against 30,441 from BEGD; the LVF stream recovered 1,355 and 1,314 units from NSD, against 3,934 from BEGD. The close agreement between the two NSD replications indicates that the pipeline is reproducible across random initializations of SDL. The cross-dataset heterogeneity is expected, as BEGD is TBE-generated over a much larger image pool per participant (approximately 123k images against approximately 10k in NSD).

Semantic visual representations

The HVF stream recovered 104,111 enriched semantic features across 34,910 units. Each recovered feature is a semantic visual concept over-represented among the unit's top-20 activating images, and is validated in the unit's held-out top-21–40 images. On average, 2.98 semantic features are recovered per unit, and every recovered feature of a unit contributes to its representation and its explanation. Recovery is robust to stricter significance thresholds: 75,191 features (72.2%) hold at $q \leq .01$ and 46,825 (45.0%) at $q \leq .001$ under Benjamini – Hochberg

correction. Beyond enriched features, the HVF stream recovered 9,603 depleted features across 3,058 units. Depleted features are semantic features unexpectedly absent from a unit's top-20 activations relative to their prevalence, and are likewise validated in the held-out top-21–40 images. They are recovered with comparable selectivity (median depletion 0.45) and higher interpretability (median AIS 0.85, against 0.75 for enriched features). Recovery criteria that test only for over-representation are one-tailed by construction and cannot recover features encoded by conspicuous absence, of which 271 units here are recovered by depletion alone.

Across the HVF stream, the recovered units are both selective and interpretable. At the unit level, 22,031 units (63.1%) are selective for a feature at 3 times the base rate and 14,191 (40.7%) at 5 times, while 93.6% of units reach an AIS of 0.5 and 41.2% reach 0.8. HVF selectivity is hardly explained by dataset, ROI, or participant (model $R^2 = .072$), whereas the quantity of recovered features is dominated by dataset (partial deviance = .659; Poisson pseudo- $R^2 = .798$). The asymmetry is most pronounced in primary visual areas, where the stream recovers 11,972 of 114,688 BEGD units (10.4%) but only 177 and 159 of the corresponding NSD units (0.15% and 0.14%). Whether unit selectivity generalizes beyond the top-20 and top-21–40 images was further examined, since interpretation based only on the most strongly activating sample can misrepresent a unit. Units remain selective for a majority of recovered features beyond the head sample, from 62.4% at ranks 41–60 to 56.7% at 81–100 in BEGD, and from 54.8% to 48.6% in NSD. Semantic visual representations were therefore discovered and interpreted across the visual hierarchy, with 104,111 enriched and 9,603 depleted representations over 34,910 and 3,058 units, respectively.

Primary visual representations

The LVF stream recovered 6,870 primary visual features across 6,603 units. Every recovered feature is a primary visual feature extracted from the tensor maps of all images guided by a unit's activation, and has been validated by its signal saliency, biological predictability, and enrichment among its unit's top-activating images. The discovered LVF units are near monosemantic, at approximately 1.04 features per unit. Feature recovery is robust to stricter significance thresholds: 4,828 features (70.3%) hold at $q \leq .01$ and 3,412 (49.7%) at $q \leq .001$ under Benjamini – Hochberg correction. Recovery is also strong in effect size: 3,640 features (53.0%) reach an enrichment of 3 times their base rate among their unit's top-activating images, and 1,216 (17.7%) reach 5 times.

Which of the factors of dataset, ROI, participant identity, and channel identity drove the quantity and selectivity of recovered features was tested. Selectivity is dominated by channel

identity, which explains 47% of the variance in feature-wise enrichment (partial $\eta^2 = .468$, model $R^2 = .569$), while quantity is co-dominated by ROI and channel identity (ROI partial deviance = .213, channel partial deviance = .201; Poisson pseudo- $R^2 = .393$). Recovery is stable across ROIs with the exception of hV4 (364 features, 5.3%), but distributes unevenly across channels, from 1,976 relative depth features (28.8%) to 494 opponent colour features (7.2%). The recovered skeleton features are the most selective, at a median enrichment of 6.28 times over base rate against 3.51 times for the second best, structured edges. The recovered depth features, though the most numerous, are the least selective, at a median of 2.57 times.

Recovered primary features are mostly interpretable. Across the LVF stream, 4,310 recovered units (65.3%) reach an apparent interpretability score (AIS) of 0.5 and 3,085 (46.7%) reach 0.8, with a median AIS of 0.700. Interpretability is stable across datasets, slightly higher in BEGD (69.4% at $\text{AIS} \geq 0.5$) than in the two NSD replications (58.2% and 60.2%). Interpretability is likewise dominated by channel identity (partial $\eta^2 = .098$, $F = 119.6$, $p < .001$), with dataset, ROI, and participant contributing little. The most interpretable features come from the oriented energy and relative depth channels (median AIS 1.000 and 0.825), while the recovered skeleton features, though the most selective, are the least interpretable, at a median AIS of 0.350. Primary visual representations were likewise discovered and interpreted, at 6,870 representations over 6,603 units, distributed across all seven expert proxy channels and all tested regions. LVF feature-wise and unit-wise results are summarised in Tables 4 and 5.

Table 4

Low-Level Visual Feature Recovery and Interpretation by Channel (Feature-Wise)

Channel	Recovery		Interpretation		
	Features	Med. enr.	Med. AIS	$\text{AIS} \geq .5$	$\text{AIS} \geq .8$
Skeleton	1,023	6.28	0.350	39.9%	21.3%
Structured edges	1,542	3.51	0.500	53.6%	34.0%
Symmetry	509	3.19	1.000	81.5%	79.4%
Oriented energy / SF	756	3.08	1.000	80.4%	68.1%
Opponent colour	494	2.78	0.850	83.0%	58.1%
Saliency	570	2.60	0.600	62.1%	37.9%
Relative depth	1,976	2.57	0.850	75.4%	54.9%

Channel	Recovery		Interpretation		
	Features	Med. enr.	Med. AIS	AIS $\geq .5$	AIS $\geq .8$
Overall	6,870	3.08	0.700	65.3%	46.7%

Note. Results are pooled across the two datasets. Med. enr. = median enrichment over base rate; Med. AIS = median apparent interpretability score; SF = spatial frequency.

Table 5

Low-Level Visual Feature Recovery and Interpretation by Channel (Unit-Wise)

Channel	Recovery		Interpretation		
	Units	Med. enr.	Med. AIS	AIS $\geq .5$	AIS $\geq .8$
Skeleton	1,023	6.28	0.350	39.9%	21.3%
Structured edges	1,509	3.52	0.500	53.3%	33.9%
Symmetry	502	3.18	1.000	81.5%	79.3%
Oriented energy / SF	642	3.06	1.000	79.9%	67.3%
Opponent colour	466	2.78	0.850	83.5%	58.4%
Saliency	537	2.60	0.600	62.8%	38.0%
Relative depth	1,924	2.57	0.825	75.3%	54.6%
Overall	6,603	3.09	0.700	65.3%	46.7%

Note. Results are pooled across the two datasets, and each recovered unit is represented by its feature with the highest enrichment. Med. enr. = median enrichment over base rate, expressed as a multiple; Med. AIS = median apparent interpretability score; SF = spatial frequency.

Dissociation of selectivity and interpretability in primary visual features (H1)

The selectivity and interpretability of primary visual features are essentially uncoupled across the 6,870 recovered primary features, with Spearman $\rho = -.084$ between AUROC and AIS and $\rho = -.090$ between enrichment and AIS. A double dissociation is observed upon further inspection. Taking the median AUROC (0.755) and median AIS (0.700) as reference points, 1,706 features are recovered selectively yet explained poorly, and 1,900 are explained well yet recovered unselectively, against 1,729 and 1,535 features that are high or low on both. The dissociation is most evident within specific channels: of the 1,023 recovered skeleton features, 735 are selectively recovered but poorly explained, and only 4 are the reverse, while relative

depth shows the opposite pattern, with 859 of its 1,976 features well explained but unselectively recovered against 145 in the reverse case. Channel identity dominates selectivity and interpretability independently, but with distinct effects: channel alone explains $R^2 = .378$ of the variance in AUROC and $R^2 = .119$ in AIS (partial $\eta^2 = .329$ and $.099$, controlling for dataset, ROI, and participant).

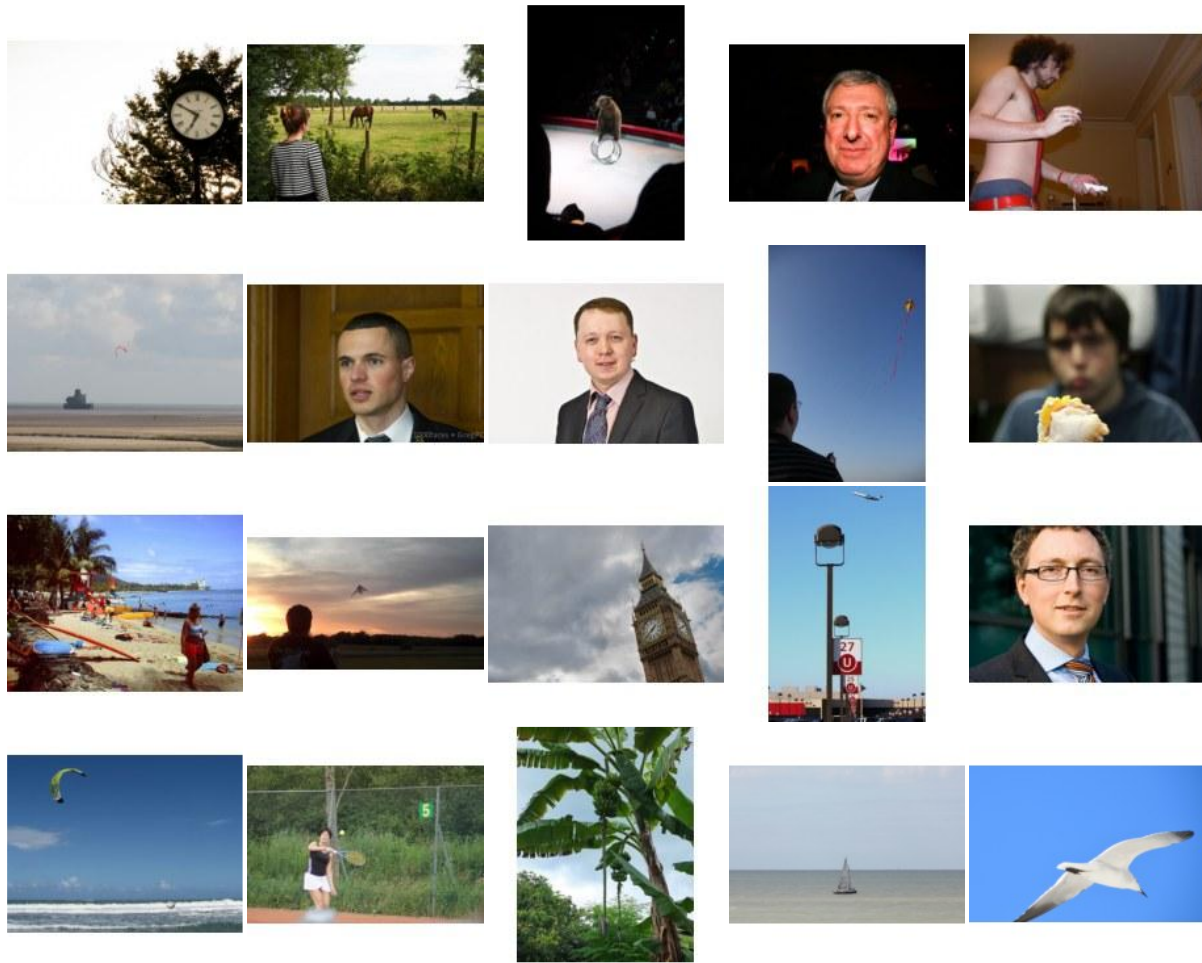
Interpreting primary features as semantic risks misattribution (H2)

To examine whether interpreting primary visual features as semantic poses a genuine risk to feature attribution and authentic interpretability, the units learnt from primary areas were separately examined under the HVF stream. Across the seven primary and intermediate regions, the stream recovered 30,107 enriched semantic features over 12,308 units, of which 177 and 159 came from the two NSD replications and 11,972 from BEGD, showing strong dataset bias. The recovered features also lack variety, being dominated by one semantic feature: the visual concept of *sky* was recovered in 6,744 of the 12,308 units (54.8%), against 2,386 units for the next most frequent feature, outdoor natural scene (19.4%), and 2,133 for face (17.3%). Against a global prevalence of .093 across the image bank, *sky* was recovered at a median enrichment of 7.24 times over the base rate, exceeding the median enrichment of the HVF stream as a whole, and reached a median AIS of 0.65, qualifying as a benchmark semantic representation discoverable in primary areas that is both selective and interpretable.

To test whether *sky* is the optimal visual feature for these fMRI features to be attributed to, the explanatory power of *sky* is examined in comparison with primary features. Of the 6,744 units from which *sky* was semantically recovered, 388 also yielded a total of 411 recovered primary features under the LVF stream. Across these 411 primary-semantic feature pairs, in the median case, two-thirds of the *sky*-absent images among a unit's top-20 activating images still express its primary feature (0.667), and the primary feature is present in more of the top-activating sample than *sky* is (0.800 against 0.650). The semantic feature therefore accounts for less of a unit's top-activating sample than the primary feature recovered from the same unit. Figures 2 and 3 qualitatively demonstrate the finding within a unit, between its recovered semantic (*sky*) and primary visual features (edge), and the corresponding explanations. The primary feature is more interpretable and explainable compared with the semantic feature.

Figure 2

Top-20 Activating Images of a Unit Learnt From a Primary Visual Area, Recovered Under the HVF Stream



Note. The unit was learnt from V2d of participant 07 in BEGD (u325). The panel shows the 20 images that most strongly activate the unit, ordered by activation strength. Under the HVF stream, a single enriched semantic feature was recovered from this unit, sky, at an enrichment of 5.57 times over base rate ($q = .007$), with no depleted features. The natural-language explanation generated from it reads "Images with open blue sky or broad sky-dominated outdoor views" (AIS = 0.450). Sky is present in 10 of the 20 images; the remainder include indoor human portraits. HVF = high-level visual feature; AIS = apparent interpretability score; BEGD = Brain Encoder Generated Dataset.

Figure 3

Signal-Expression Maps of the Top-20 Activating Images of the Same Unit, Recovered Under the LVF Stream



Note. The unit is the same as in Figure 2 (BEGD, participant 07, V2d, u325), examined under the LVF stream through the structured-edges channel. The figure shows the signed spatial expression map of the recovered feature over the same images, in which red marks local contribution above and blue below the channel prior, normalised to the 99th percentile and independent of the unit's activation, with the shared colour bar below. The recovered feature is enriched 3.82 times over the base rate ($q = 1.57 \times 10^{-5}$) with AUROC = 0.809. The natural-language explanation reads "A dense pattern of fine horizontal and vertical boundary fragments, strongest in broad textured background regions and around thin upright contour structures, with little emphasis on closed object outlines" (AIS = 1.000). The primary explanation succeeded in describing the unit's activation on all instances, while images in which the contrasting background is not sky include an indoor floor at rank 3, a uniform colour field at ranks 4, 5, and 8, and a continuous texture at ranks 7 and 17. LVF = low-level visual feature; AUROC = area under the receiver operating characteristic curve; AIS = apparent interpretability score; BEGD = Brain Encoder Generated Dataset.

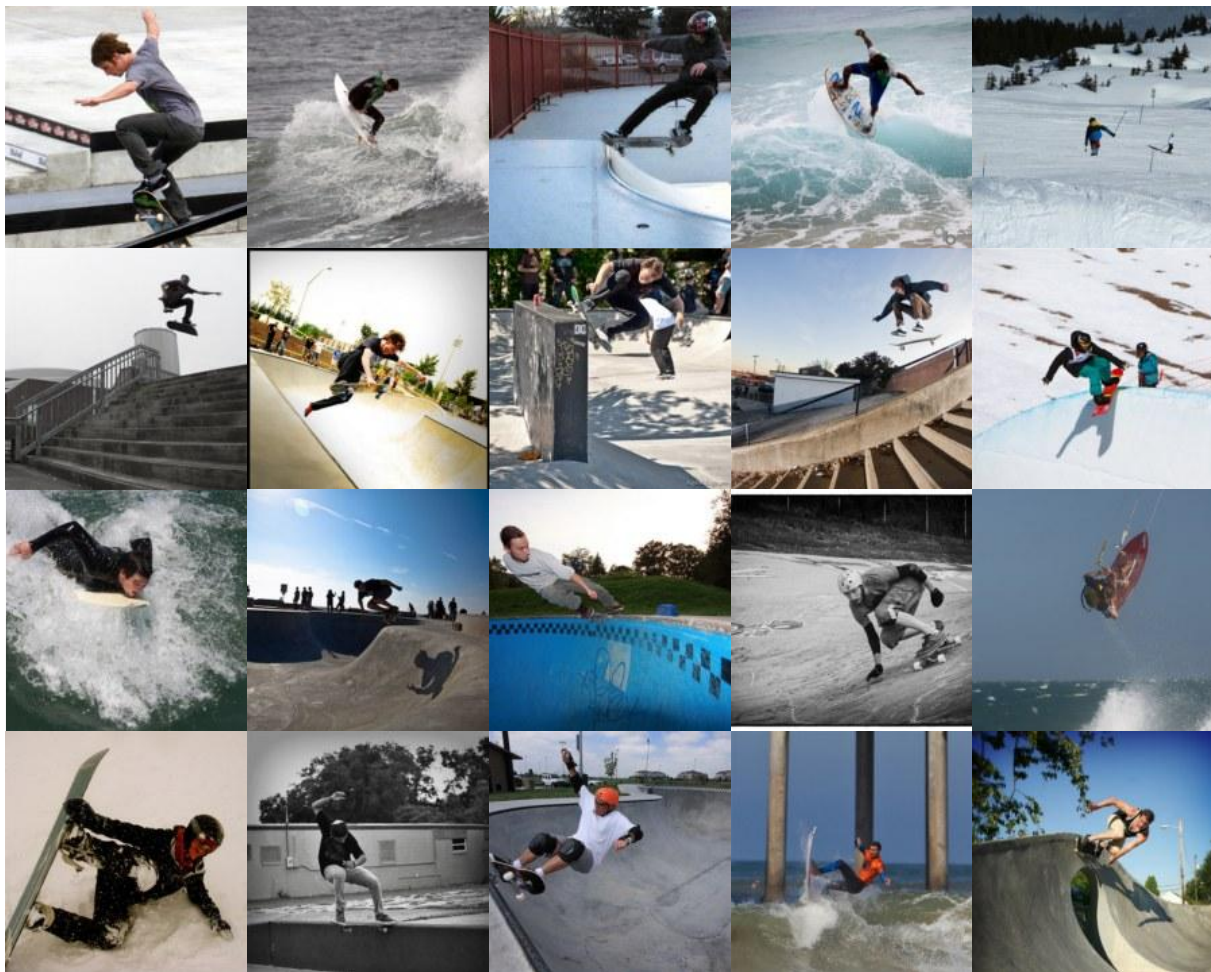
Qualitative Demonstrations of Discovered Visual Representations

To showcase individual discovered visual representations, two units were selected for detailed presentation (Figures 4 and 5). Each is shown by the 20 images that most strongly activate it, ordered by activation strength, together with the features recovered from it, the natural-language explanation for the recovered visual features, and the apparent interpretability score. The primary visual feature is additionally visualized by the spatial expression map of the recovered feature over the same images. The corresponding fMRI activity pattern of each unit

is further shown as a cortical surface map (Figure 6), with its activation visualized and located within its source ROI. The full set of showcases, spanning across both streams and ROIs, is provided in the Appendices. The semantic representations recovered from each category-selective region were additionally summarised by pooling the recovered semantic features of all units learnt from that region in NSD. Figure 7 shows the most frequently recovered features per region and Figure 8 the most strongly enriched ones.

Figure 4

Top-20 Activating Images of a Unit Learnt From a Category-Selective Region

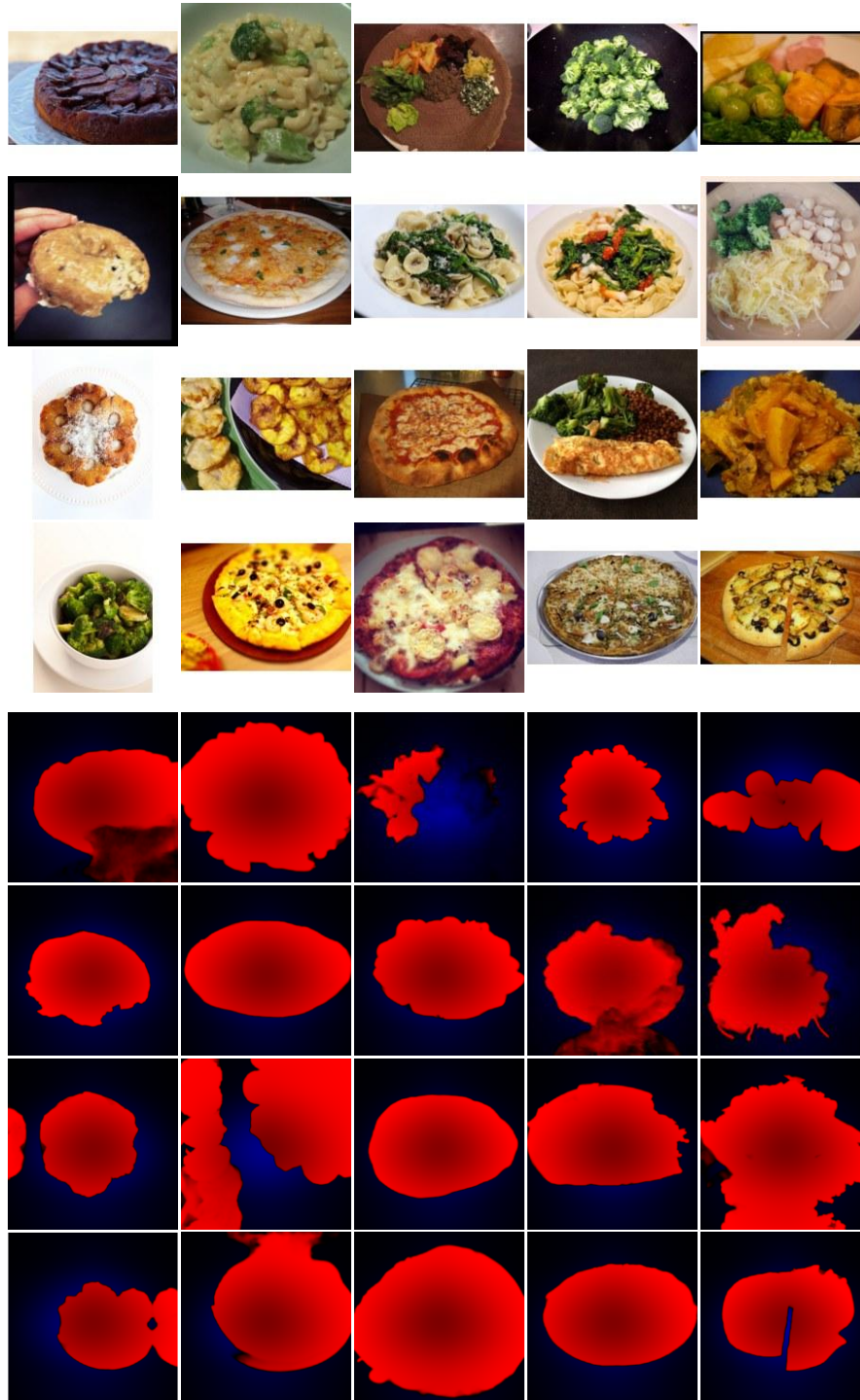


Note. The unit was learnt from the extrastriate body area of participant 05 in NSD (u34), examined under the HVF stream. The panel shows the 20 images that most strongly activate the unit, ordered by activation strength. Fourteen enriched and two depleted semantic features were recovered. The enriched features span body and person content (human/person, face, body part/hand, person-centered), posture and action (locomotion, biological motion, visible non-static motion, object-directed manipulation, reaching/grasp-related posture, lowered posture), scene structure (navigable/path-like scene), and object identity (skateboard, skateboarding, one). Enrichment ranges from 1.76 to 30.11 times over base rate, the highest being skateboarding at a prevalence of .020 and the lowest one at a prevalence of .511. Both depleted features, static and static/stationary, are absent from the top-activating sample entirely despite prevalences of .695 and .697. The natural-language explanation reads "The unit responds to images of people in active, body-focused motion, especially skateboarding, surfing, and similar reaching, leaning, or manipulating

movements in motion-filled, path-like settings, while tending not to activate for static, brown scenes" (AIS = 1.000). HVF = high-level visual feature; AIS = apparent interpretability score; NSD = Natural Scenes Dataset.

Figure 5

Top-20 Activating Images and Signal-Expression Maps of a Unit Learnt From a Primary Visual Area

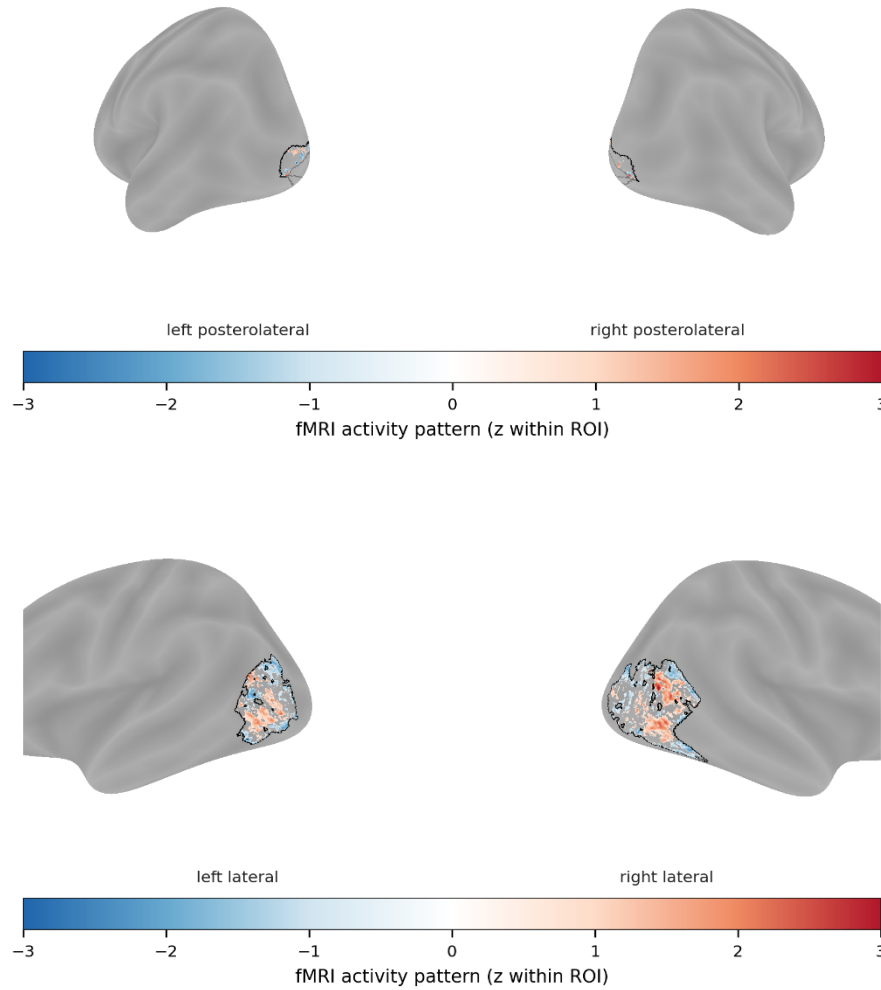


Note. The unit was learnt from V2d of participant 02 in BEGD (u125), examined under the LVF stream through the saliency channel. Top: the 20 images that most strongly activate the unit, ordered by activation strength. Bottom: the signed spatial expression map of the recovered feature over the same images, in which red marks local contribution above and blue below the

channel prior, normalised to the 99th percentile of the absolute value and independent of the unit's activation. The recovered feature is enriched 4.89 times over the base rate ($q = 1.14 \times 10^{-11}$) with AUROC = 0.946. The natural-language explanation reads "A strong figure-ground signal concentrated on the main rounded food mass, emphasizing the outer silhouette and central piled surface while suppressing the background" (AIS = 0.850). LVF = low-level visual feature; AUROC = area under the receiver operating characteristic curve; AIS = apparent interpretability score; BEGD = Brain Encoder Generated Dataset.

Figure 6

Cortical Localisation of the fMRI Activity Pattern of the Two Showcasing Units



Note. Each map shows the fMRI activity pattern denoted by one sparse dictionary unit, plotted as the unit's contribution to reconstructing the response of its source region and standardised within that region. Red marks vertices whose response increases when the unit is active and blue marks those whose response decreases. Top row: the V2d unit of Figure 5 (BEGD, participant 02, u125; 1,243 vertices), shown in posterolateral view with the camera 35° lateral of the pure posterior axis and 5° above the horizontal, mirrored between hemispheres. Vertices below $|z| = 0.88$ are unmarked, leaving 27% of the region's vertices marked; the pattern is heavy-tailed (excess kurtosis = 5.42), so a minority of vertices carries most of the weight. Bottom row: the extrastriate body area unit of Figure 4 (NSD, participant 05, u34; 9,359 vertices), shown in lateral view. Vertices below $|z| = 0.61$ are unmarked, leaving 55% marked, with the weight distributed broadly across the region (excess kurtosis = -0.25). Thresholds are set at one quarter of each unit's own 99th percentile, so the marked extent is comparable in meaning across units despite the difference in the

underlying weight distributions. Surfaces are the inflated fsaverage template with sulcal depth as a greyscale underlay. The source region boundary is drawn in black; adjacent V1v and V1d boundaries are drawn in grey in the top row. Colour limits are shared at ± 3 . Rendered with Nilearn 0.12.1. ROI = region of interest; BEGD = Brain Encoder Generated Dataset; NSD = Natural Scenes Dataset.

Figure 7

Most Frequently Recovered Semantic Features by Category-Selective Region



Note. Each panel pools the semantic features recovered from all units learnt from one region across the eight NSD participants (12 of the 13 category-selective regions, excluding mfs-words), with the 14 most frequently recovered features shown per region. Font size is proportional to the number of units from which a feature was recovered; word position, orientation and colour carry no information. Features recovered from BEGD are excluded, so the figure reflects measured responses only. NSD = Natural Scenes Dataset; BEGD = Brain Encoder Generated Dataset.

Figure 8

Most Strongly Enriched Semantic Features by Category-Selective Region

Discussion

The Selectivity and Interpretability of Visual Features are Doubly Dissociated

H1 was tested and confirmed, as selectivity and interpretability are essentially uncoupled across the 6,870 recovered primary visual features, and under two different selectivity measures. How selectively a unit encodes a primary visual feature is therefore hardly

informative of how well that feature can be interpreted. The dissociation runs in both directions and is sharpest within channels: of the recovered skeleton features, 735 are selectively recovered but poorly explained against only 4 in the reverse case, while relative depth shows the opposite pattern, with 859 features well explained but unselectively recovered against 145 in the reverse. Upon further analysis, the interpretability and selectivity of visual features are revealed as separate inherent characteristics of feature modality, whose impacts can be disentangled. These findings validate the architectural designs of the present work as well as flag a genuine concern for interpretability work: that discovering a visual feature and interpreting it are distinct objectives to be pursued. Furthermore, interpretation should be rather set-dependent on visual latent discovery to ensure the genuineness of discovery, instead of the reverse.

H1 was tested because prior automated interpretability makes recovery conditional on interpretability. In the prevailing automatic interpretability paradigm, a unit's most strongly activating images are shown to a VLM, and the VLM is tasked with generating an explanation under an open-ended prompt. As the explanation is expected to have captured the visual feature the unit encodes, a visual feature must be salient, understandable, and namable to the VLM before it can be recovered. Arguably, this methodology is defensible only if the visual feature is highly interpretable to be noticed and comprehended by the VLM. Due to agentic bias, VLMs are empirically validated in capturing semantic visual categories (Tavakoli et al., 2017), while being shown to be less sensitive to primary features such as orientation, spatial frequency, and colour (Dahou et al., 2025; Rahmanzadehgervi et al., 2024). This makes decoupling interpretation and feature recovery a precondition for discovering primary visual representations. This consideration underpins novel methods primarily relying on statistics to resolve and validate the recovered visual features to reduce the impact of agentic bias in representation discovery. The interpretation is also segregated from feature recovery as post-hoc sense-making to emphasize genuine interpretability above mere plausibility (Jacovi & Goldberg, 2020).

Misattribution of Primary Representations Under a Semantic Proxy

H2 was quantitatively examined in the example of *sky*, the most frequently recovered semantic feature in units learnt from primary visual areas. This semantic concept was recovered in more than half of them, was strongly enriched, and yielded coherent explanations. Taken at face value, this serves as evidence of semantic representation of *sky* in V1 through hV4. Prior knowledge of the visual system, however, argues against such an encoding, since the primary visual cortex is characterised by tuning to orientation, spatial frequency, and position (Hubel &

Wiesel, 1962; Kay et al., 2008), whereas category-level selectivity emerges in downstream extrastriate regions (Kanwisher et al., 1997). Past literature that only discovers semantic representations has no way to register the conflict (Wasserman et al., 2025), because it wrongly assumes representational modality while having no alternative account to compare against. As further quantitative analysis has shown, *sky* does not explain a non-trivial part of the stimuli to which these fMRI activity patterns in primary areas respond. In the meantime, primary visual features are proven more selective and interpretable when accounting for the stimuli. As a result, the concept of *sky* is more likely to be a semantic covariate of other primary features, for example, as a prototypical exemplar of a continuous smooth space low in spatial frequency, or by contrasting contours against a uniform background.

H2 was tested to verify another risk to genuine representation discovery. Where the dissociation causes representations to be missed, misattribution causes them to be found but explained incorrectly, while nothing in the evaluating metrics indicates that anything has gone astray. A unit high in both selectivity and interpretability may still be misattributed. When a unit is probed under a semantic proxy while its representation is in fact low-level, it may be explained by a semantic exemplar of the low-level feature rather than by the feature itself. This phenomenon is referred to here as masquerade, confirmable if a recovered feature does not meet the a priori expectation of knowledge, and if it lacks explanatory power compared with another more plausible source of attribution. The present work therefore suggests masquerade as another obstacle to genuine discovery and interpretation of visual representations. Consistent with the opinions posted in earlier studies, interpretability alone cannot serve as evidence that the correct feature has been recovered (Jacovi & Goldberg, 2020), and neither can selectivity. A genuine discovery also requires its methodology to be falsifiable on grounds of a priori knowledge and alternative hypotheses.

Relation to Previous Research and Contribution

Overall, the present work has contributed to a holistic, data-driven approach to discovering and interpreting visual representations in human fMRI data, particularly addressing research gaps in discovering and interpreting primary visual representations. While the present literature in this field is scarce, the closest prior work is BrainExplore (Wasserman et al., 2025), which shares this study's dataset, its usage of SAE to decompose fMRI activity patterns, and its unit of analysis. While both the present work and BrainExplore aim at automatic discovery of visual representations, BrainExplore assumed, searched for, and reported only semantic representations across the visual hierarchy. The prior work also observed strong heterogeneity in the interpretable fMRI patterns discovered across ROIs: the proportion of explainable fMRI

patterns was 0.6% in V1, 0.9% in V2, and 1.3% in V3, against 14.5% in the EBA. This finding was interpreted as higher-level regions showing greater interpretability than early visual areas, and was attributed to those regions encoding higher-level semantics that are both easier to decompose and easier to describe in natural language within their pipeline. Motivated by this finding, the present work set out to discover primary visual features with a method that does not condition recovery on an agent's ability to describe a pattern. The present results indicate that decomposition through SDL is not what limits recovery in early visual areas. As 6,870 validated primary representations were discovered under the same SDL framework, the strategies for discovering primary visual representations largely complemented the scope of visual representation discovery in previous research (Wasserman et al., 2025). The present results are consistent with prior work in finding few interpretable patterns in primary areas under a semantically grounded pipeline, and further indicate that this outcome is attributable to the visual feature recovery strategies.

Furthermore, the present work has made three methodological contributions relative to the prior work (Wasserman et al., 2025). First, the present work has contributed to a feature discovery pipeline that separates feature interpretability from feature presence. By doing so, primary and subtle semantic representations such as texture and color can be more fairly discovered and explained. Second, encoding is treated as bidirectional in the present work. The present work treats negative encoding as a plausible way of feature presentation following literature findings (Franke et al., 2025), and tests bidirectional feature representations, which is an absent design in previous literature. Under mirrored feature evaluation criteria, the present work also discovered 9,603 depleted semantic features across 3,058 units. Among these, 271 units approximate pure negative encoding by carrying no enriched features at all. Despite a relatively small quantity, the present work has also contributed preliminary methods and findings on negative semantic encoding through fMRI activity. Third, the present work has explicitly addressed the ambiguity in feature attribution. Attributing fMRI activity to certain conceptual representations can typically be ambiguous since genuine polysemantic representations can be mixed with natural statistical covariation between features. In the LVF stream, the present work has proposed a method to resolve this ambiguity by validating each possible source of attribution by measuring its residualized biological predictability through MVPA. This successfully identified a standalone, optimized source of attribution (1.04 features on average) per unit. Also, because such a basis for MVPA is lacking for the binary semantic features, the present work has preserved all validated semantic features for interpretation, rather than seeking a single best-matching one.

Limitations

The most substantial limitation concerns the evaluation of feature selectivity, as a feature’s natural prevalence is a strong determinant of its measured selectivity. The natural prevalence of features correlates strongly with feature selectivity across channels (channel-level Spearman $\rho = .77$): the rarer a feature type naturally is, the more selectively the unit that recovers it activates on the images expressing it. This is a result of the architectural design of SAE: when a feature is far denser than the sparsity constraint of SAE, the activation of the unit capturing it tends to be less selective toward all images containing the feature, and its selectivity is impaired. Depth and saliency features are typical examples. These two types of features are highly prevalent in the stimuli of the present work, because the images (MS-COCO; Lin et al., 2014) consist of complex everyday scenes with multiple objects set against visible, cluttered backgrounds, so a foreground-background depth contrast and a figure-ground saliency structure are present in almost every image. Although the source unit still fires on only a sparse set of images, that set is comparatively less selective toward how strongly the signal is present. Nonetheless, the recovered features remain genuine and highly interpretable. While the architectural sparsity constraint drives the discovered fMRI patterns to be monosemantic (Bricken et al., 2023), which benefits their interpretability, the present work also suggests that sparsity favours those that represent a visual feature with a natural prevalence in the dataset that aligns with it.

The coverage of primary visual features is also not exhaustive. Seven expert proxy channels were used, denoting all types of primary visual features the present work has set out to discover. Other visual properties, such as texture and curvature, were not examined despite evidence of their encoding in the human visual cortex (Freeman et al., 2013; Yue et al., 2020). As a result, the scope of discovery in the present work is restricted to the feature modalities it has expected, and the units falling short of validation cannot be fully treated as noise. The explanations generated for each recovered feature were evaluated by a separate VLM rather than by human raters, and no inter-rater agreement between human and model judgements was estimated. This is a result of the massive scale of the dataset, which is beyond what could be manually rated under the constraints of resources. The apparent interpretability score should therefore be understood as indicating how much a discovered visual representation can be interpreted by a VLM instead of a human.

Finally, the sample size of the present work was small ($N = 8$). Although the number of trials per participant is high, the anatomical differences of participants resulted in different voxel mappings, hindering cross-participant modeling for a more generalizable result. The

present study should therefore be viewed as a participant-wise, in-depth decoding work, which discovers individualized visual representations with limited generalizability.

Future Directions

The present work has planned an extension to visual reconstruction as a major future direction, with a held-out set of 100 image–fMRI pairs per participant reserved from both datasets. Reconstructing visual inputs from fMRI activity is an established line of work, from early decoding of retinotopic patterns and simple shapes (Miyawaki et al., 2008; Thirion et al., 2006) to current approaches that map responses into a latent space and generate images with a diffusion model (Ozcelik & VanRullen, 2023; Takagi & Nishimoto, 2023). A recurring criticism of these approaches is their lack of mechanistic interpretability (Rakhimberdina et al., 2021; Rudin, 2019). As the mapping from brain activity to latent representation is typically learnt end-to-end, a reconstruction can be accurate while its internal mechanism and causal contribution remain unclear. The past reconstruction literature is also focused on reconstructing semantic visual features, while the primary features are under-mobilised (Ozcelik et al., 2022; He et al., 2026). The dictionary built by the present work has high potential in addressing these two concerns at once. Because each unit encodes a separately validated and interpreted visual representation, and because reconstruction is linear in the units, the contribution of any unit to a reconstructed image can be isolated. Units can be removed or added individually, and the causal consequence for the reconstruction can be observed. These advantages could empower an interpretable reconstruction model that allows ablation tests of causal mechanisms. The primary visual features may also provide critical information, such as the spatial layout and shape of visual contents, to improve the fidelity of visual reconstruction.

Conclusion

The present work discovered a total of 120,584 primary and semantic visual representations encoded in human fMRI activity across the visual hierarchy through sparse dictionary learning. Localised fMRI response patterns were decomposed by unsupervised learning, and the visual features encoded by each pattern were recovered under objective statistical criteria, before being validated, explained, and visualised. The discovered representations are generally selective and interpretable, revealing the diversity and richness of the visual representations carried by population-level neural activity in the human visual system. The work further addressed a gap in the discovery of primary visual representations, and contributes a method that reduces the dependence of discovery on an interpreting agent and emphasises the genuine interpretability of features.

Declaration of Generative AI and AI-Assisted Technologies

Generative AI and AI-assisted technologies were used in the present work for two distinct purposes. Vision-language and language models were used as instruments of the analysis itself, for image captioning, ontology learning and canonicalization, obtaining natural-language explanations, and evaluating those explanations. The detailed usages, including the model versions and the prompts, are specified in the Method section and in the Appendices.

Separately, during the preparation of this thesis, the author used ChatGPT 5.4 and Claude Opus 5 to support the systematization of background literature, programming and code checking, drafting and restructuring text, checking the logical coherence of arguments, and proofreading. After using these tools, the author critically reviewed, edited, and independently revised the content as necessary, verified every cited source against the original publication, and took full responsibility for the content of the thesis. No confidential or participant-identifying material was entered into any tool.

References

- Adeli, H., Minni, S., & Kriegeskorte, N. (2025). *Transformer brain encoders explain human high-level visual responses*. arXiv. <https://doi.org/10.48550/arXiv.2505.17329>
- Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., & Kay, K. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1), 116–126. <https://doi.org/10.1038/s41593-021-00962-x>
- Averbeck, B. B., Latham, P. E., & Pouget, A. (2006). Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, 7(5), 358–366. <https://doi.org/10.1038/nrn1888>
- Ayzenberg, V., Kamps, F. S., Dilks, D. D., & Lourenco, S. F. (2022). Skeletal representations of shape in the human visual cortex. *Neuropsychologia*, 164, Article 108092. <https://doi.org/10.1016/j.neuropsychologia.2021.108092>
- Bao, P., She, L., McGill, M., & Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, 583(7814), 103–108. <https://doi.org/10.1038/s41586-020-2350-5>
- Barnett, M. A., Aguirre, G. K., & Brainard, D. H. (2021). A quadratic model captures the human V1 response to variations in chromatic direction and contrast. *eLife*, 10, Article e65590. <https://doi.org/10.7554/eLife.65590>
- Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3319–3327). IEEE. <https://doi.org/10.1109/CVPR.2017.354>
- Bereska, L., & Gavves, E. (2024). *Mechanistic interpretability for AI safety: A review*. arXiv. <https://doi.org/10.48550/arXiv.2404.14082>
- Bersch, D., Vilas, M. G., Saba-Sadiya, S., Schaumlöffel, T., Dwivedi, K., Sartzetaki, C., Cichy, R. M., & Roig, G. (2025). Net2Brain: A toolbox to compare artificial vision models with human brain responses. *Frontiers in Neuroinformatics*, 19, Article 1515873. <https://doi.org/10.3389/fninf.2025.1515873>
- Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., Sutskever, I., Leike, J., Wu, J., & Saunders, W. (2023). *Language models can explain neurons in language models*. OpenAI. <https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html>

- Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askeell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., ... Olah, C. (2023). *Towards monosemanticity: Decomposing language models with dictionary learning*. Transformer Circuits Thread. <https://transformer-circuits.pub/2023/monosemantic-features/index.html>
- Brouwer, G. J., & Heeger, D. J. (2009). Decoding and reconstructing color from responses in human visual cortex. *The Journal of Neuroscience*, 29(44), 13992–14003. <https://doi.org/10.1523/JNEUROSCI.3577-09.2009>
- Bussmann, B., Nabeshima, N., Karvonen, A., & Nanda, N. (2025). Learning multi-level features with Matryoshka sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning* (Vol. 267, pp. 6077–6101). PMLR.
- Cohen, L., Dehaene, S., Naccache, L., Lehéricy, S., Dehaene-Lambertz, G., Hénaff, M.-A., & Michel, F. (2000). The visual word form area. *Brain*, 123(2), 291–307. <https://doi.org/10.1093/brain/123.2.291>
- Cui, J., Zhang, Q., Wang, Y., & Wang, Y. (2025). *On the limits of sparse autoencoders: A theoretical framework and reweighted remedy*. arXiv. <https://doi.org/10.48550/arXiv.2506.15963>
- Cunningham, H., Ewart, A., Riggs, L., Huben, R., & Sharkey, L. (2023). *Sparse autoencoders find highly interpretable features in language models*. arXiv. <https://doi.org/10.48550/arXiv.2309.08600>
- Dahou, Y., Huynh, N. D., Le-Khac, P. H., Para, W. R., Singh, A., & Narayan, S. (2025). Vision-language models can't see the obvious. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 24159–24169). IEEE. <https://doi.org/10.1109/ICCV51701.2025.02239>
- Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D., & DiCarlo, J. J. (2020). Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. *Advances in Neural Information Processing Systems*, 33, 13073–13087.
- De Valois, R. L., Albrecht, D. G., & Thorell, L. G. (1982). Spatial frequency selectivity of cells in macaque visual cortex. *Vision Research*, 22(5), 545–559. [https://doi.org/10.1016/0042-6989\(82\)90113-4](https://doi.org/10.1016/0042-6989(82)90113-4)
- DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3), 415–434. <https://doi.org/10.1016/j.neuron.2012.01.010>

- Dollár, P., & Zitnick, C. L. (2013). Structured forests for fast edge detection. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1841–1848). IEEE.
<https://doi.org/10.1109/ICCV.2013.231>
- Dollár, P., & Zitnick, C. L. (2015). Fast edge detection using structured forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8), 1558–1570.
<https://doi.org/10.1109/TPAMI.2014.2377715>
- Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, 293(5539), 2470–2473.
<https://doi.org/10.1126/science.1063414>
- Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., & Olah, C. (2022). *Toy models of superposition*. arXiv.
<https://doi.org/10.48550/arXiv.2209.10652>
- Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, 392(6676), 598–601. <https://doi.org/10.1038/33402>
- Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1), 1–47. <https://doi.org/10.1093/cercor/1.1.1>
- Felsen, G., & Dan, Y. (2005). A natural approach to studying vision. *Nature Neuroscience*, 8(12), 1643–1646. <https://doi.org/10.1038/nn1608>
- Franke, K., Karantzas, N., Willeke, K., Diamantaki, M., Ramakrishnan, K., Bedel, H. A., Elumalai, P., Restivo, K., Fahey, P., Nealley, C., Shinn, T., Garcia, G., Patel, S., Ecker, A., Walker, E. Y., Froudarakis, E., Sanborn, S., Sinz, F. H., & Tolias, A. S. (2025). *Dual-feature selectivity enables bidirectional coding in visual cortical neurons*. bioRxiv. <https://doi.org/10.1101/2025.07.16.665209>
- Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience*, 16(7), 974–981. <https://doi.org/10.1038/nn.3402>
- Gifford, A. T., Lahner, B., Saba-Sadiya, S., Vilas, M. G., Lascelles, A., Oliva, A., Kay, K., Roig, G., & Cichy, R. M. (2023). *The Algonauts Project 2023 challenge: How the human brain makes sense of natural scenes*. arXiv.
<https://doi.org/10.48550/arXiv.2301.03198>
- Goddard, E., Mannion, D. J., McDonald, J. S., Solomon, S. G., & Clifford, C. W. G. (2010). Combination of subcortical color channels in human visual cortex. *Journal of Vision*, 10(5), Article 25. <https://doi.org/10.1167/10.5.25>

- Gujral, O., Bafna, M., Alm, E., & Berger, B. (2025). Sparse autoencoders uncover biologically interpretable features in protein language model representations. *Proceedings of the National Academy of Sciences*, 122(34), Article e2506316122. <https://doi.org/10.1073/pnas.2506316122>
- Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539), 2425–2430. <https://doi.org/10.1126/science.1063736>
- He, W., Wang, J., Li, H., Chi, H., & Zhang, B. (2026). Balancing semantic and structural decoding for fMRI-to-image reconstruction. *Expert Systems with Applications*, 298, Article 129836. <https://doi.org/10.1016/j.eswa.2025.129836>
- Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1), 106–154. <https://doi.org/10.1113/jphysiol.1962.sp006837>
- Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6), 1210–1224. <https://doi.org/10.1016/j.neuron.2012.10.014>
- Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4198–4205). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.386>
- Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8(5), 679–685. <https://doi.org/10.1038/nn1444>
- Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, 17(11), 4302–4311. <https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997>
- Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, 452(7185), 352–355. <https://doi.org/10.1038/nature06713>
- Keefe, B. D., Gouws, A. D., Sheldon, A. A., Vernon, R. J. W., Lawrence, S. J. D., McKeefry, D. J., Wade, A. R., & Morland, A. B. (2018). Emergence of symmetry selectivity in the visual areas of the human brain: fMRI responses to symmetry presented in both

- frontoparallel and slanted planes. *Human Brain Mapping*, 39(10), 3813–3826.
<https://doi.org/10.1002/hbm.24211>
- Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6), 1126–1141.
<https://doi.org/10.1016/j.neuron.2008.10.043>
- Lewis, P. M., Ackland, H. M., Lowery, A. J., & Rosenfeld, J. V. (2015). Restoration of vision in blind individuals using bionic devices: A review with a focus on cortical visual prostheses. *Brain Research*, 1595, 51–73.
<https://doi.org/10.1016/j.brainres.2014.11.020>
- Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollár, P. (2014). *Microsoft COCO: Common objects in context*. arXiv. <https://doi.org/10.48550/arXiv.1405.0312>
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman.
- Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M., Morito, Y., Tanabe, H. C., Sadato, N., & Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5), 915–929.
<https://doi.org/10.1016/j.neuron.2008.11.004>
- Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, 56(2), 400–410.
<https://doi.org/10.1016/j.neuroimage.2010.07.073>
- Nasr, S., Polimeni, J. R., & Tootell, R. B. H. (2016). Interdigitated color- and disparity-selective columns within human visual cortical areas V2 and V3. *The Journal of Neuroscience*, 36(6), 1841–1857. <https://doi.org/10.1523/JNEUROSCI.3518-15.2016>
- Nasr, S., & Tootell, R. B. H. (2018). Visual field biases for near and far stimuli in disparity-selective columns in human visual cortex. *NeuroImage*, 168, 358–365.
<https://doi.org/10.1016/j.neuroimage.2016.09.012>
- Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10(9), 424–430. <https://doi.org/10.1016/j.tics.2006.07.005>
- Oikarinen, T., & Weng, T.-W. (2022). *CLIP-Dissect: Automatic description of neuron representations in deep vision networks*. arXiv.
<https://doi.org/10.48550/arXiv.2204.10965>

- Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23), 3311–3325.
[https://doi.org/10.1016/S0042-6989\(97\)00169-7](https://doi.org/10.1016/S0042-6989(97)00169-7)
- Ozcelik, F., Choksi, B., Mozafari, M., Reddy, L., & VanRullen, R. (2022). *Reconstruction of perceived images from fMRI patterns and semantic brain exploration using instance-conditioned GANs*. arXiv. <https://doi.org/10.48550/arXiv.2202.12692>
- Ozcelik, F., & VanRullen, R. (2023). Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, 13(1), Article 15666.
<https://doi.org/10.1038/s41598-023-42891-8>
- Panzeri, S., Macke, J. H., Gross, J., & Kayser, C. (2015). Neural population coding: Combining insights from microscopic and mass signals. *Trends in Cognitive Sciences*, 19(3), 162–172. <https://doi.org/10.1016/j.tics.2015.01.002>
- Papale, P., Leo, A., Cecchetti, L., Handjaras, G., Kay, K. N., Pietrini, P., & Ricciardi, E. (2018). Foreground-background segmentation revealed during natural image viewing. *eNeuro*, 5(3), Article ENEURO.0075-18.2018.
<https://doi.org/10.1523/ENEURO.0075-18.2018>
- Paulo, G., Mallen, A., Juang, C., & Belrose, N. (2025). Automatically interpreting millions of features in large language models. In *Proceedings of the 42nd International Conference on Machine Learning* (Vol. 267, pp. 48393–48421). PMLR.
- Peirce, J. W. (2015). Understanding mid-level representations in visual processing. *Journal of Vision*, 15(7), Article 5. <https://doi.org/10.1167/15.7.5>
- Poltoratski, S., Maier, A., Newton, A. T., & Tong, F. (2020). Resolving the spatial profile of figure enhancement in human V1 through population receptive field modeling. *The Journal of Neuroscience*, 40(16), 3292–3303.
<https://doi.org/10.1523/JNEUROSCI.2377-19.2020>
- Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., & Jagersand, M. (2020). U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognition*, 106, Article 107404. <https://doi.org/10.1016/j.patcog.2020.107404>
- Qiu, C., Burton, P. C., Kersten, D., & Olman, C. A. (2016). Responses in early visual areas to contour integration are context dependent. *Journal of Vision*, 16(8), Article 19.
<https://doi.org/10.1167/16.8.19>
- Rahmanzadehgervi, P., Bolton, L., Taesiri, M. R., & Nguyen, A. T. (2024). *Vision language models are blind: Failing to translate detailed visual features into words*. arXiv. <https://doi.org/10.48550/arXiv.2407.06581>

- Rakhimberdina, Z., Jodelet, Q., Liu, X., & Murata, T. (2021). Natural image reconstruction from fMRI using deep learning: A survey. *Frontiers in Neuroscience*, 15, Article 795488. <https://doi.org/10.3389/fnins.2021.795488>
- Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 12179–12188). IEEE. <https://doi.org/10.1109/ICCV48922.2021.01196>
- Roth, Z. N., Kay, K., & Merriam, E. P. (2022). Natural scene sampling reveals reliable coarse-scale orientation tuning in human V1. *Nature Communications*, 13, Article 6469. <https://doi.org/10.1038/s41467-022-34134-7>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Rust, N. C., & Movshon, J. A. (2005). In praise of artifice. *Nature Neuroscience*, 8(12), 1647–1650. <https://doi.org/10.1038/nn1606>
- Seo, A., Kim, B., Kwak, S., & Cho, M. (2022). Reflection and rotation symmetry detection via equivariant learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9539–9548). IEEE. <https://doi.org/10.1109/CVPR52688.2022.00932>
- Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., Biderman, S., Garriga-Alonso, A., Conmy, A., Nanda, N., Rumbelow, J., Wattenberg, M., Schoots, N., Miller, J., Michaud, E. J., ... McGrath, T. (2025). *Open problems in mechanistic interpretability*. arXiv. <https://doi.org/10.48550/arXiv.2501.16496>
- Sharpee, T. O. (2013). Computational identification of receptive fields. *Annual Review of Neuroscience*, 36(1), 103–120. <https://doi.org/10.1146/annurev-neuro-062012-170253>
- Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24(1), 1193–1216. <https://doi.org/10.1146/annurev.neuro.24.1.1193>
- Simoncelli, E., Young, R., Broderick, W., Fiquet, P.-É., Wang, Z., Kadkhodaie, Z., Parthasarathy, N., & Ward, B. (2025). *Pyrtools: Tools for multi-scale image processing* [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.17342148>
- Smet, K. A. G. (2020). Tutorial: The LuxPy Python toolbox for lighting and color science. *LEUKOS*, 16(3), 179–201. <https://doi.org/10.1080/15502724.2018.1518717>

- Stevens, S., Chao, W.-L., Berger-Wolf, T., & Su, Y. (2025). *Interpretable and testable vision features via sparse autoencoders*. arXiv. <https://doi.org/10.48550/arXiv.2502.06755>
- Takagi, Y., & Nishimoto, S. (2023). High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14453–14463). IEEE. <https://doi.org/10.1109/CVPR52729.2023.01389>
- Tavakoli, H. R., Shetty, R., Borji, A., & Laaksonen, J. (2017). Paying attention to descriptions generated by image captioning models. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2506–2515). IEEE. <https://doi.org/10.1109/ICCV.2017.272>
- Taylor, J., & Xu, Y. (2024). Using fMRI to examine nonlinear mixed selectivity tuning to task and category in the human brain. *Imaging Neuroscience*, 2. https://doi.org/10.1162/imag_a_00354
- Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Tamkin, A., Durmus, E., Hume, T., Mosconi, F., Freeman, C. D., ... Henighan, T. (2024). *Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet*. Transformer Circuits Thread. <https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html>
- Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J.-B., Lebihan, D., & Dehaene, S. (2006). Inverse retinotopy: Inferring the visual content of images from brain activation patterns. *NeuroImage*, 33(4), 1104–1116. <https://doi.org/10.1016/j.neuroimage.2006.06.062>
- van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., & Yu, T. (2014). scikit-image: Image processing in Python. *PeerJ*, 2, Article e453. <https://doi.org/10.7717/peerj.453>
- von der Heydt, R., Peterhans, E., & Baumgartner, G. (1984). Illusory contours and cortical neuron responses. *Science*, 224(4654), 1260–1262. <https://doi.org/10.1126/science.6539501>
- Wasserman, N., Cosarinsky, M., Golbari, Y., Oliva, A., Torralba, A., Rott Shaham, T., & Irani, M. (2025). *BrainExplore: Large-scale discovery of interpretable visual representations in the human brain*. arXiv. <https://doi.org/10.48550/arXiv.2512.08560>

- Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365.
<https://doi.org/10.1038/nn.4244>
- Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.
<https://doi.org/10.1073/pnas.1403112111>
- Yue, X., Robert, S., & Ungerleider, L. G. (2020). Curvature processing in human visual cortical areas. *NeuroImage*, *222*, Article 117295.
<https://doi.org/10.1016/j.neuroimage.2020.117295>
- Zhang, L.-Q., Mao, J., Aguirre, G. K., & Stocker, A. A. (2025). The tilt illusion arises from an efficient reallocation of neural coding resources at the contextual boundary. *Proceedings of the National Academy of Sciences*, *122*(17), Article e2421565122.
<https://doi.org/10.1073/pnas.2421565122>
- Zhou, H., Friedman, H. S., & von der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *The Journal of Neuroscience*, *20*(17), 6594–6611.
<https://doi.org/10.1523/JNEUROSCI.20-17-06594.2000>

Appendix A: Model Prompts

Structured Captioning Prompt

Every image was converted into a structured caption with the schema below before semantic feature recovery.

You are given one natural image.

Return ONLY one JSON object.
Do not write any text outside the JSON.
Do not use markdown.
Keep every field short and concrete.
If a field is unsupported by the image, use "".

Goal:
produce a compact, semantically useful visual description for later high-level feature aggregation.
Prioritize clearly visible scene, object, body, action, appearance, and text-related information.
Do NOT report exact object coordinates.
Do NOT report foreground/midground/background depth layering.
Do NOT focus on low-level edge geometry.

JSON schema:

```
{
  "caption": "2-3 short sentences. Summarize the main scene, salient objects, people/animals if present, what is happening, the setting/context, and any visible text or logo.",
  "primary_focus": "the single dominant object, entity, or coherent object group that the image is mainly about; if no single focus dominates, use ''",
  "objects": "comma-separated salient items, including food/tableware/text-bearing objects when present; include short qualifiers when useful",
  "major_objects_count": "none | one | two | three_or_more",
  "environment_type": "indoor | outdoor | none",
  "framing_type": "close-up | object-centered | person-centered | scene-centered | none",
  "scene_category": "short semantic scene label such as lake, street, kitchen, park, forest path, office, storefront, beach, living room, classroom, roadway, dining table, transit platform; otherwise ''",
  "dominant_colors": "comma-separated color words for the visually dominant colors or color themes",
  "activity": "short phrase describing the visible action, event, or interaction; if mostly static, describe the visible state",
  "background": "short phrase describing the surrounding context behind or around the main focus; do not repeat scene category unless needed for clarity",
  "body_postures": "short phrase about visible body/face configuration if present: pose, orientation, gesture, facial visibility, interaction with objects; otherwise ''",
  "movement_tendency": "short phrase about visually implied motion, action readiness, directional posture, or dynamic tension if clearly supported; do not describe depth layering; if mostly static, say so briefly",
  "appearance_form_cues": "short phrase about clearly visible surface/material appearance, curved-vs-straight form tendency, garment/object appearance detail, or other semantically useful appearance cues; avoid raw low-level geometry terms",
  "text_and_symbol_presence": "short phrase describing visible text, logo, signage, label, package text, or screen text, including if it is partial, small, or unreadable; otherwise ''"
}
```

Rules:

- * Use only visually supported information.
- * Report subtle but potentially meaningful cues if they are visible, even if they are not dominant.
- * `primary_focus` should reflect the main semantic focus of the image, not the whole scene inventory.
- * For `major_objects_count`, count only focal objects/entities at the same semantic level as `primary_focus`. Do not count background scene elements, sky, water, trees, buildings, or clutter unless they are themselves the main focus.
- * If one object/entity clearly dominates attention, use "one".
- * If there are two co-dominant focal entities, use "two".
- * Use "three_or_more" only when several focal entities are equally central.
- * `environment_type` and `framing_type` are different: `environment_type` is indoor/outdoor; `framing_type` is about what the image mainly centers on.
- * `scene_category` should name the kind of place or setting as specifically as the image supports.
- * For `movement_tendency`, do not hallucinate actual motion; report only visible movement implication or say the scene is mostly static.
- * For `appearance_form_cues`, focus on semantically useful appearance or form tendencies, not LVF-style edge/shape decomposition.
- * Keep outputs concise, specific, and reusable.

Canonicalization Verification Prompt

For each image, candidate features shortlisted from the canonical dictionary were verified against the image's caption with the schema below.

You are verifying HVF atoms from FULL_CAPTION and FIELD_TEXTS.

Task:

Return exactly one 0/1 label for every residual candidate, in the exact order provided.

Decision rule:

1 = the atom is supported by FULL_CAPTION or FIELD_TEXTS.

0 = the atom is not supported.

Recall-sensitive rule for concrete and human-related evidence:

- Mark 1 for explicit mentions.
- Mark 1 for close synonyms, paraphrases, singular/plural variants, and ordinary wording differences.
- For human/person: mark 1 when the caption or field text states people, person, human, man, woman, child, rider, skier, player, or similar human role.
- For human-contained visible features: mark 1 for face, hand, arm, leg, body part, clothing, posture, standing, sitting, smiling, facing, or grasp/hold-related atoms when the caption or field texts describe visible people and the atom is field-appropriate.
- Field-appropriate means: do not mark a human/background item as primary_focus unless the caption or field text makes humans the main focus.

Harsher rule for abstract terms:

- Abstract atoms are 1 only when directly supported by explicit caption/field evidence or by a necessary visible entailment.
- Do not mark broad abstract categories just because they are generally possible.
- Tool/manipulable object, vehicle, scene, posture, social/action, texture, and form abstractions require direct support from the caption or field texts.
- For scene abstractions, require explicit scene evidence, not neighboring scene concepts.

Movement-tendency rule:

- Be more recall-sensitive for movement tendency than for other abstractions.
- Mark 1 when movement is explicit or directly implied by action wording, posture-in-action, or described dynamic context.
- Accept direct implications such as walking, gesturing, posing dynamically, mimicking an action, balancing on a movable object, riding/standing on skateboard/surfboard/ski-like equipment, implied pedestrian movement, or visible action readiness.
- Still mark 0 for unrelated motion types, merely possible motion, or motion not supported by the caption/field texts.

Do not hallucinate:

- Mark 0 for sibling concepts, unrelated subtypes, or scene-neighbor concepts not entailed.
- Mark 0 for abstract categories when no stated instance supports them.
- Mark 0 when the evidence is genuinely uncertain.

Output only JSON:

```
{"image_id": "...", "chunk_index": K, "n_label_slots": N, "labels": [0, 1, ...]}
```

Semantic Explainer Prompt: Enriched-Only Units

For units with enriched features only, the explainer was prompted with the single-list schema below. {ATOMS} is replaced by the unit's recovered enriched features.

You are given a montage of 20 images, ordered left-to-right and top-to-bottom. These are the 20 images that most strongly activate one learned unit from a model of visual-cortex responses.

Analysis of these 20 images found that the following visual concepts each recur across them.

IMPORTANT: these concepts were each found to be independently over-represented -- they do NOT necessarily co-occur in the same image. The unit responds to images exhibiting ANY of them, not images where all appear together. Treat this list as the unit's conceptual scope:

{ATOMS}

Your task: in ONE natural-language sentence, describe the unit's overall selectivity -- the common theme that spans these concepts -- as a fluent description, as if telling a person what kind of images this unit tends to respond to.

Rules:

- Describe the shared theme across the concepts, NOT a single scene requiring all of them at once.
- Prefer a unifying description (e.g. "people and their body parts in active poses") over a conjunction that demands every concept be present simultaneously.
- Stay within the conceptual scope above: do not invoke any object, scene, action, or category outside these concepts, even if visible in the images.
- Do NOT copy the concept terms verbatim or list them as tags. Paraphrase into natural language.
- Write a sentence, not a label or comma-separated set.
- Do NOT single out specific images ("especially image 3", "such as the one with X").
- Do NOT speculate about why the unit responds or what brain region it is.
- If the concepts are broad, the sentence stays broad; do not invent specificity they lack.
- One fluent sentence.

Return ONLY one JSON object:

```
{"explanation":"one natural sentence describing the unit's shared selectivity"}
```

Semantic Explainer Prompt: Units Carrying Depleted Features

For units with at least one depleted feature, the explainer was prompted with the two-list schema below. {ENRICHED} and {DEPLETED} are replaced by the unit's enriched and depleted features respectively. When a unit has no enriched features, {ENRICHED} is set to “(none, this unit is defined only by what it avoids)”.

You are given a montage of 20 images, ordered left-to-right and top-to-bottom. These are the 20 images that most strongly activate one learned unit from a model of visual-cortex responses.

Analysis of these 20 images found two kinds of visual concepts:

PRESENT -- concepts that recur across these images (the unit responds TO these). These were each independently over-represented; they do not necessarily co-occur in the same image: {ENRICHED}

CONSPICUOUSLY ABSENT -- concepts that are unusually MISSING from these images compared to images in general (the unit specifically does NOT respond to these; their absence is part of what defines the unit): {DEPLETED}

Your task: in ONE natural-language sentence, describe the unit's selectivity, capturing BOTH what kind of images it responds to AND what it conspicuously avoids.

Rules:

- The CONSPICUOUSLY ABSENT concepts are NOT present in the images. Describe them as things the unit avoids or that are absent -- never describe them as if they appear (e.g. if "static" is absent, say "...and tend not to be static", NOT "...static scenes").
- If the PRESENT list is empty, describe the unit purely by what it avoids: the unit responds to images that are NOT characterized by the absent concepts.
- Stay within the listed concepts; do not invoke objects, scenes, actions, or categories outside them, even if visible.
- Do NOT copy concept terms verbatim or list them as tags. Paraphrase into natural language; write a sentence, not a label.
- Do NOT single out specific images. Describe the recurring pattern across the set.
- For PRESENT concepts: describe the shared theme, not a scene requiring all of them at once.
- One fluent sentence.

Return ONLY one JSON object:

```
{"explanation":"one natural sentence covering what the unit responds to and what it avoids"}
```

Semantic Evaluator Prompt

The evaluator was prompted with the schema below. {NLE} is replaced by the explanation generated at the first stage. The evaluator was not shown the unit's recovered features. The apparent interpretability score is the proportion of positive labels among the 20 images.

You are given a montage of 20 images, ordered left-to-right and top-to-bottom.

You are also given ONE explanation describing the kind of images a learned unit tends to respond to.

Explanation:
{NLE}

Your task: for each of the 20 images independently, decide whether it plausibly belongs to the kind of images the explanation describes -- whether it fits the unit's overall theme.

Rules:

- Return 1 if the image fits the general theme of the explanation, even if not every detail in the explanation is present in that image. A partial but clear thematic match counts as 1.
- Return 0 if the image is absent of the theme, unclear, or only weakly or tangentially related.
- Do NOT require that every element mentioned in the explanation appear together in the image.
- Judge each image on its own; do not let the overall set bias an individual image.

Return exactly 20 binary digits and nothing else (e.g. 10110...), one digit per image in montage order.

Primary Explainer Prompt

The explainer was prompted per channel. {signal_modality} names the channel, {modality_block} is the channel-specific description drawn from the modality prompt bank, and {unit_metadata(row)} is optional per-unit metadata.

You are given two aligned montages of 20 images, ordered left-to-right and top-to-bottom.

Montage A shows the raw top-activating images for one learned fMRI/SAE unit.

Montage B shows the corresponding low-level visual signal expression maps for the same 20 images, using the {signal_modality} modality.

{modality_block}

The colored overlay is not a natural image color signal. It visualizes where the unit-specific {signal_modality} signal is expressed in each image. Red/bright structures indicate stronger localized expression of the recovered unit-specific signal; darker regions indicate weaker expression.

Your task is to infer ONE best explanation sentence describing the visual signal pattern this unit appears to represent.

Rules:

- Use Montage A only to understand visible image structure.
- Use Montage B to identify the recurring recovered signal.
- The explanation must describe the signal pattern itself, not the object category, scene category, material, color name, action, or named image content.
- If raw-image contents are visible, do not name them; describe only the shared visual structure expressed by the signal.

- The explanation may include spatial layout when it helps describe where the signal is expressed.
- Do not describe only the generic presence of this signal modality; identify the recurring signal structure that distinguishes this unit from other possible units in the same modality.
- Do not give examples, case lists, object lists, or X/Y/Z-style expansions.
- Do not describe the overlay color itself as the explanation.
- Do not force the explanation to fit every image.
- Keep the explanation compact and interpretable as a low-level visual feature representation.

Optional unit metadata:
 {unit_metadata(row)}

Return ONLY one JSON object:
 {"explanation":"one sentence, concise but meaningful"}

Primary Evaluator Prompt

The evaluator was prompted with the same channel-specific substitutions.

{explanation} is replaced by the explanation generated at the first stage.

You are given two aligned montages of 20 images, ordered left-to-right and top-to-bottom.

Montage A shows the raw images.

Montage B shows the corresponding {signal_modality} signal expression maps for those same 20 images.

{modality_block}

The colored overlay is not a natural image color signal. It visualizes where the unit-specific {signal_modality} signal is expressed in each image. Red/bright structures indicate stronger localized expression of the recovered unit-specific signal; darker regions indicate weaker expression.

You are also given one explanation describing the visual signal pattern a learned unit may represent.

Explanation:
 "{explanation}"

Your task is to evaluate each image independently and decide whether the explanation is supported in that image.

- Rules:
- Use Montage A to judge visible low-level structure.
 - Use Montage B to judge whether the same signal pattern is expressed in that image.
 - Judge the signal pattern named in the explanation, including its spatial layout if stated.
 - Do not mark an image positive merely because it shows any activity typical of this signal modality; mark 1 only when the map supports the specific recurring signal structure stated in the explanation.
 - Do not treat object category, scene category, material, color name, action, or named image content as the target.
 - Return 1 if the explanation is clearly or reasonably supported by the visible structure and the signal map.
 - Return 0 if the explanation is absent, unclear, only weakly related, spatially mismatched, or not supported by the signal map.
 - Do not force positives.
 - Keep the output aligned to montage order: image 1 to image 20.
 - Return exactly 20 binary digits and nothing else.

Appendix B: Demonstration of Discovered Semantic Visual Representations

Two units are shown for each of the 12 of the 13 category-selective regions of interest, excluding mfs-words, all drawn from the Natural Scenes Dataset. Each figure shows the 20 images that most strongly activate the unit, ordered by activation strength from left to right and top to bottom. The footer of each montage lists the semantic features recovered from the unit, separated into enriched and depleted features where both are present, the natural-language explanation generated from them, and the apparent interpretability score. Regions are ordered by category: body-selective, face-selective, place-selective, and word-selective.

EBA: Extrastriate Body Area

Figure B1

Top-20 Activating Images of an EBA Unit (participant 05, replication 1, u30)

Top-20 activating images — GT/subj05/rep1/EBA u30

raw



enriched semantic features: {body part / hand, face, human/person, object-directed manipulation, playing, social action, reaching/grasp-related posture, upright posture, person-centered}

depleted semantic features: {one}

"The unit prefers images of people, especially men, engaged in social, hands-on action with visible faces and arms in upright or reaching postures, and it tends to avoid angular, rectilinear forms and solitary one-person scenes." (AIS = 0.700)

Figure B2

Top-20 Activating Images of an EBA Unit (participant 05, replication 2, u504)

Top-20 activating images — GT/subj05/rep2/EBA u504

raw



enriched semantic features: {body part / hand, face, human/person, tool/manipulable object, court, player, tennis racket, tennis court, object-directed manipulation, tennis, reaching/grasp-related posture, biological motion, visible non-static motion, person-centered}

depleted semantic features: {static, static/stationary}

"The unit favors images of people in active tennis or board-sport movements, especially reaching or handling equipment in outdoor court or path-like settings, and it tends to avoid scenes that are static or still." (AIS = 0.800)

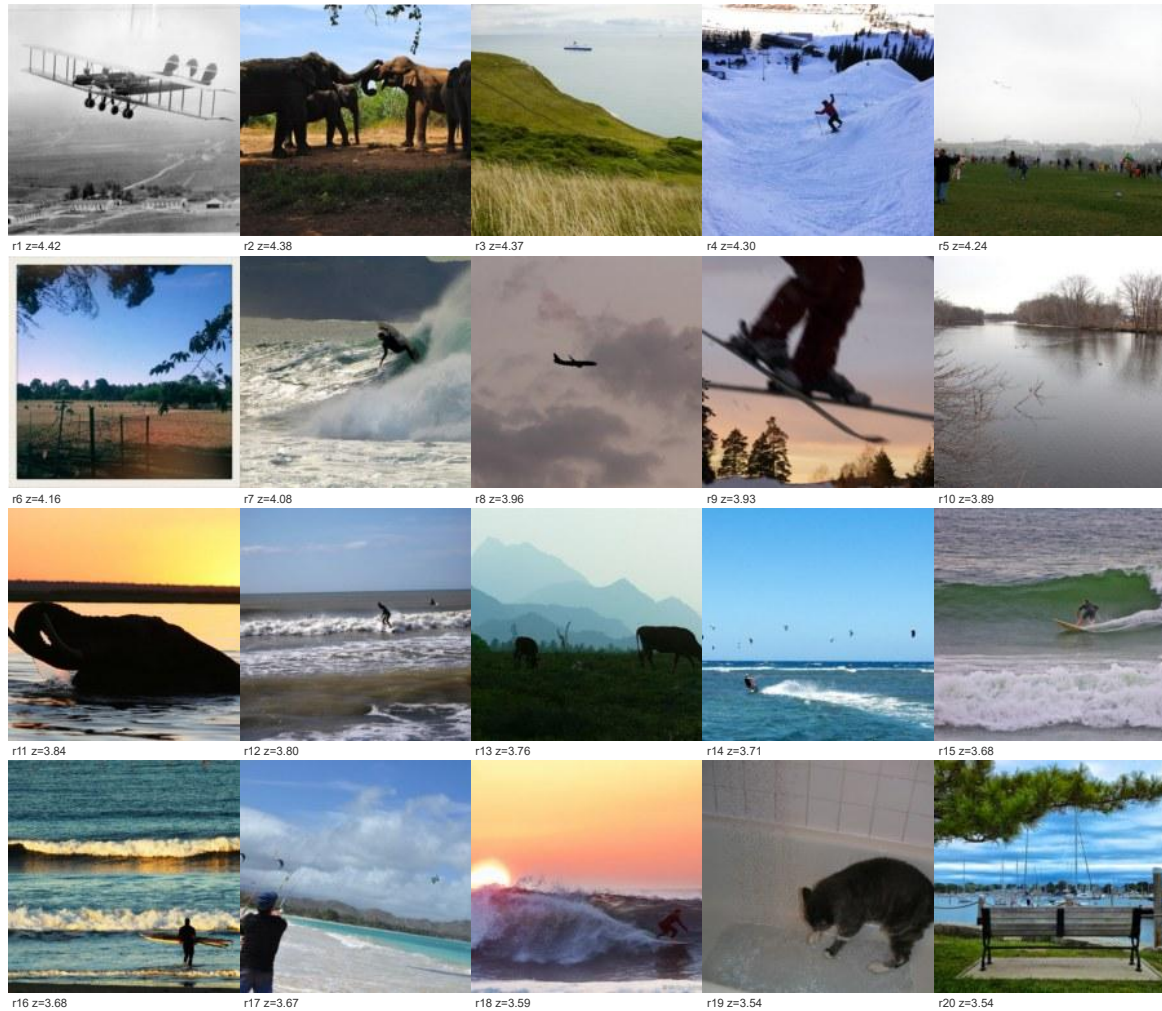
FBA-1: Fusiform Body Area 1

Figure B3

Top-20 Activating Images of an FBA-1 Unit (participant 05, replication 1, u91)

Top-20 activating images — GT/subj05/rep1/FBA-1 u91

raw



enriched semantic features: {sky, locomotion, visible non-static motion}

depleted semantic features: {}

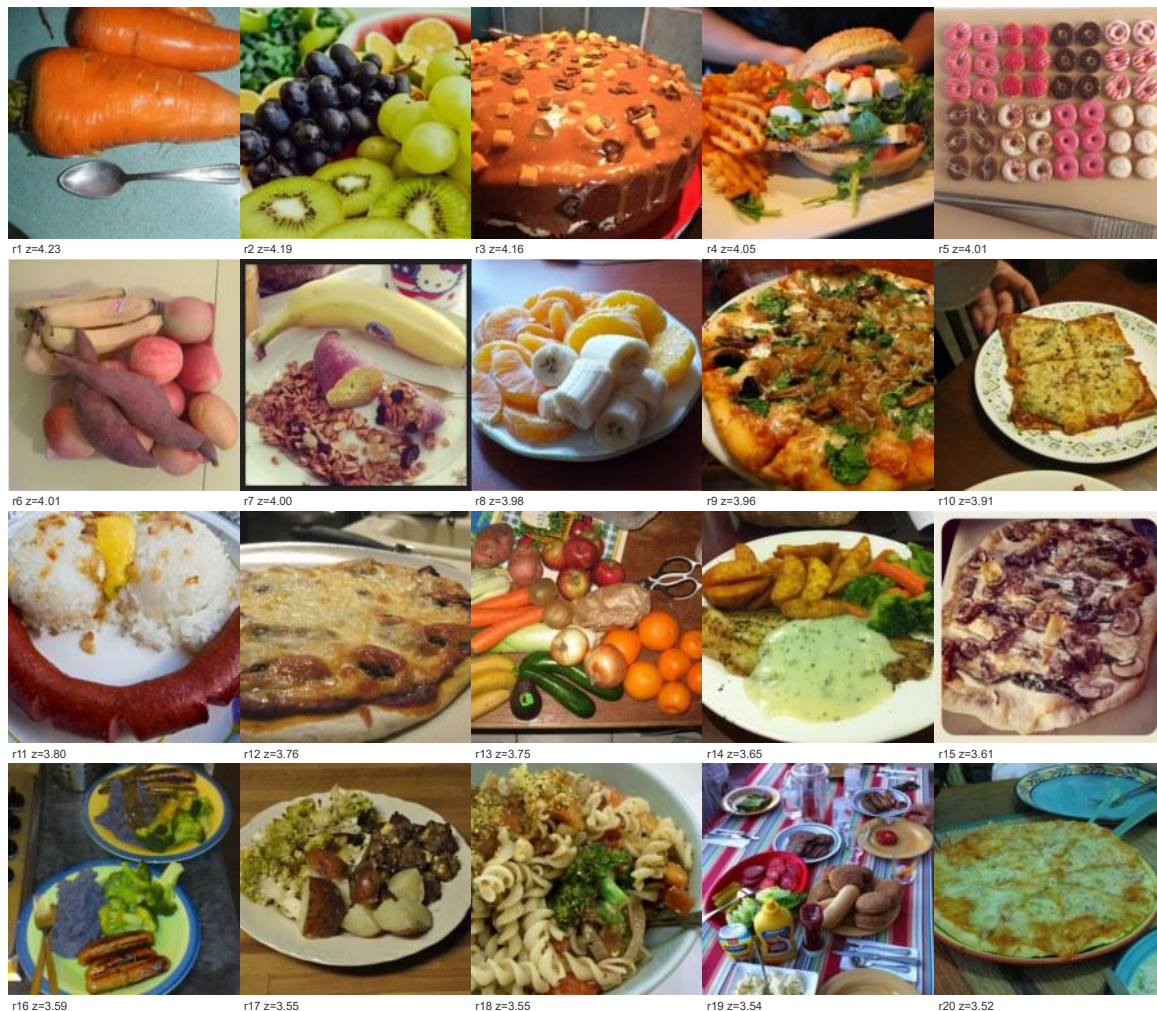
"Images of things moving through open natural settings, often under open sky or on water, including active outdoor sports and other motion-filled scenes." (AIS = 0.800)

Figure B4

Top-20 Activating Images of an FBA-1 Unit (participant 05, replication 2, u59)

Top-20 activating images — GT/subj05/rep2/FBA-1 u59

raw



enriched semantic features: {food, plate, indoor scene, object-centered}

depleted semantic features: {standing, upright posture}

"This unit responds to images of food presented in indoor, object-centered settings, often on plates or against textured surfaces, and it tends not to respond when black tones or upright standing forms are prominent." (AIS = 0.850)

FBA-2: Fusiform Body Area 2

Figure B5

Top-20 Activating Images of an FBA-2 Unit (participant 05, replication 1, u864)

Top-20 activating images — GT/subj05/rep1/FBA-2 u864

raw



enriched semantic features: {body part / hand, face, human/person, man, object-directed manipulation, social action, reaching/grasp-related posture, upright posture, person-centered}

depleted semantic features: {}

"Images of people in active, upright poses, often using their hands to interact with objects or with others, including gestures, reaching, and other visible motion." (AIS = 0.950)

Figure B6

Top-20 Activating Images of an FBA-2 Unit (participant 07, replication 1, u18)

Top-20 activating images — GT/subj07/rep1/FBA-2 u18

raw



enriched semantic features: {man, facing forward}

depleted semantic features: {}

"Images showing a single person or face, typically a person centered and looking toward the camera." (AIS = 0.650)

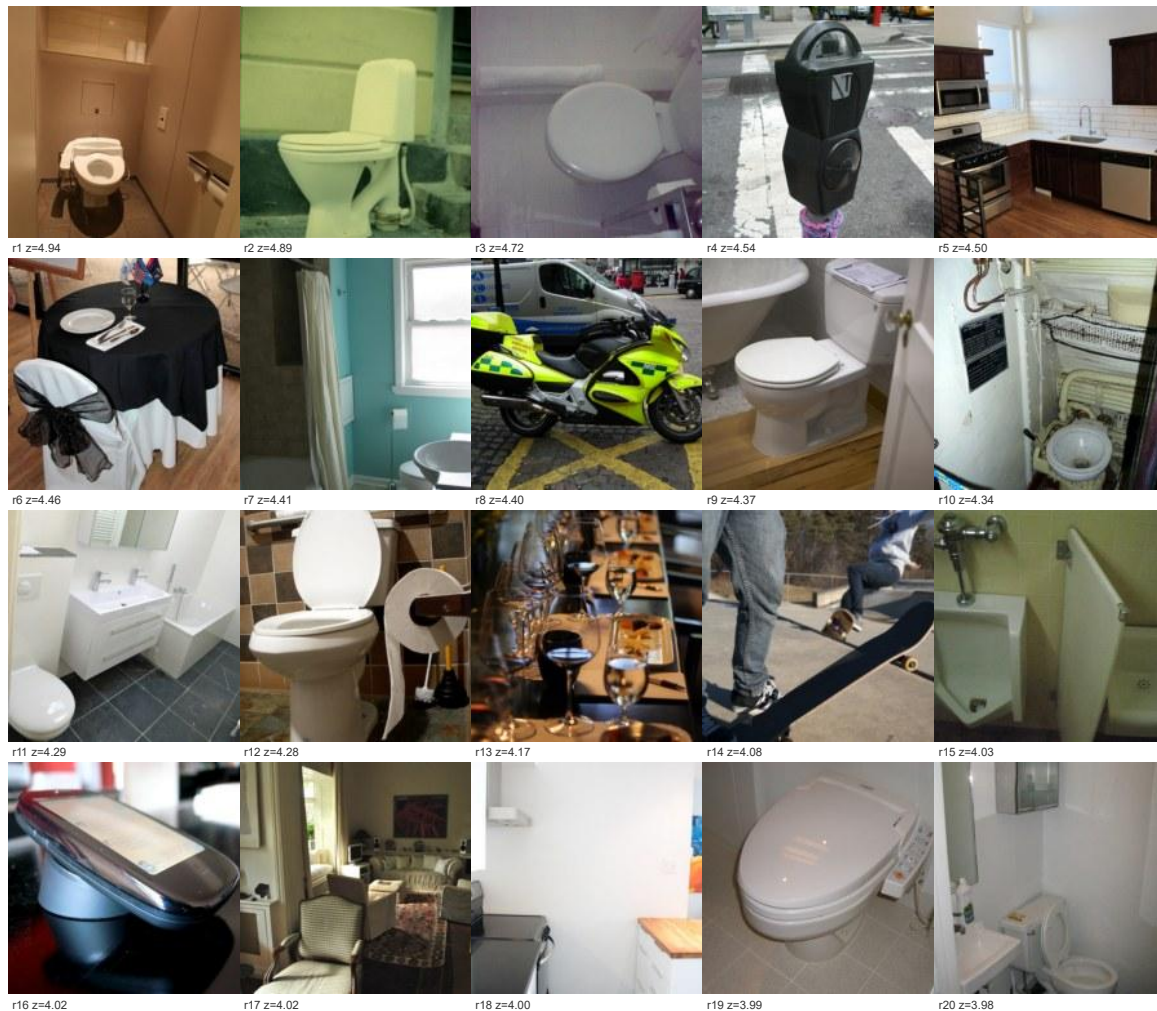
OFA: Occipital Face Area

Figure B7

Top-20 Activating Images of an OFA Unit (participant 05, replication 1, u13)

Top-20 activating images — GT/subj05/rep1/OFA u13

raw



enriched semantic features: {toilet, indoor scene}

depleted semantic features: {}

"An indoor bathroom fixture, especially a toilet, in a room setting." (AIS = 0.500)

Figure B8

Top-20 Activating Images of an OFA Unit (participant 05, replication 2, u78)

Top-20 activating images — GT/subj05/rep2/OFA u78

raw



enriched semantic features: {face, human/person, looking, facing forward, person-centered}

depleted semantic features: {}

"Images of people—often shown from the front and in upright poses—especially when their face, hands, or arms are prominent in gestures of holding, reaching, or handling something." (AIS = 1.000)

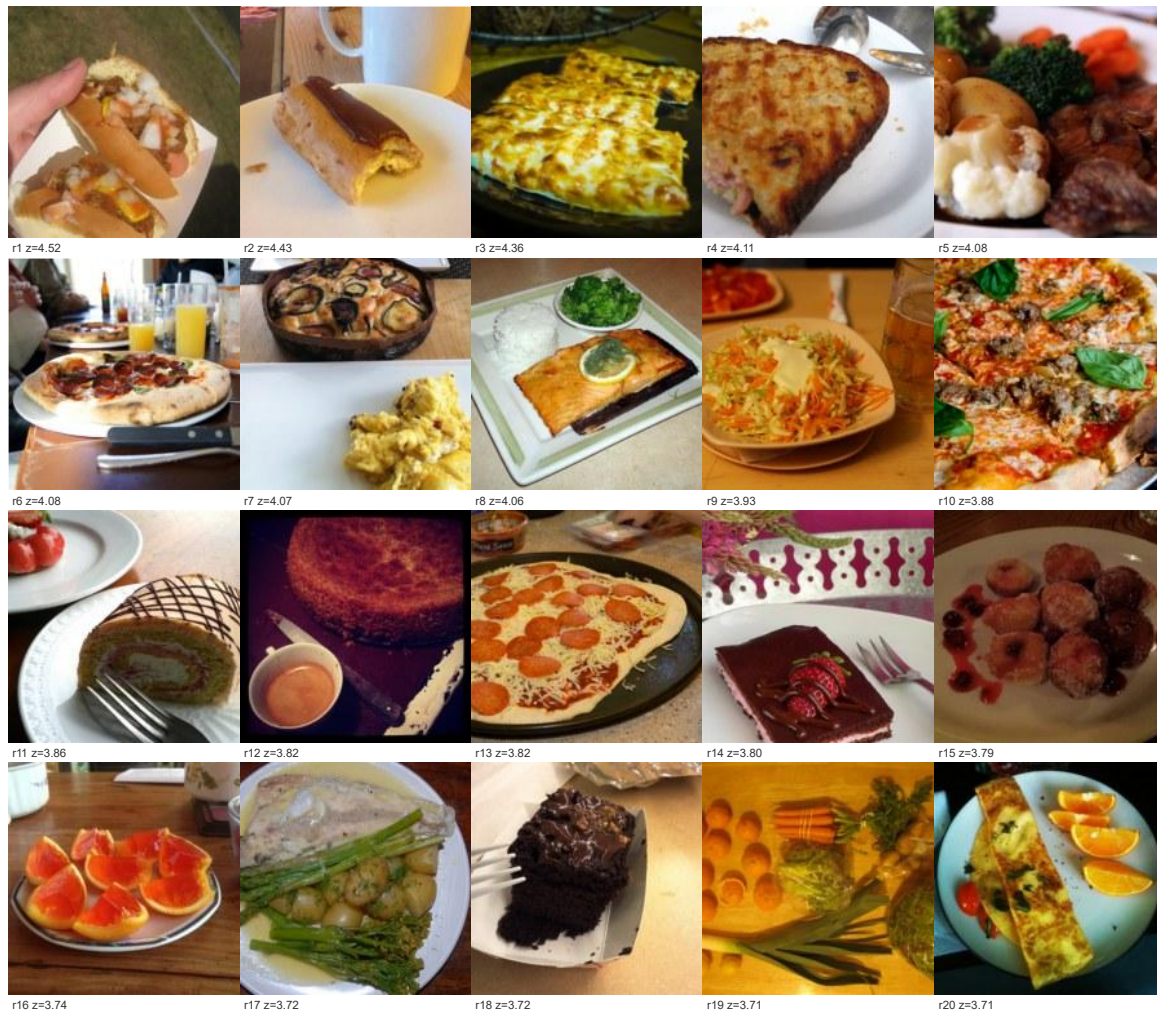
FFA-1: Fusiform Face Area 1

Figure B9

Top-20 Activating Images of an FFA-1 Unit (participant 05, replication 1, u64)

Top-20 activating images — GT/subj05/rep1/FFA-1 u64

raw



enriched semantic features: {food, plate, object-centered}

depleted semantic features: {standing}

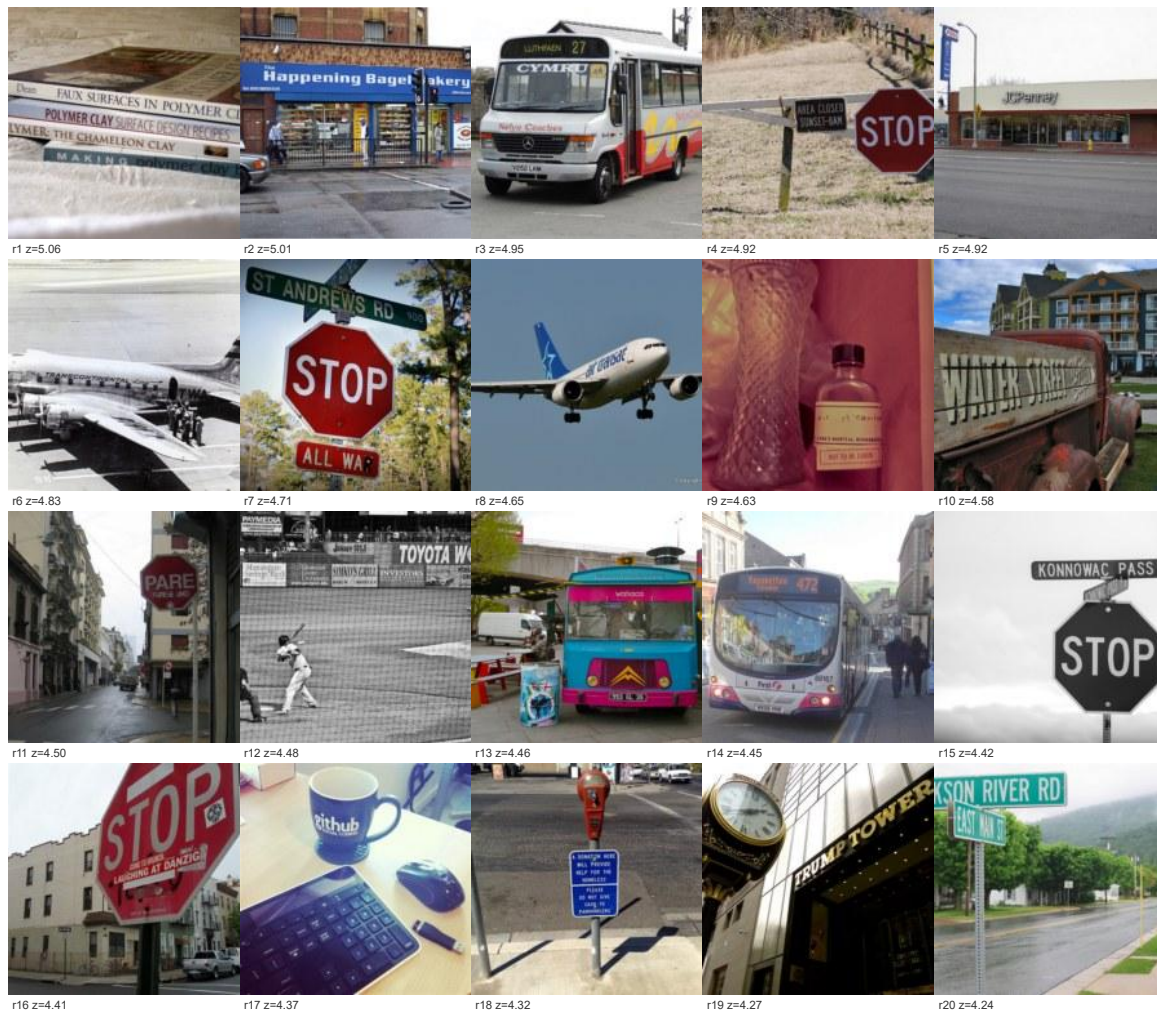
"This unit prefers close-up views of plated food with lots of surface texture and object-centered compositions, and it tends to avoid blue tones, faces, standing people, hand-held tools, and upright postures." (AIS = 0.850)

Figure B10

Top-20 Activating Images of an FFA-1 Unit (participant 05, replication 2, u81)

Top-20 activating images — GT/subj05/rep2/FFA-1 u81

raw



enriched semantic features: {street, alphabetic/word-like text, sign/label/logo mark, text present}

depleted semantic features: {}

"Images containing visible written signage, labels, or other text, especially in street and roadside scenes." (AIS = 0.950)

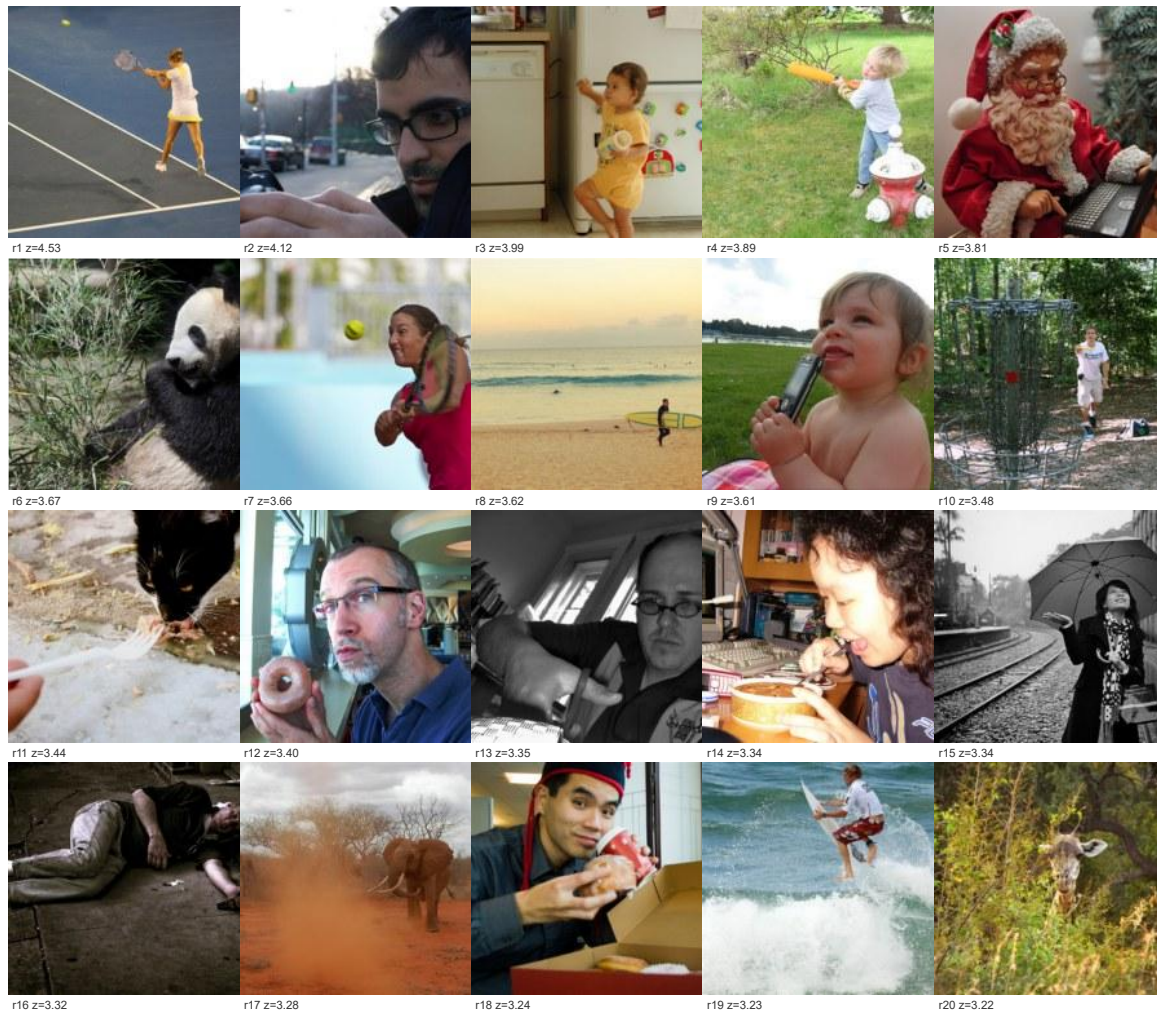
FFA-2: Fusiform Face Area 2

Figure B11

Top-20 Activating Images of an FFA-2 Unit (participant 02, replication 2, u72)

Top-20 activating images — GT/subj02/rep2/FFA-2 u72

raw



enriched semantic features: {body part / hand, holding, object-directed manipulation, reaching/grasp-related posture}

depleted semantic features: {}

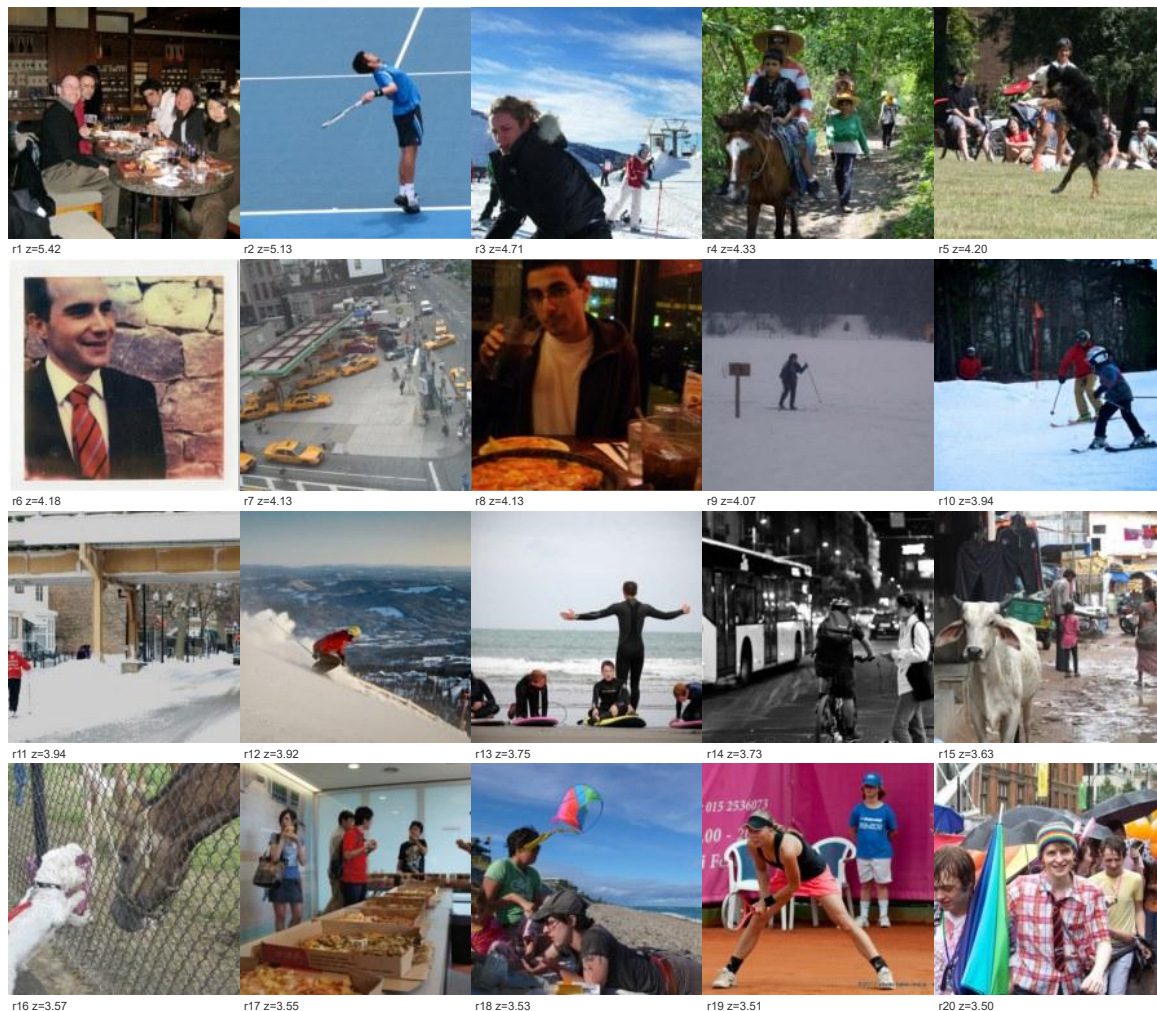
"Images of people or animals with arms, paws, or hands engaged in holding, reaching for, or otherwise manipulating something."
(AIS = 0.750)

Figure B12

Top-20 Activating Images of an FFA-2 Unit (participant 03, replication 2, u996)

Top-20 activating images — GT/subj03/rep2/FFA-2 u996

raw



enriched semantic features: {face, human/person, people, social action}

depleted semantic features: {}

"Images of people, especially upright in active movement or social interaction, including visible faces and body parts." (AIS = 0.850)

OPA: Occipital Place Area

Figure B13

Top-20 Activating Images of a OPA Unit (participant 05, replication 1, u26)

Top-20 activating images — GT/subj05/rep1/OPA u26

raw



enriched semantic features: {food, plate, brown, textured surface, object-centered}

depleted semantic features: {standing}

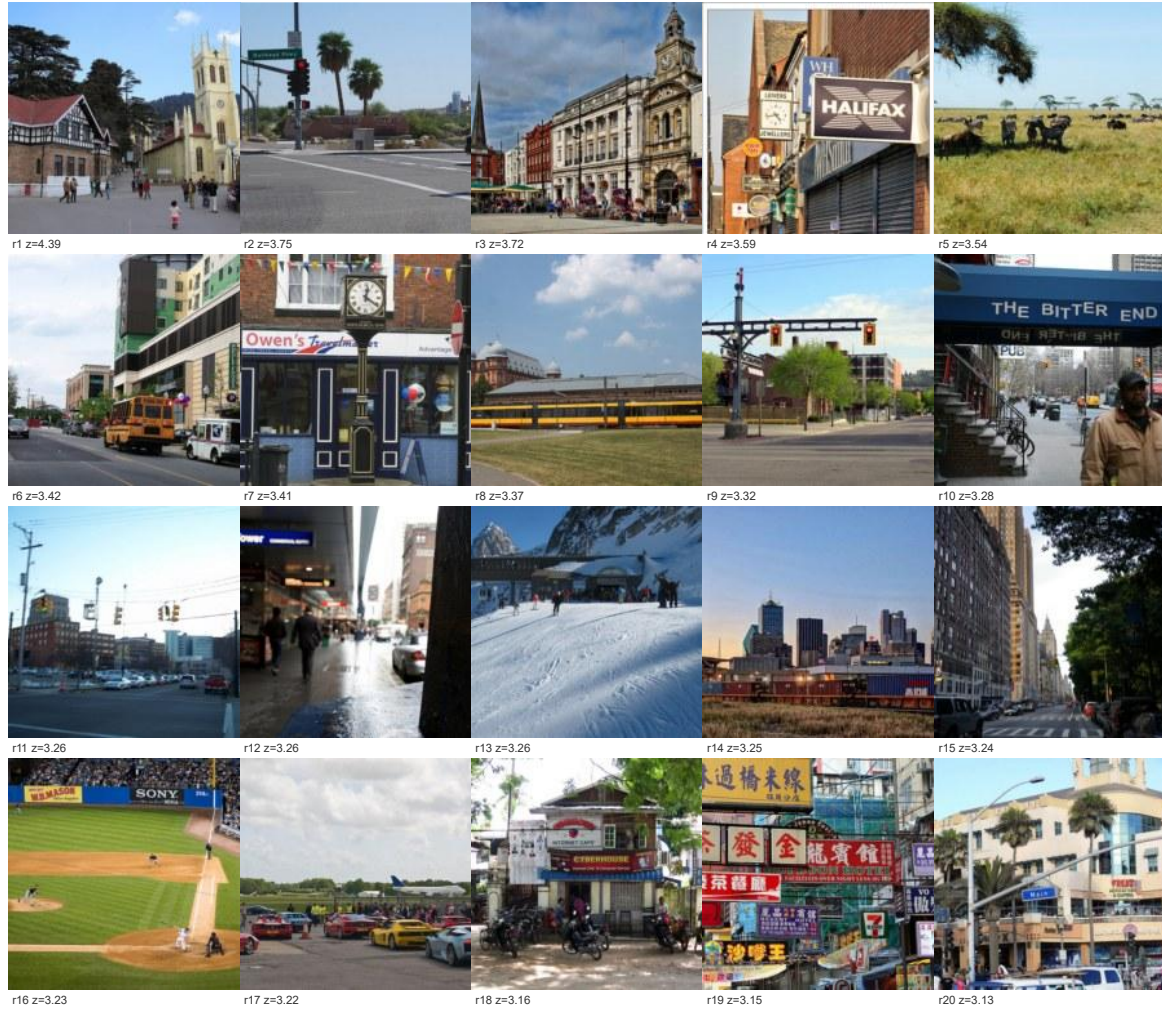
"This unit responds to close-up, object-centered views of browned, textured food on plates—especially pizza-like dishes and other prepared meals—and it tends not to activate for images with people standing." (AIS = 1.000)

Figure B14

Top-20 Activating Images of a OPA Unit (participant 05, replication 2, u6)

Top-20 activating images — GT/subj05/rep2/OPA u6

raw



enriched semantic features: {street, outdoor built scene, scene-centered}

depleted semantic features: {}

"Images of public outdoor urban streetscapes and other built environments viewed as broader scene settings." (AIS = 0.750)

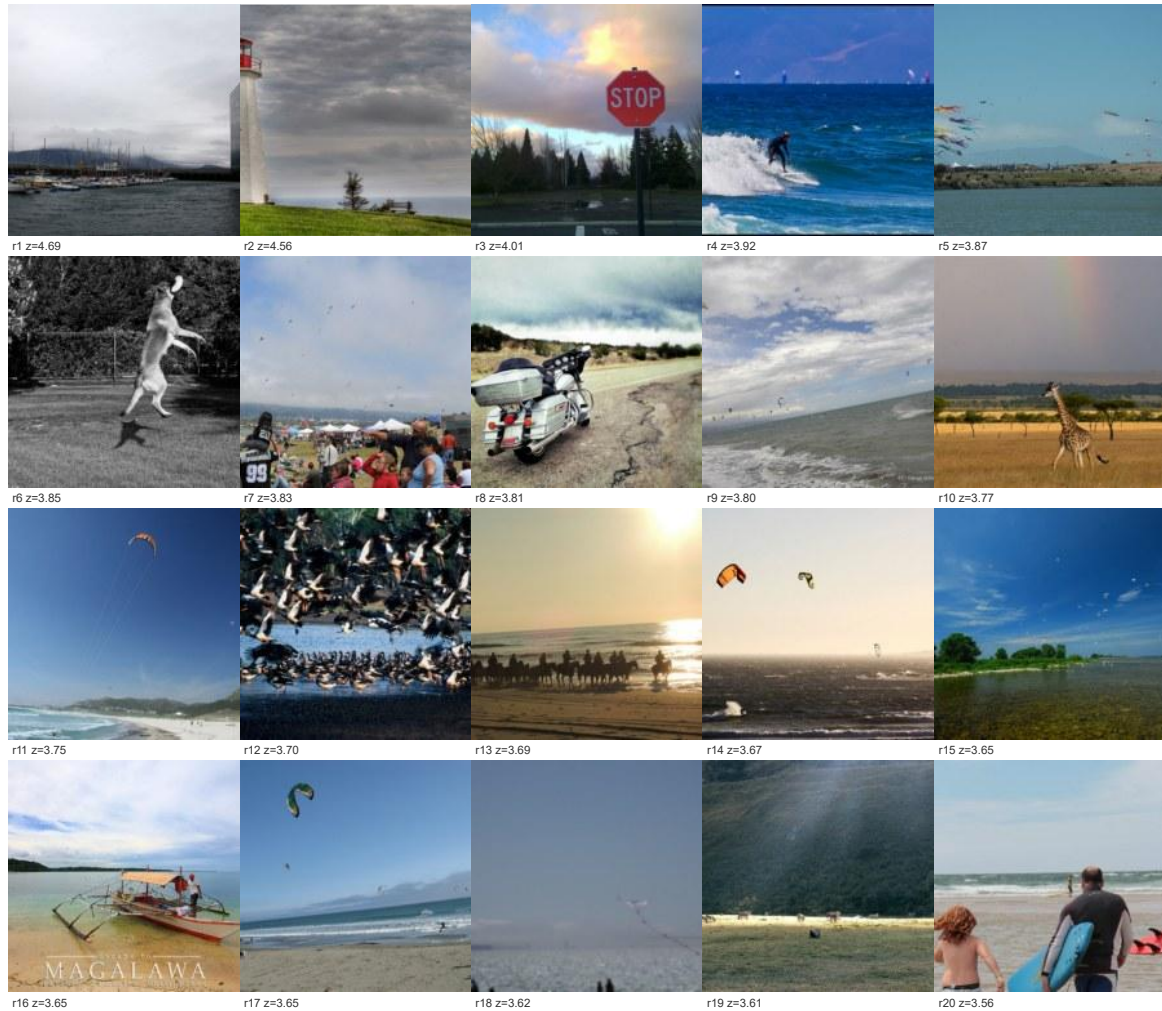
PPA: Parahippocampal Place Area

Figure B15

Top-20 Activating Images of a PPA Unit (participant 02, replication 1, u29)

Top-20 activating images — GT/subj02/rep1/PPA u29

raw



enriched semantic features: {sky, outdoor natural scene}

depleted semantic features: {}

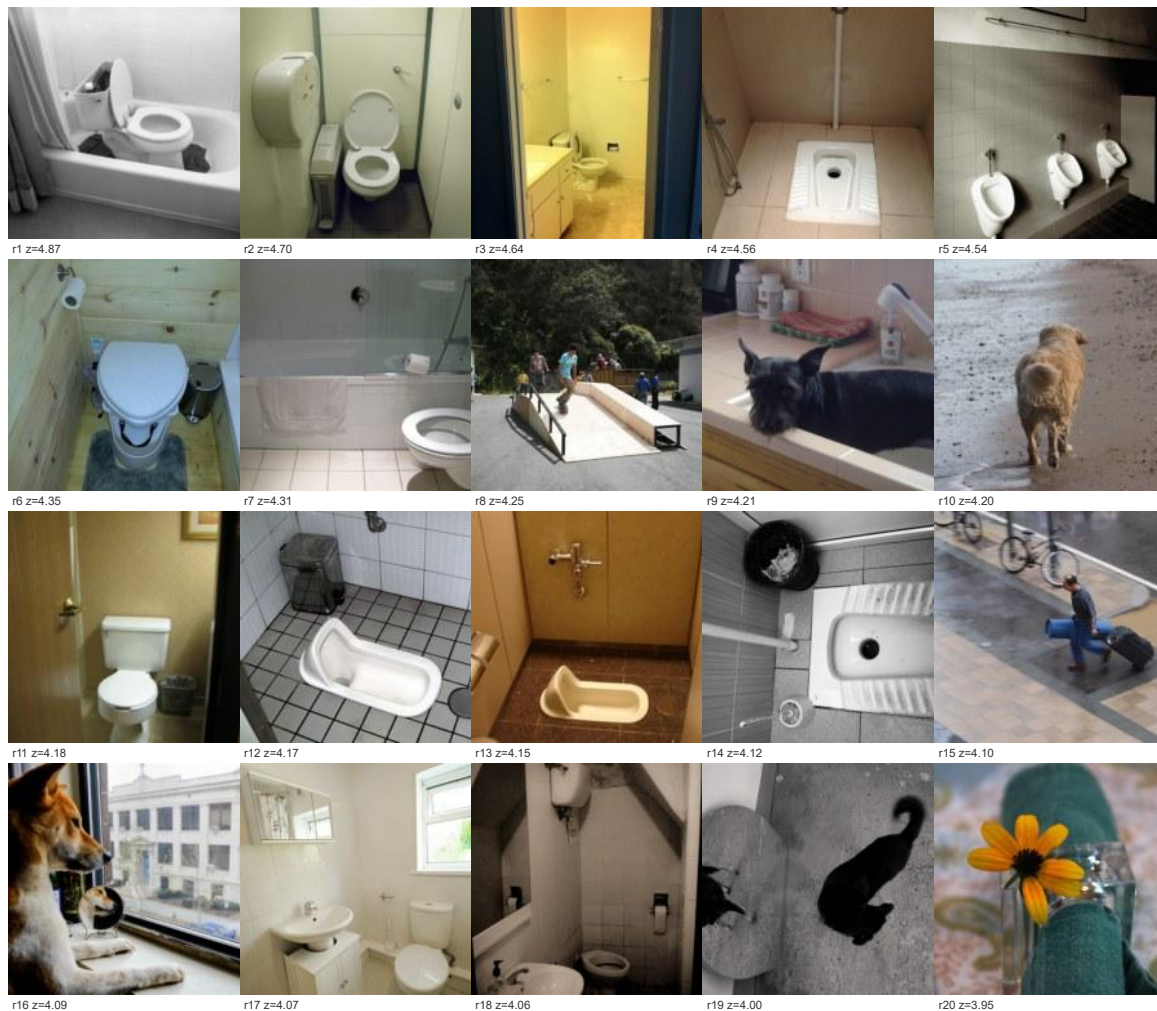
"Images of open outdoor scenery, especially broad natural landscapes with visible sky." (AIS = 0.950)

Figure B16

Top-20 Activating Images of a PPA Unit (participant 05, replication 2, u70)

Top-20 activating images — GT/subj05/rep2/PPA u70

raw



enriched semantic features: {toilet, indoor scene, smooth surface}

depleted semantic features: {}

"An indoor bathroom-like setting with ceramic fixtures and other smooth hard surfaces" (AIS = 0.650)

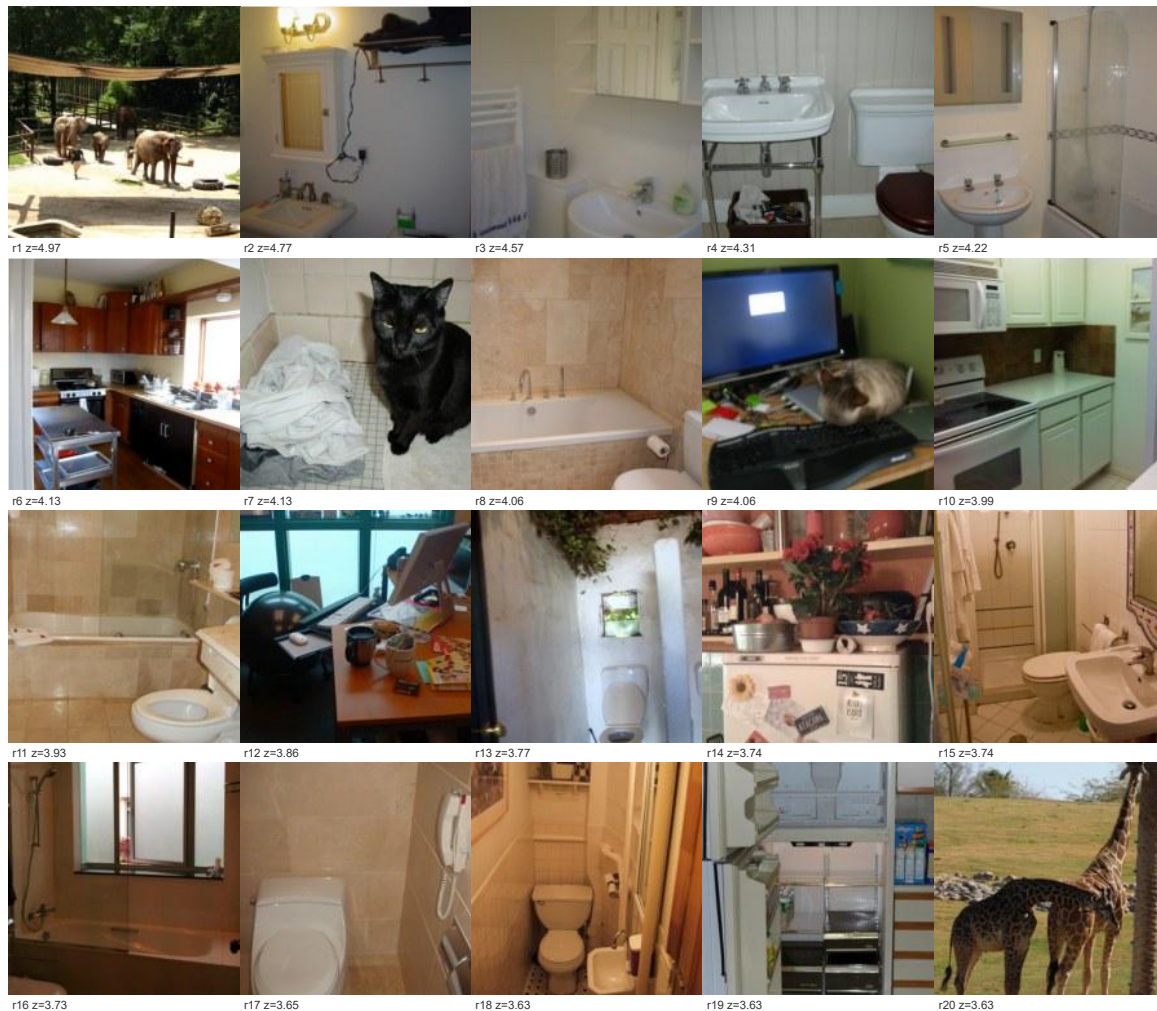
RSC: Retrosplenial Cortex

Figure B17

Top-20 Activating Images of an RSC Unit (participant 02, replication 2, u33)

Top-20 activating images — GT/subj02/rep2/RSC u33

raw



enriched semantic features: {bathroom, indoor scene, beige, smooth surface}

depleted semantic features: {}

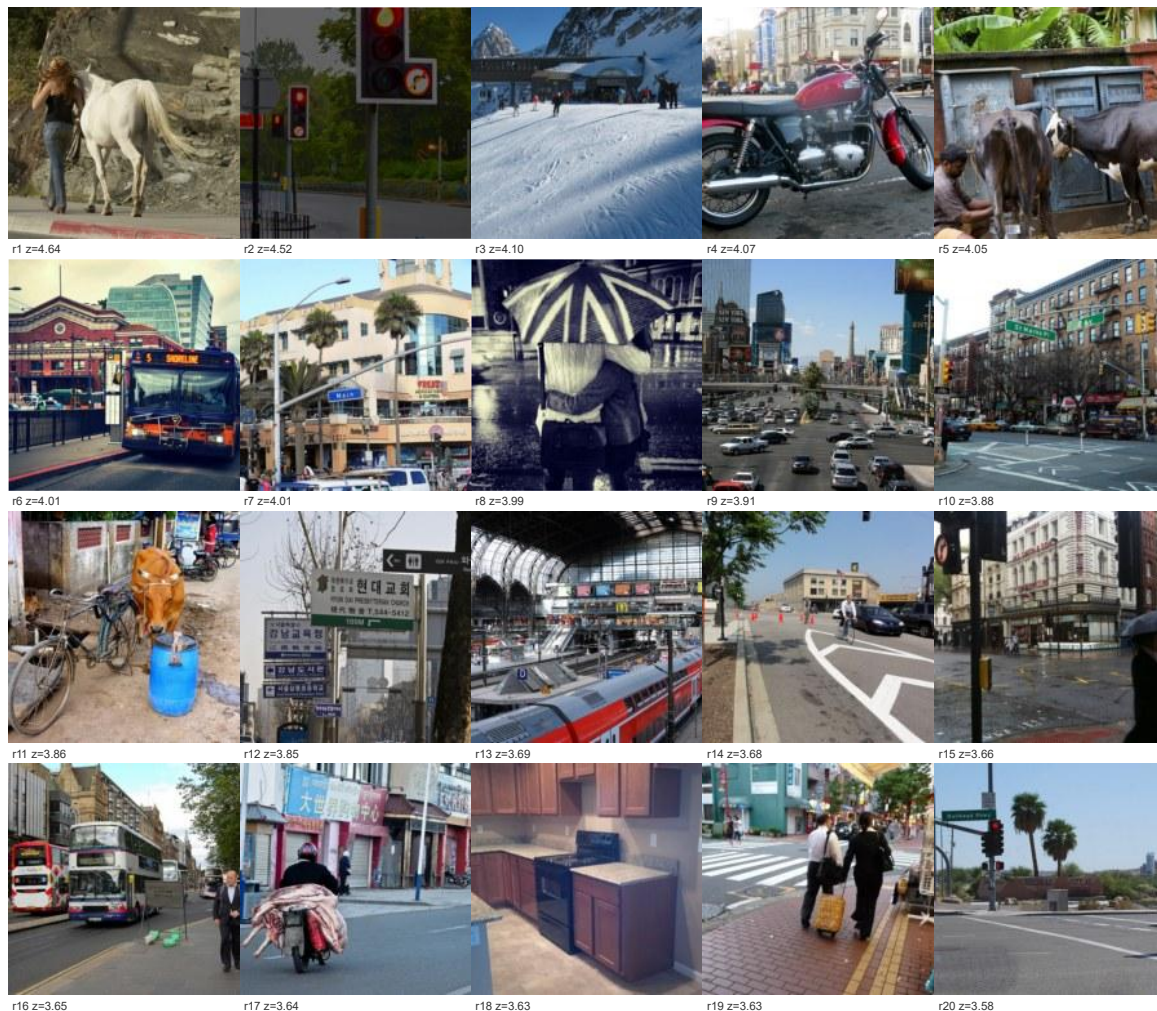
"The unit responds to mostly indoor, bathroom-like, pale and smooth-surfaced scenes, often centered on the room itself, and it tends to avoid images containing people." (AIS = 0.700)

Figure B18

Top-20 Activating Images of an RSC Unit (participant 05, replication 2, u544)

Top-20 activating images — GT/subj05/rep2/RSC u544

raw



enriched semantic features: {street, navigable/path-like scene, outdoor built scene}

depleted semantic features: {}

"Images of streets and other built outdoor places, especially scenes with roads, crossings, traffic, and similar walkable or drivable environments." (AIS = 0.800)

OWFA: Occipital Word Form Area

Figure B19

Top-20 Activating Images of an OWFA Unit (participant 05, replication 1, u246)

Top-20 activating images — GT/subj05/rep1/OWFA u246

raw



enriched semantic features: {sign, sign/label/logo mark, text}

depleted semantic features: {}

"Images featuring written words or lettering on signs, labels, and other readable displays." (AIS = 0.950)

Figure B20

Top-20 Activating Images of an OWFA Unit (participant 05, replication 1, u56)

Top-20 activating images — GT/subj05/rep1/OWFA u56

raw



enriched semantic features: {food, textured surface}

depleted semantic features: {}

"The unit responds to close-up views of plated food and other object-centered, textured edible surfaces, while avoiding images with text, people or faces, upright standing figures, biological motion, or outdoor built scenes." (AIS = 0.950)

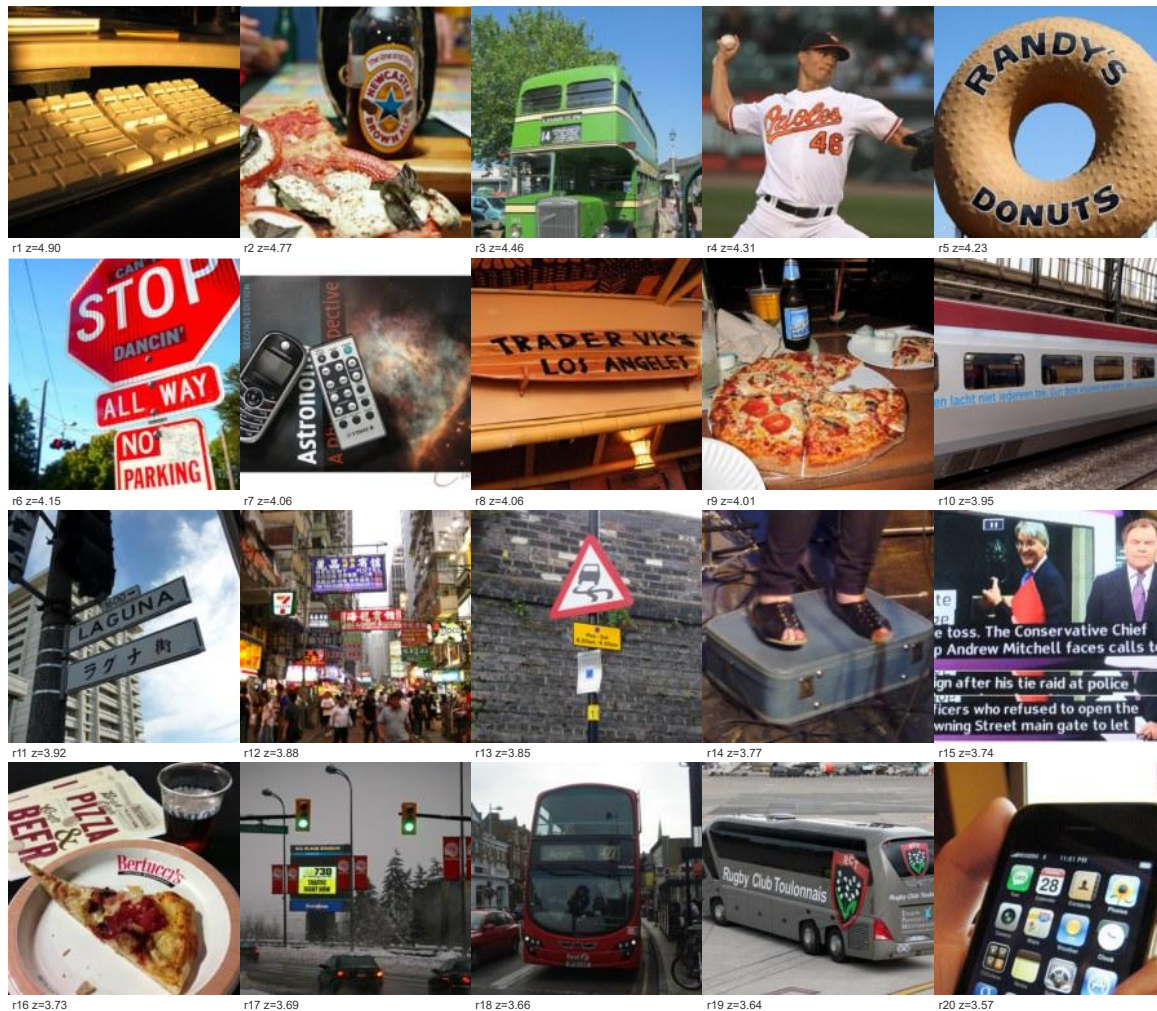
VWFA-1: Visual Word Form Area 1

Figure B21

Top-20 Activating Images of a VWFA-1 Unit (participant 02, replication 1, u201)

Top-20 activating images — GT/subj02/rep1/VWFA-1 u201

raw



enriched semantic features: {alphabetic/word-like text, numeric text, sign/label/logo mark, text present}

depleted semantic features: {}

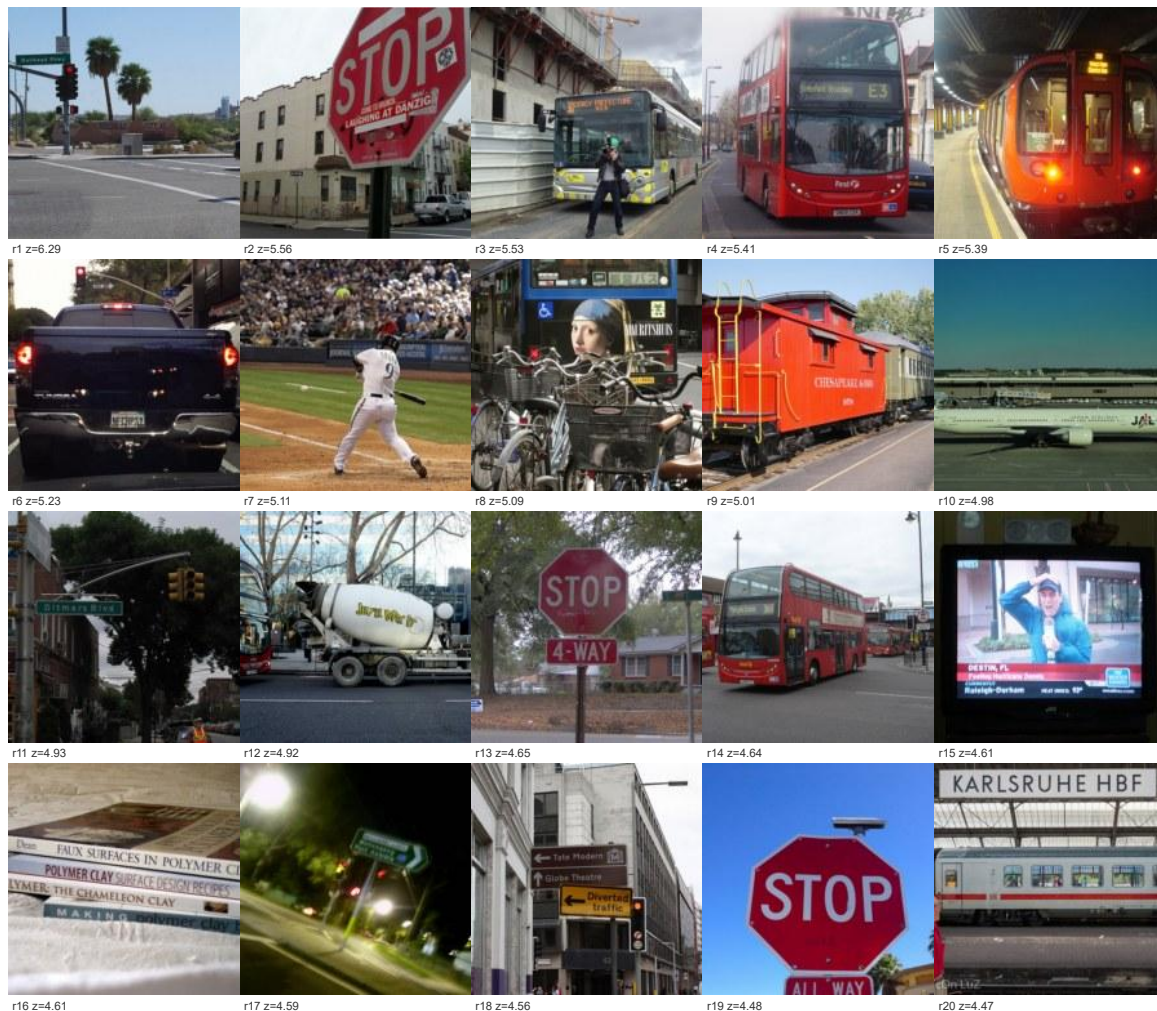
"A unit that responds to written markings and display-like graphics, including words, numbers, and logo or sign symbols." (AIS = 0.900)

Figure B22

Top-20 Activating Images of a VWFA-1 Unit (participant 05, replication 1, u23)

Top-20 activating images — GT/subj05/rep1/VWFA-1 u23

raw



enriched semantic features: {sign, street, vehicle, navigable/path-like scene, alphabetic/word-like text, numeric text, sign/label/logo mark, text present}

depleted semantic features: {}

"Images of roads and transit scenes that prominently include written signs, labels, or other text, often on vehicles or street-side markers." (AIS = 0.850)

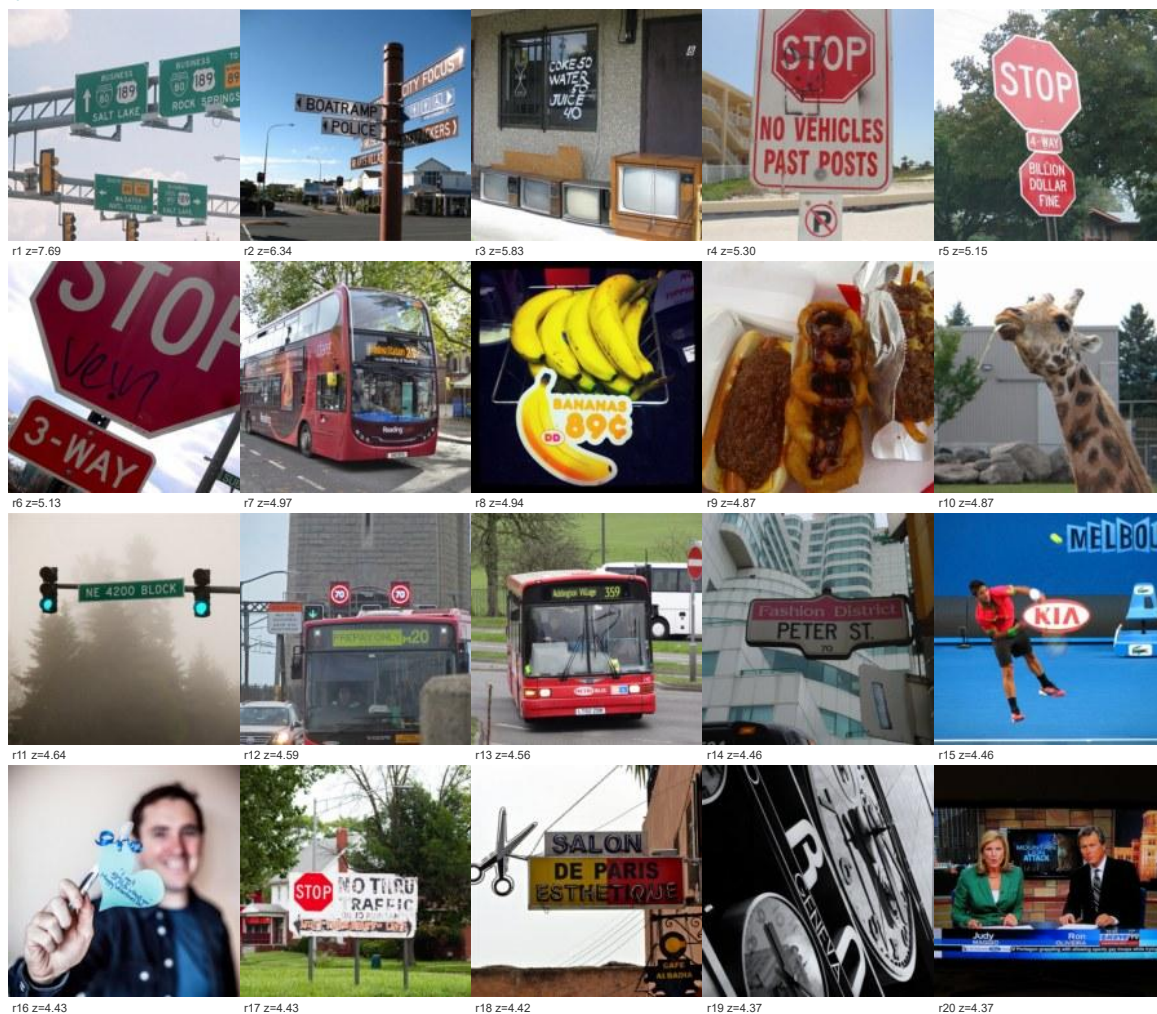
VWFA-2: Visual Word Form Area 2

Figure B23

Top-20 Activating Images of a VWFA-2 Unit (participant 07, replication 2, u69)

Top-20 activating images — GT/subj07/rep2/VWFA-2 u69

raw



enriched semantic features: {sign, alphabetic/word-like text, sign/label/logo mark}

depleted semantic features: {}

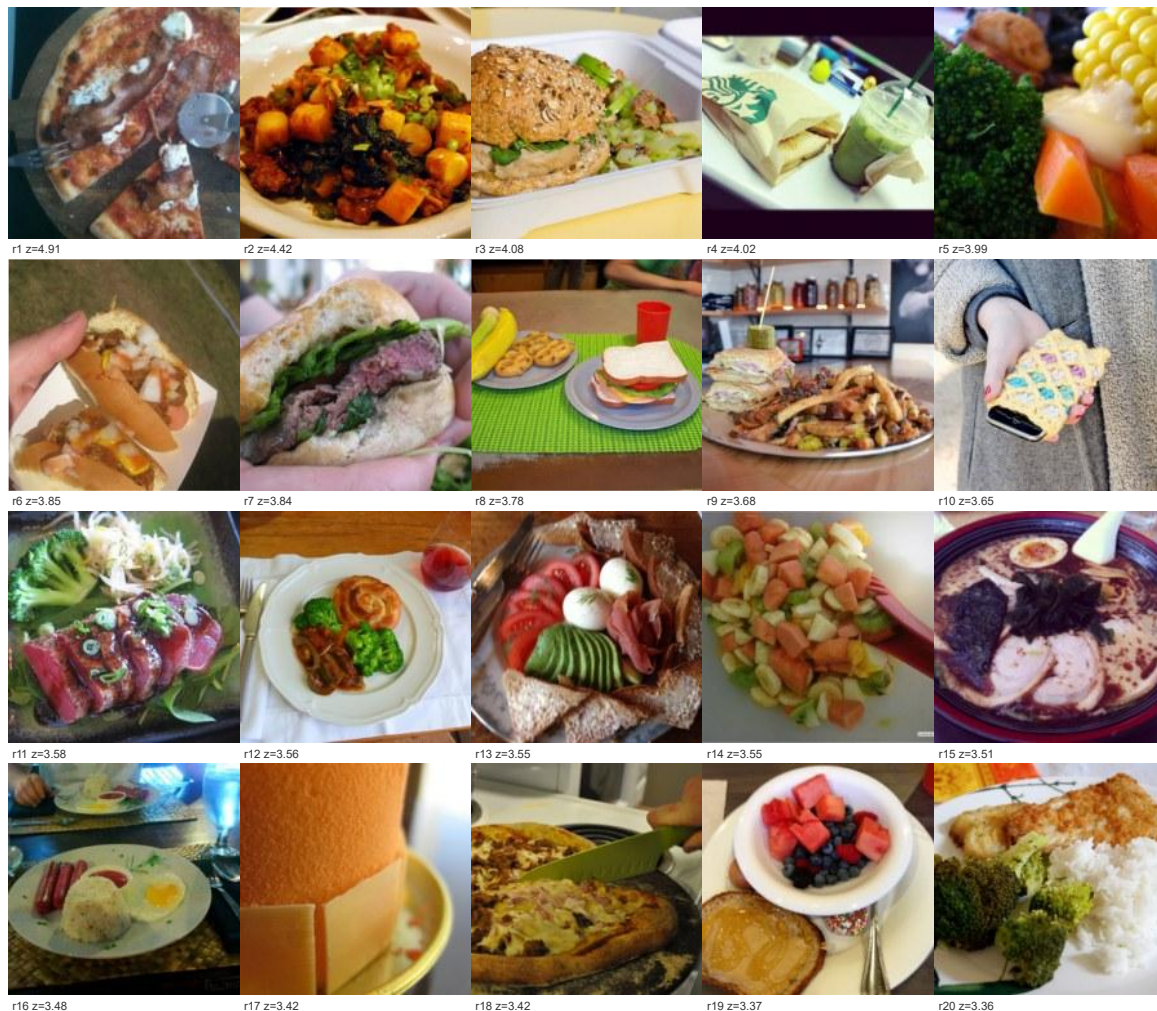
"Images containing written or printed signage, labels, and logo-like markings that convey text." (AIS = 0.800)

Figure B24

Top-20 Activating Images of a VWFA-2 Unit (participant 02, replication 1, u80)

Top-20 activating images — GT/subj02/rep1/VWFA-2 u80

raw



enriched semantic features: {food}

depleted semantic features: {outdoor built scene}

"The unit prefers images of prepared food and edible dishes, while tending not to appear in outdoor built settings or with people standing upright." (AIS = 1.000)

Appendix C: Demonstration of Discovered Primary Visual Representations

Two units are shown for each of the seven expert proxy channels. Each figure shows the 20 images that most strongly activate the unit, ordered by activation strength from left to right and top to bottom, together with the spatial expression map of the recovered feature over the same images, in which red marks local contribution above and blue below the channel prior. The footer of each montage lists the natural-language explanation generated for the recovered feature and the apparent interpretability score.

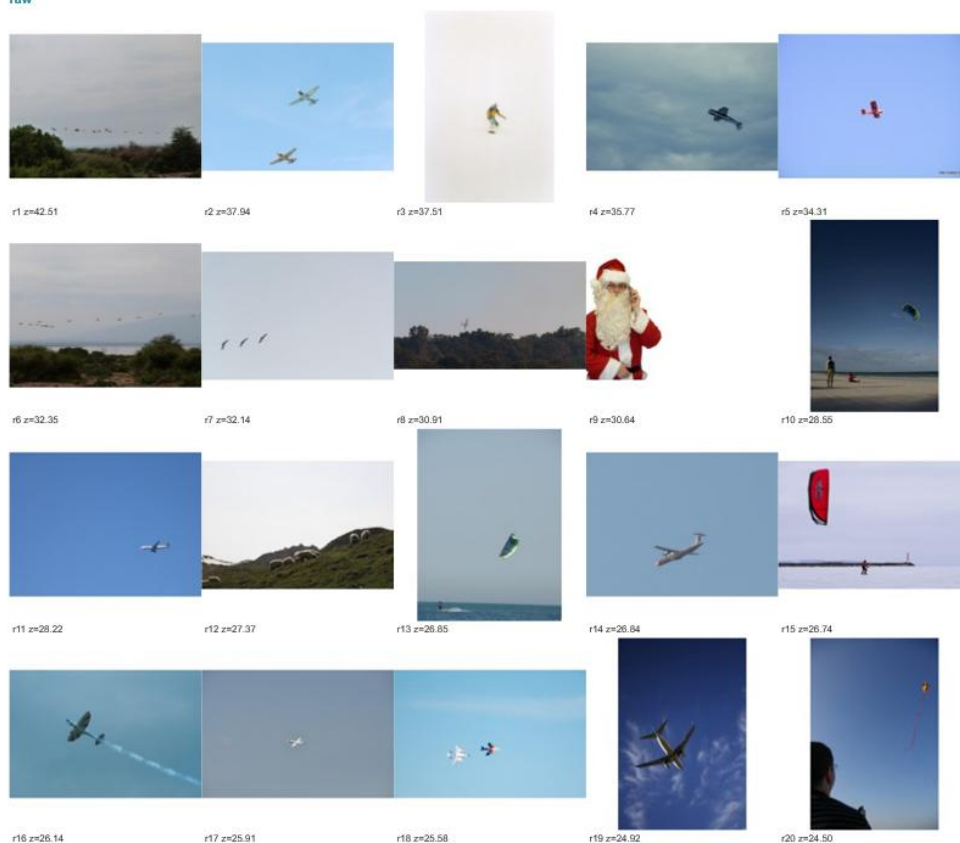
Oriented Energy and Spatial Frequency

Figure C1

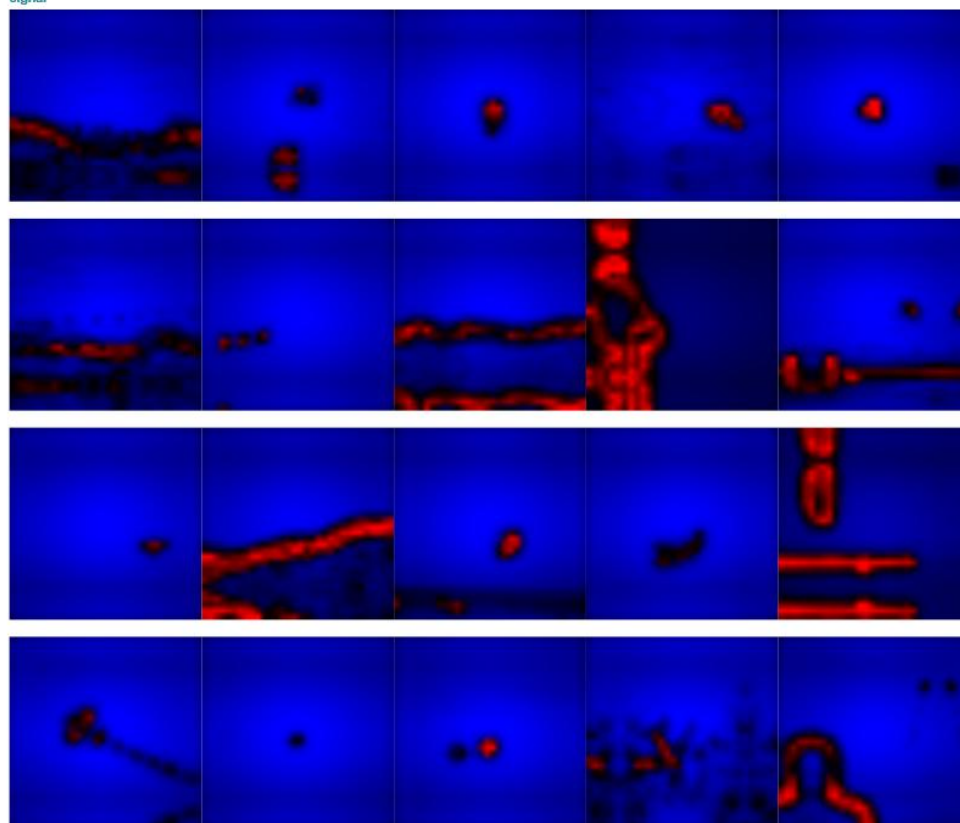
Top-20 Activating Images and Signal-Expression Maps of a V1d Unit (BEGD, participant 07, replication 1, u200)

Top-20 activating images — BEGD/subj07/rep1/V1d u200 (Oriented energy / SF)

raw



signal



~ (blue)
 "A broad horizontal band of low-frequency sky-like texture near the upper half of the image, often with a localized oriented streak or object silhouette cutting through it." (AIS = 1.000)

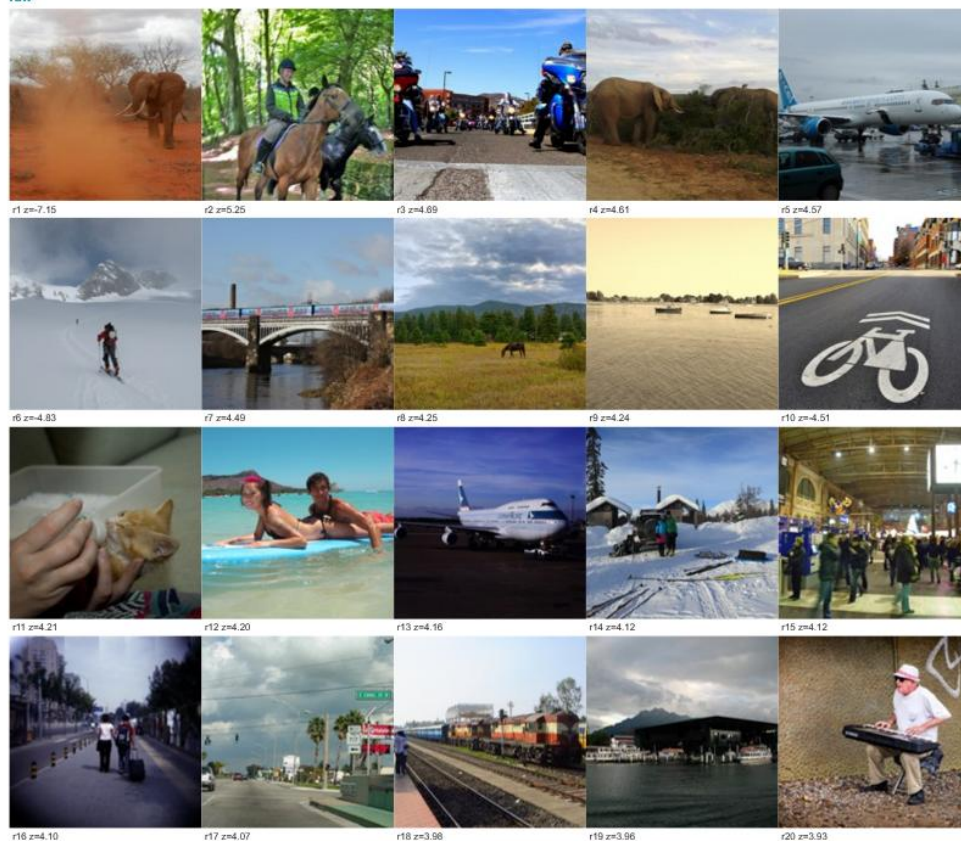
+ (red)

Figure C2

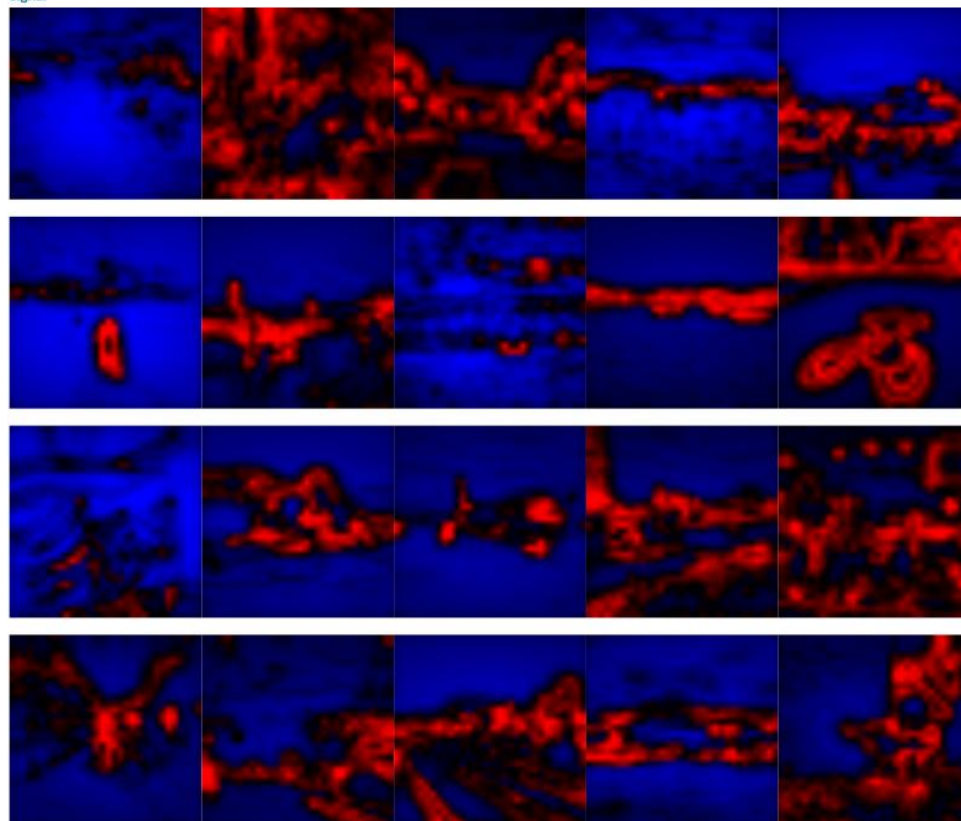
Top-20 Activating Images and Signal-Expression Maps of a V1v Unit (NSD, participant 02, replication 1, u909)

Top-20 activating images — GT/subj02/rep1/V1v u909 (Oriented energy / SF)

raw



signal



~ (blue)
 "A broad horizontal band of strong oriented texture energy concentrated across the upper half of the image, especially along sky/rooftop-like edges and other wide, layered high-frequency structures." (AIS = 1.000)

+ (red)

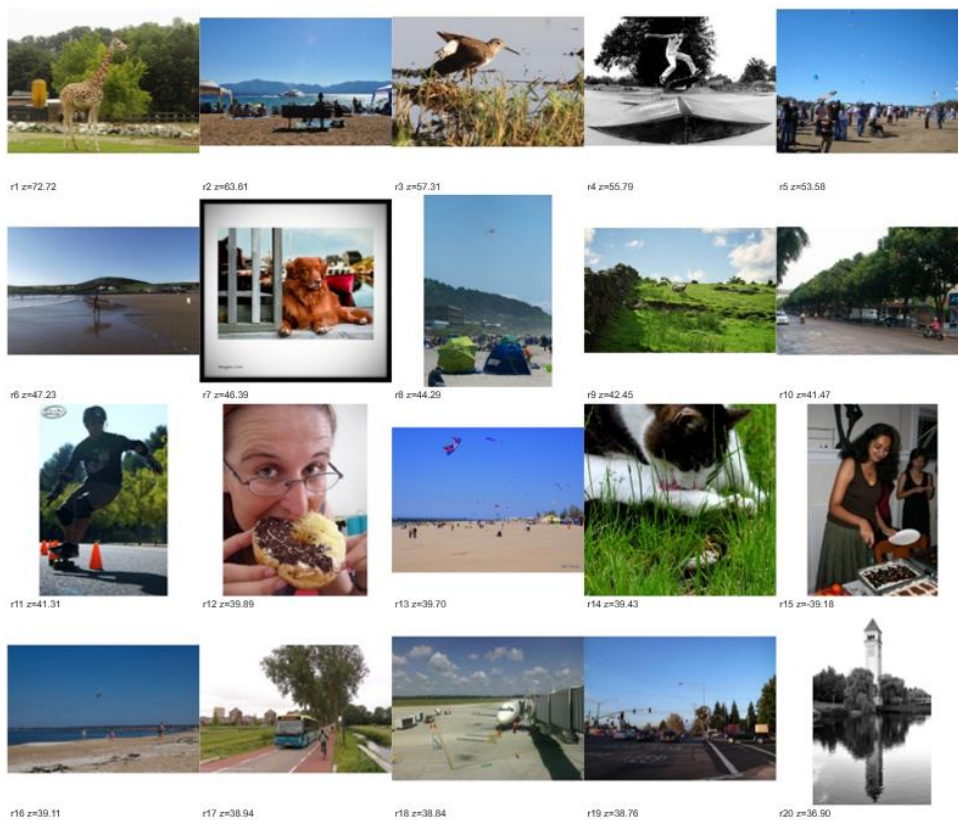
Structured Edges

Figure C3

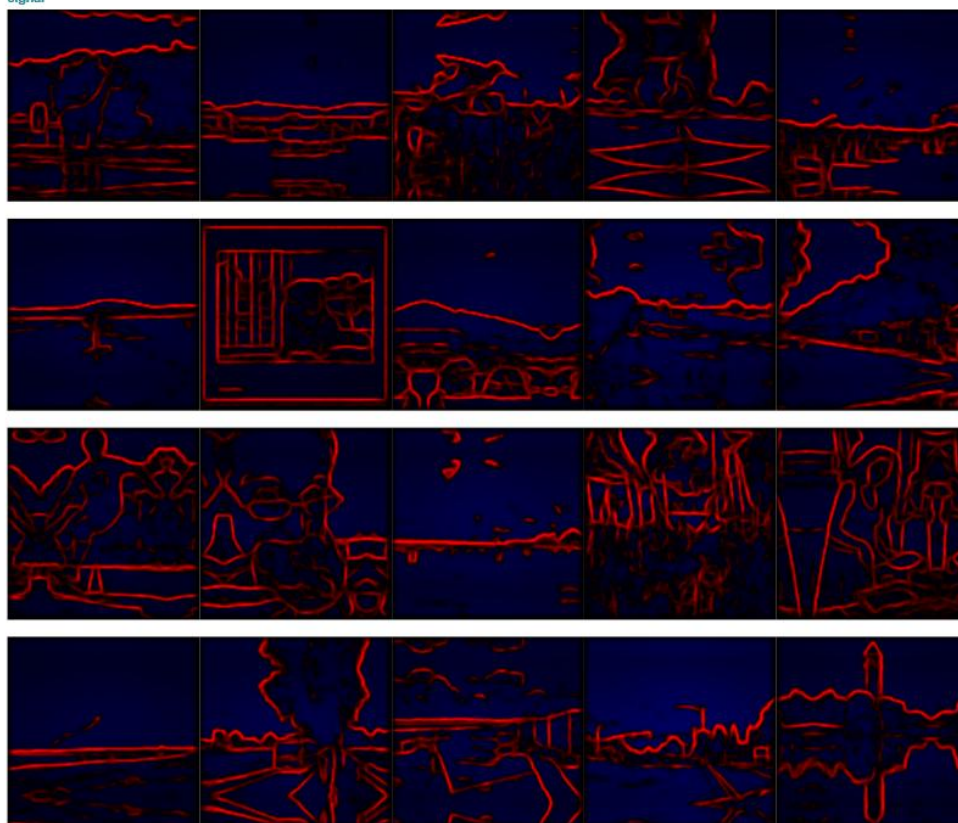
Top-20 Activating Images and Signal-Expression Maps of a V1d Unit (BEGD, participant 01, replication 1, u46)

Top-20 activating images — BEGD/subj01/rep1/V1d u46 (Structured edges)

raw



signal



~ (blue)
 "A strong horizontal boundary band across the middle of the image, often with additional thin contour fragments and junction-like structure above and below it." (AIS = 1.000)

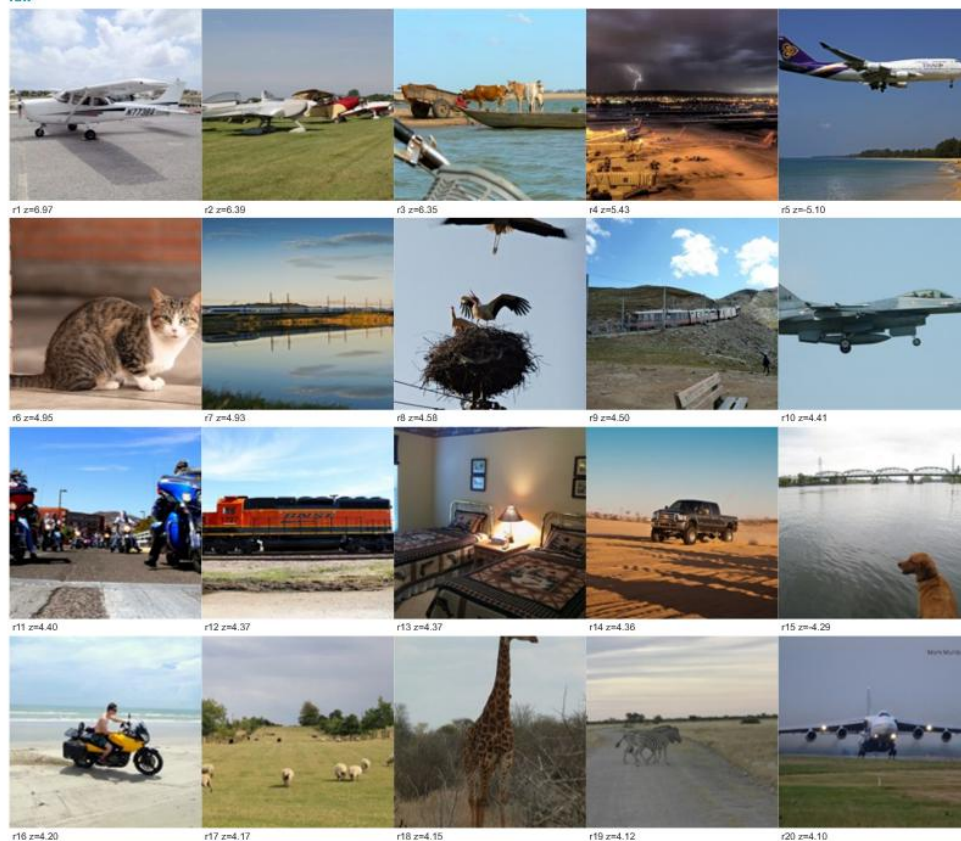
+ (red)

Figure C4

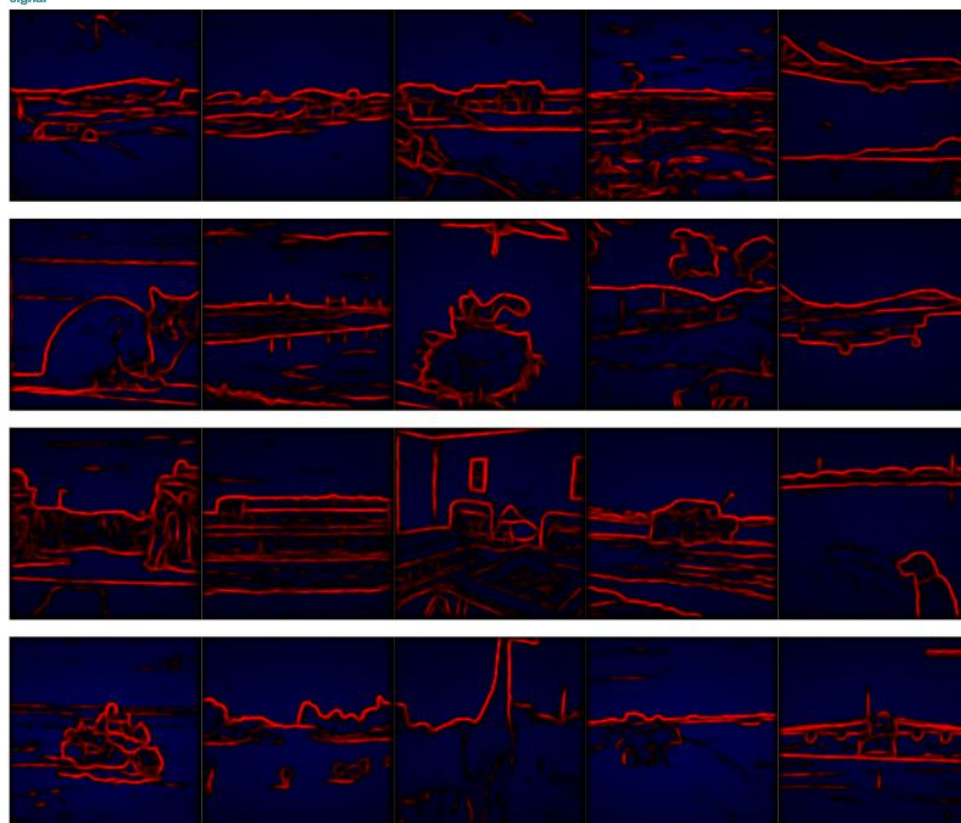
Top-20 Activating Images and Signal-Expression Maps of a V1v Unit (NSD, participant 01, replication 2, u117)

Top-20 activating images — GT/subj01/rep2/V1v u117 (Structured edges)

raw



signal



- (blue)
 "Strong horizontal boundary structure concentrated along distant horizon-like bands and elongated contour fragments, often with a few vertical support edges or thin junctions rising from the band." (AIS = 0.700)

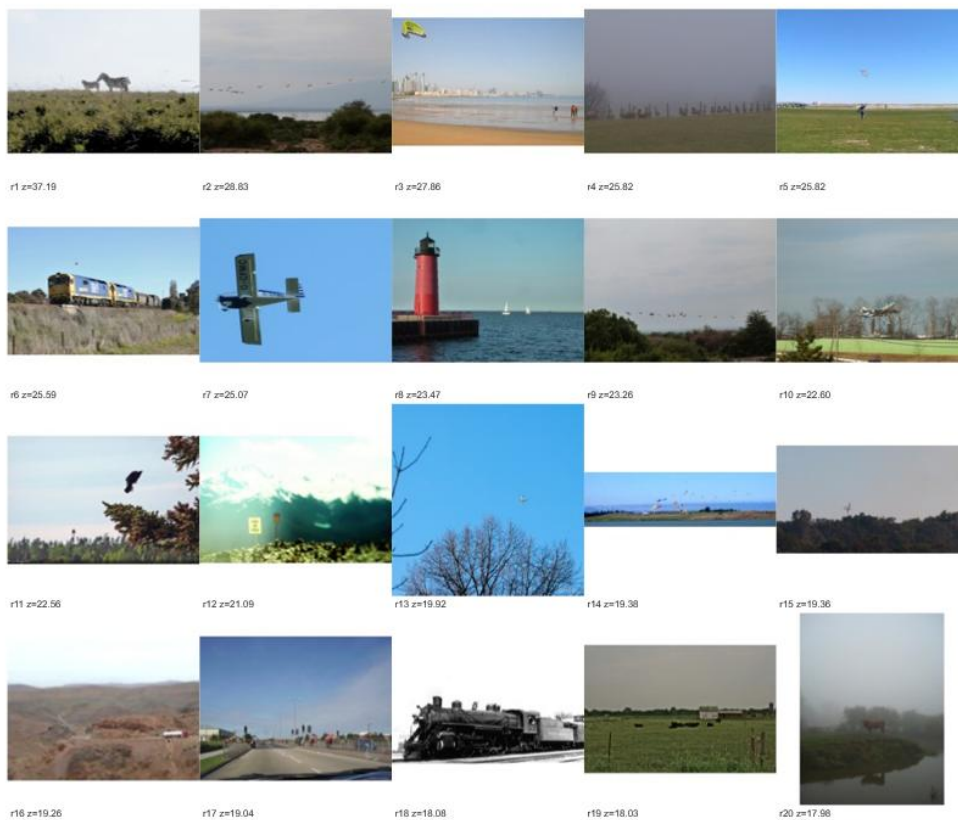
+ (red)

Opponent Colour**Figure C5**

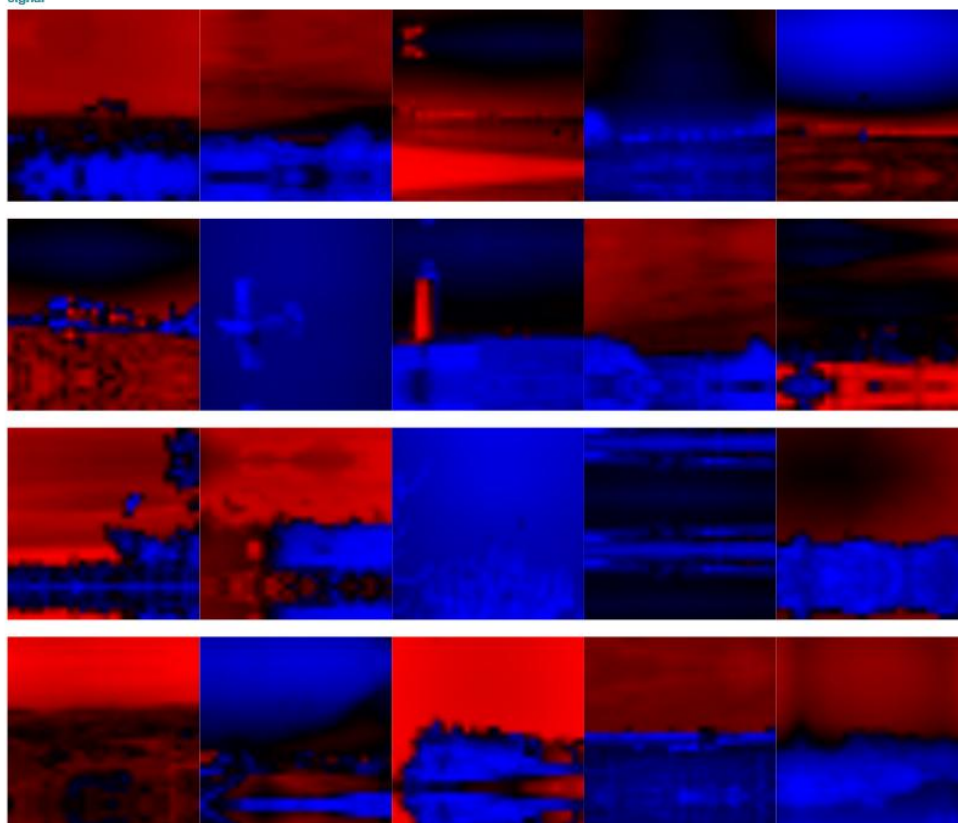
Top-20 Activating Images and Signal-Expression Maps of a hV4 Unit (BEGD, participant 07, replication 1, u210)

Top-20 activating images — BEGD/subj07/rep1/hV4 u210 (Opponent color)

raw



signal



— (blue)
 "A broad horizontal chromatic boundary or band, with strongest expression in the sky-horizon region and adjacent low-contrast color fields." (AIS = 1.000)

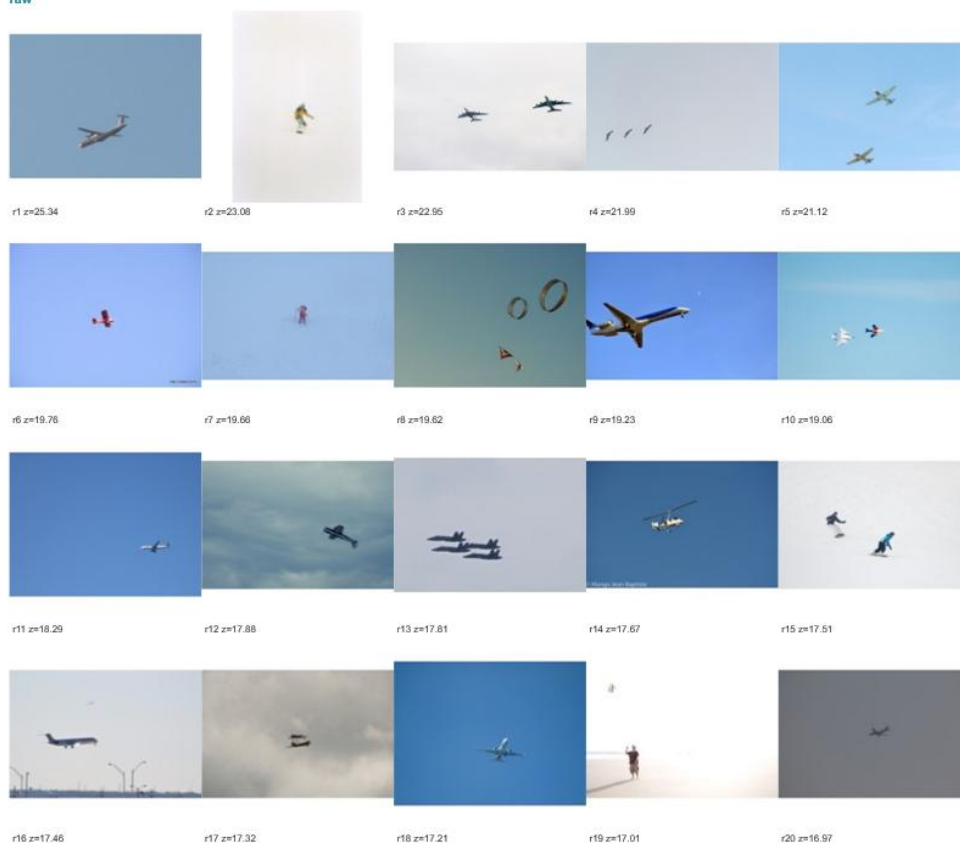
+ (red)

Figure C6

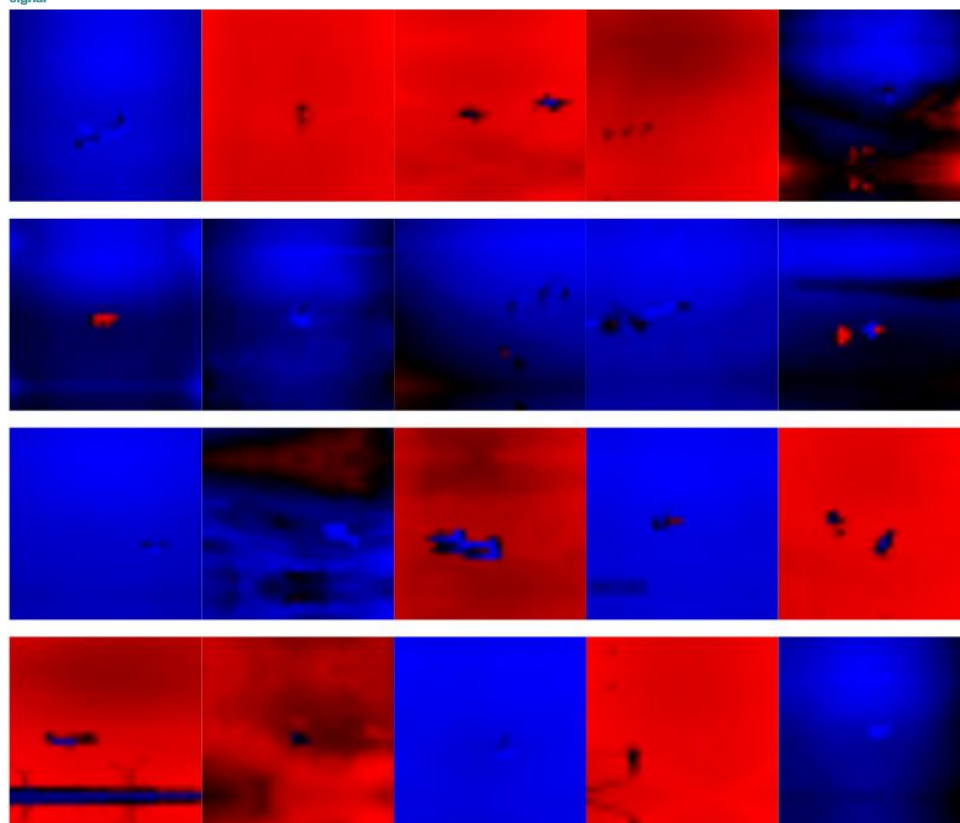
Top-20 Activating Images and Signal-Expression Maps of a V2d Unit (BEGD, participant 07, replication 1, u69)

Top-20 activating images — BEGD/subj07/rep1/V2d u69 (Opponent color)

raw



signal



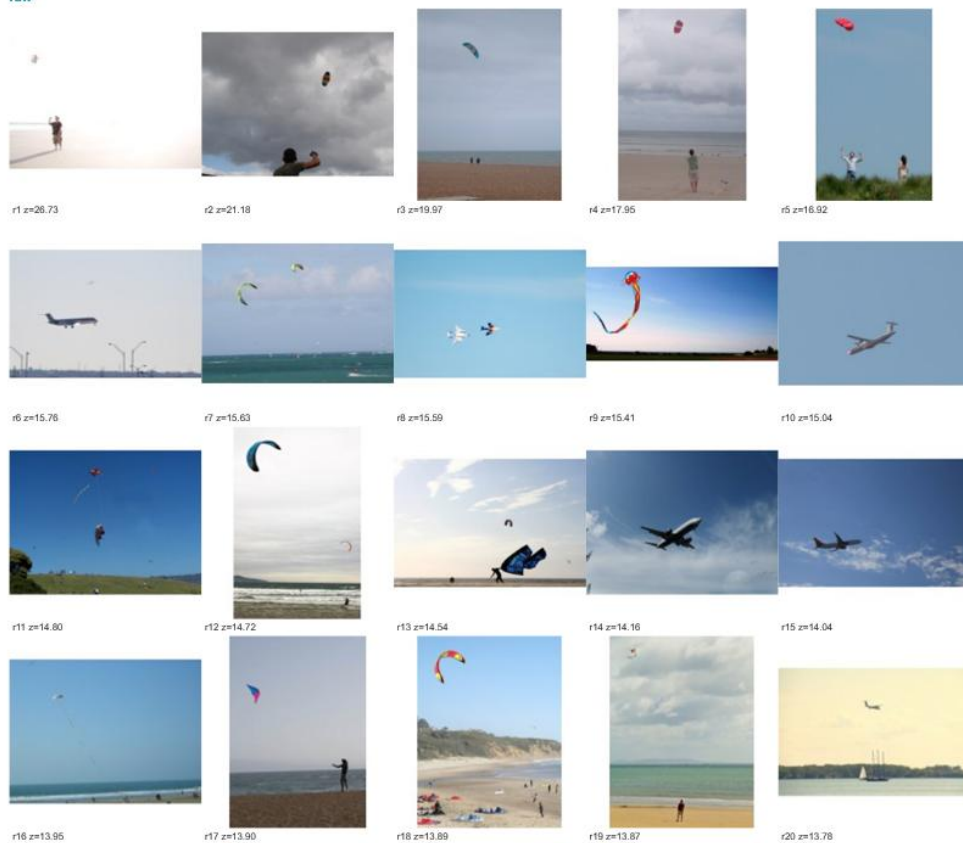
- (blue) + (red)
 "A broad opponent-color field that preferentially highlights small isolated objects and their local chromatic contrast against the surrounding sky or background, often centered on the object silhouette and nearby boundary regions." (AIS = 0.800)

Relative Depth**Figure C7**

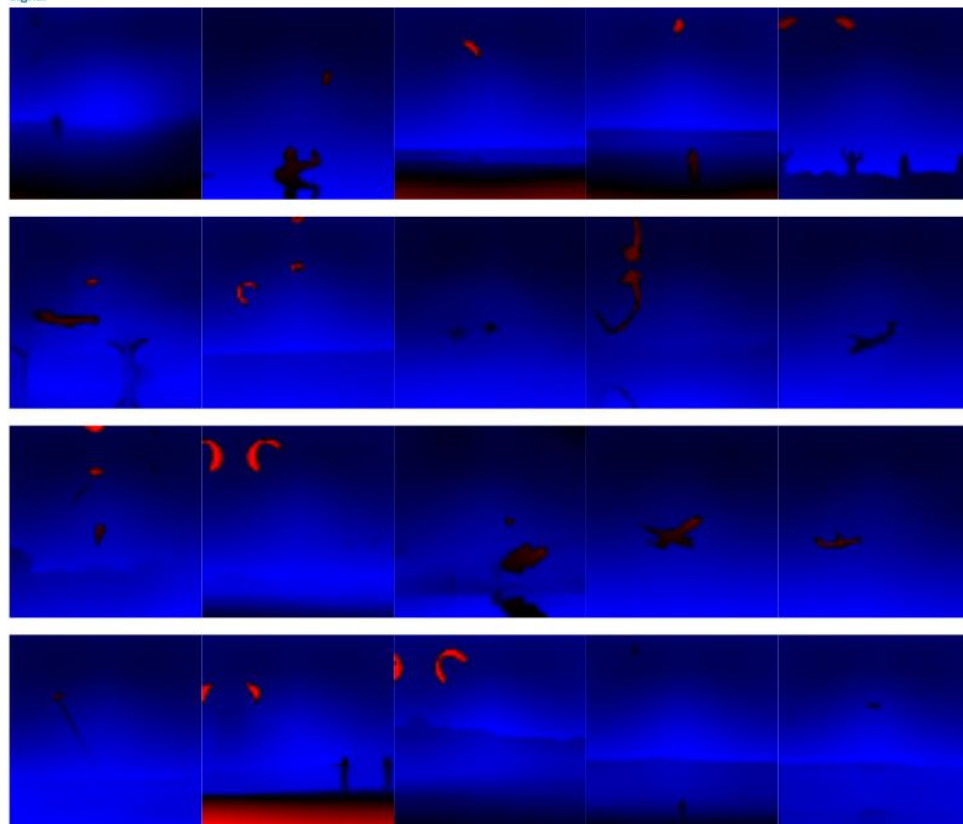
Top-20 Activating Images and Signal-Expression Maps of a V1v Unit (BEGD, participant 07, replication 1, u60)

Top-20 activating images — BEGD/subj07/rep1/V1v u60 (Relative depth)

raw



signal



~ (blue)
 "A strong depth-gradient signal concentrated along the lower horizon/ground plane, with foreground objects and vertical structures popping out against a receding sky or distant background." (AIS = 1.000)

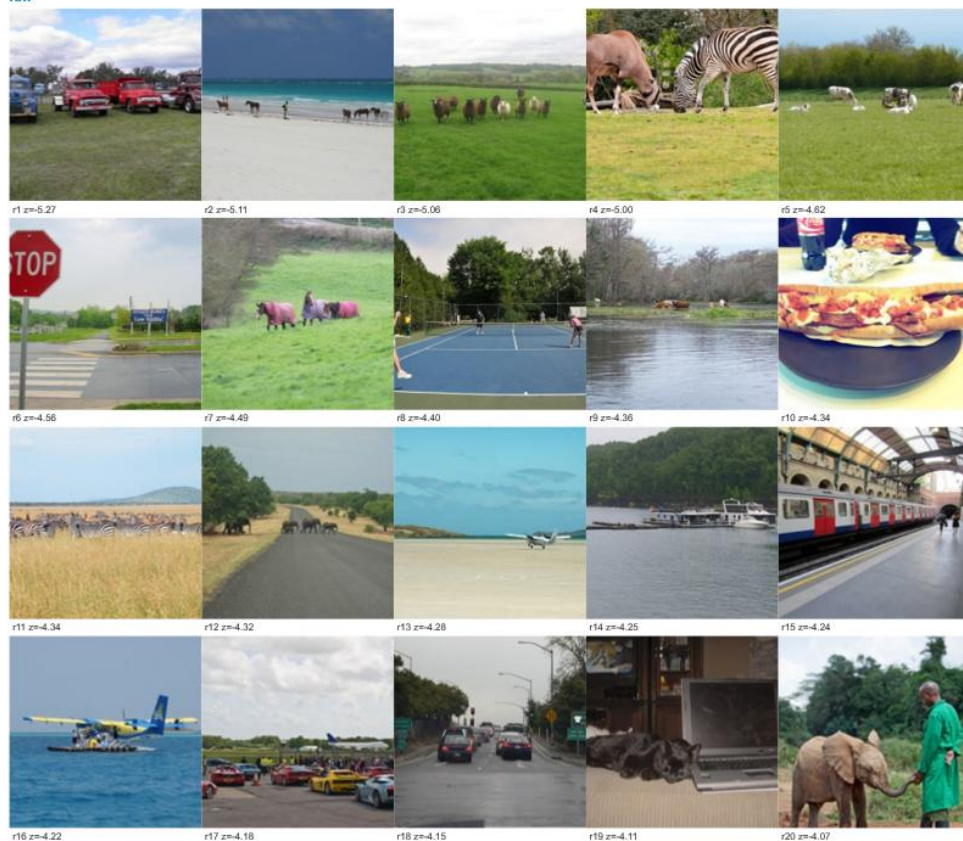
+ (red)

Figure C8

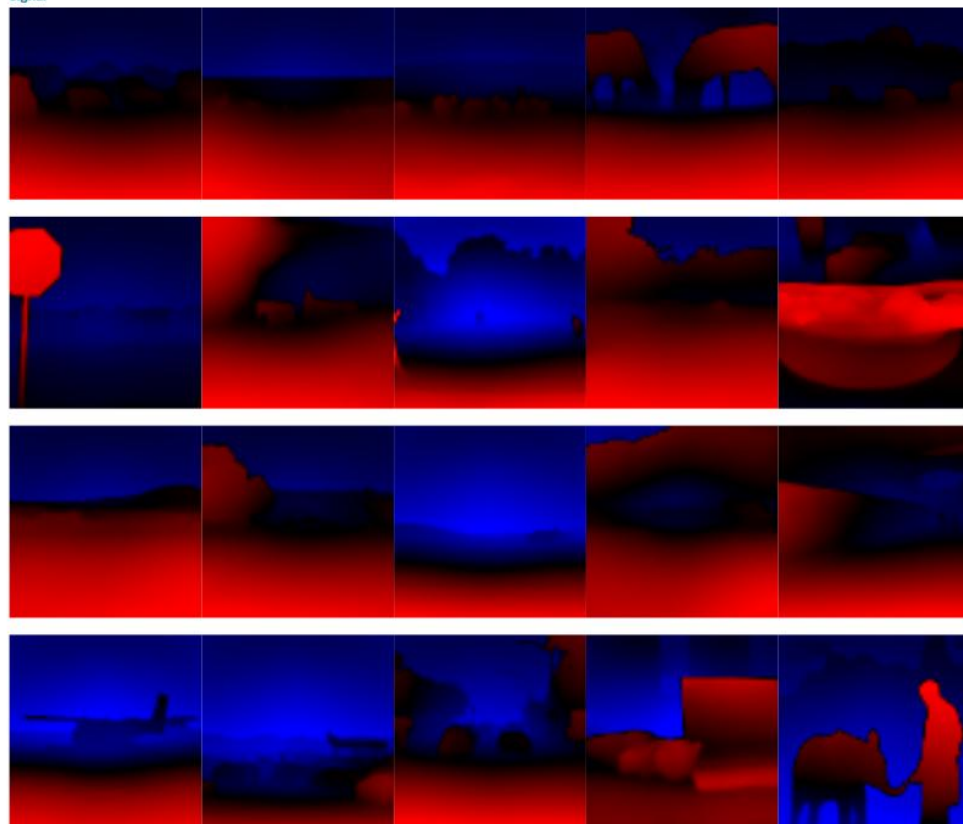
Top-20 Activating Images and Signal-Expression Maps of a V1d Unit (NSD, participant 05, replication 1, u19)

Top-20 activating images — GT/subj05/rep1/V1d u19 (Relative depth)

raw



signal



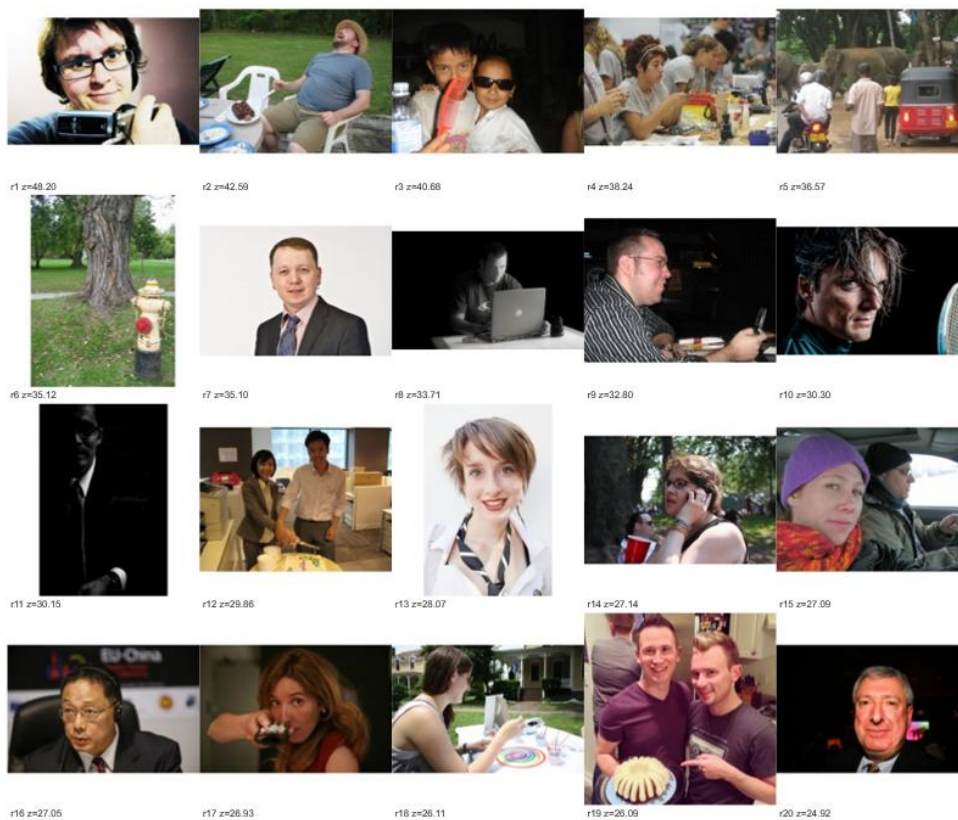
"A strong relative-depth signal concentrated in the lower foreground and along broad ground-plane or horizon transitions, emphasizing near-to-far layering and depth separation across the scene." (AIS = 1.000)

Saliency**Figure C9**

Top-20 Activating Images and Signal-Expression Maps of a hV4 Unit (BEGD, participant 07, replication 1, u91)

Top-20 activating images — BEGD/subj07/rep1/hV4 u91 (Saliency)

raw



signal



~ (blue)
 "A centered foreground person-shaped silhouette with strong emphasis on the head and upper torso, often extending into nearby hands or held objects while the background stays suppressed." (AIS = 0.850)

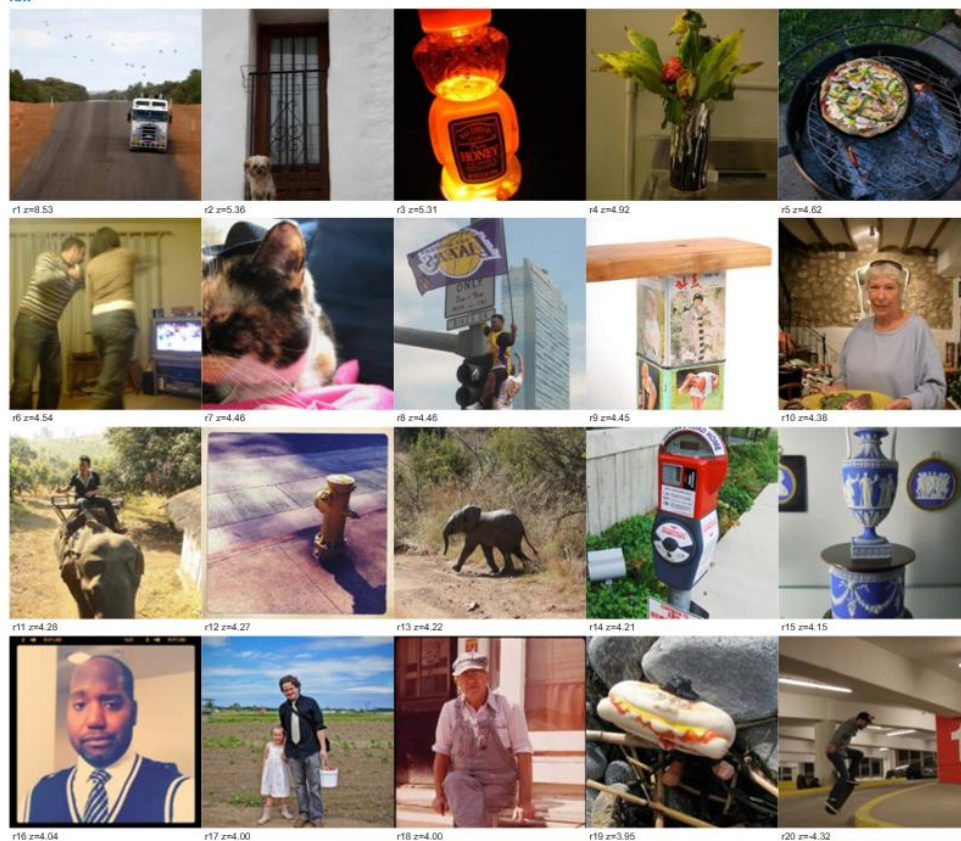
+ (red)

Figure C10

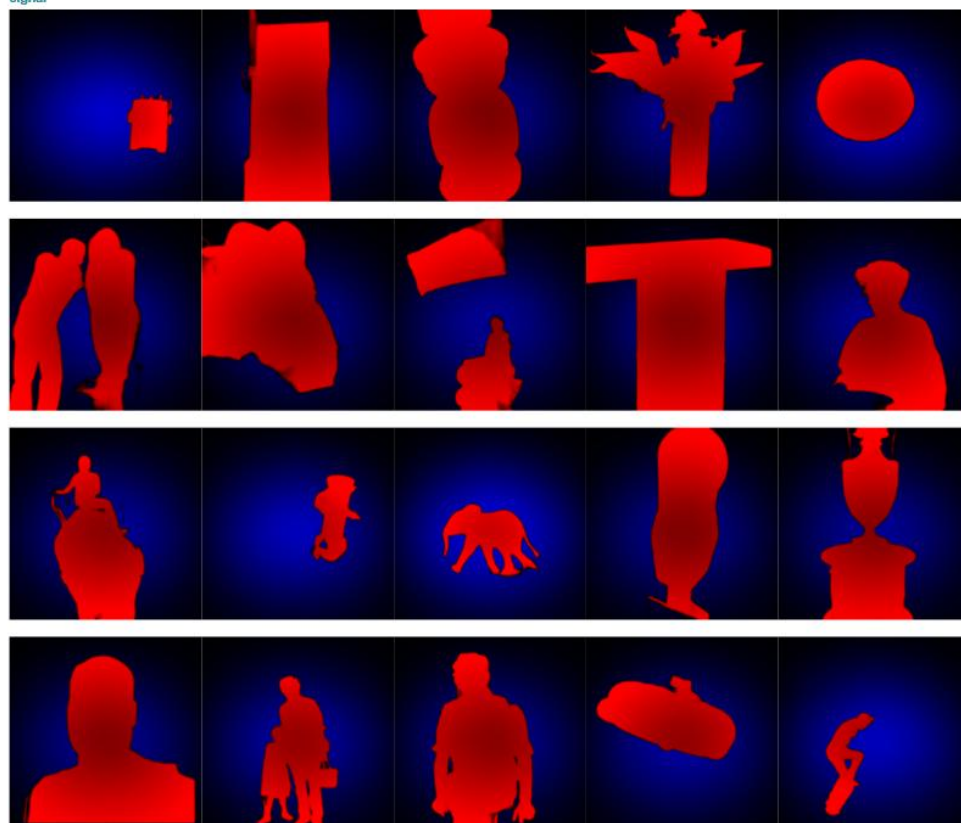
Top-20 Activating Images and Signal-Expression Maps of a V2d Unit (NSD, participant 02, replication 2, u820)

Top-20 activating images — GT/subj02/rep2/V2d u820 (Saliency)

raw



signal



— (blue)
 "A vertically centered foreground figure or object with strong silhouette-like saliency, often occupying the middle of the frame and standing out as a compact central mass against weaker surroundings." (AIS = 0.950)

+ (red)

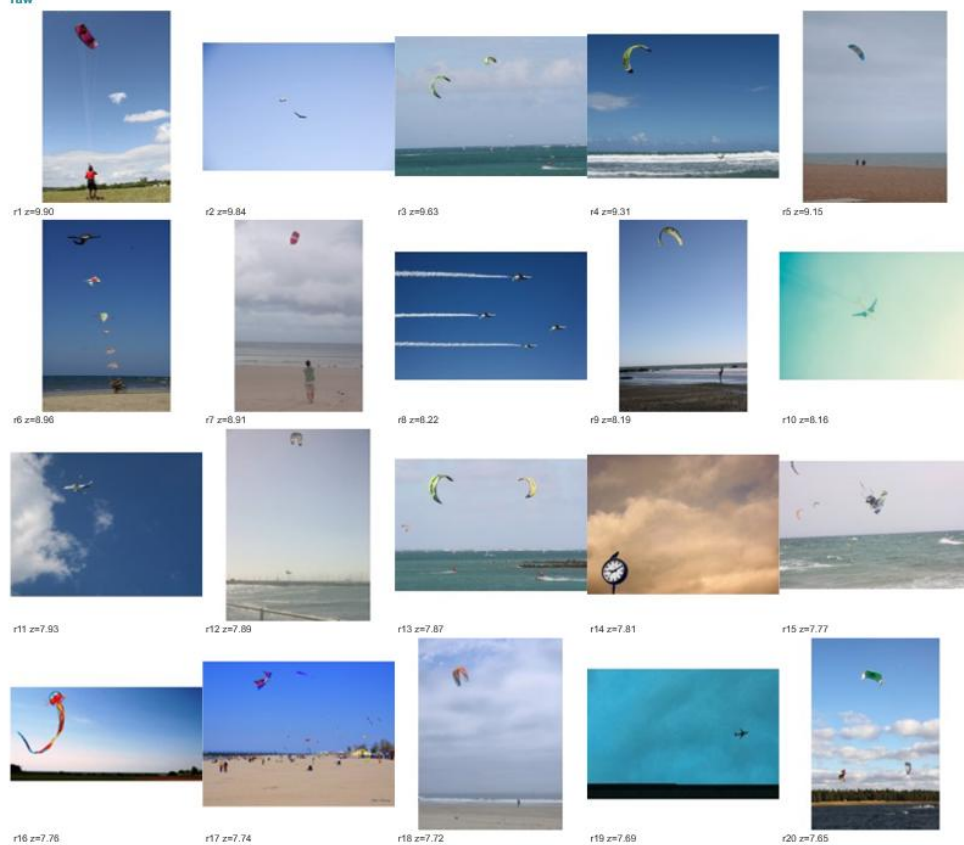
Symmetry

Figure C11

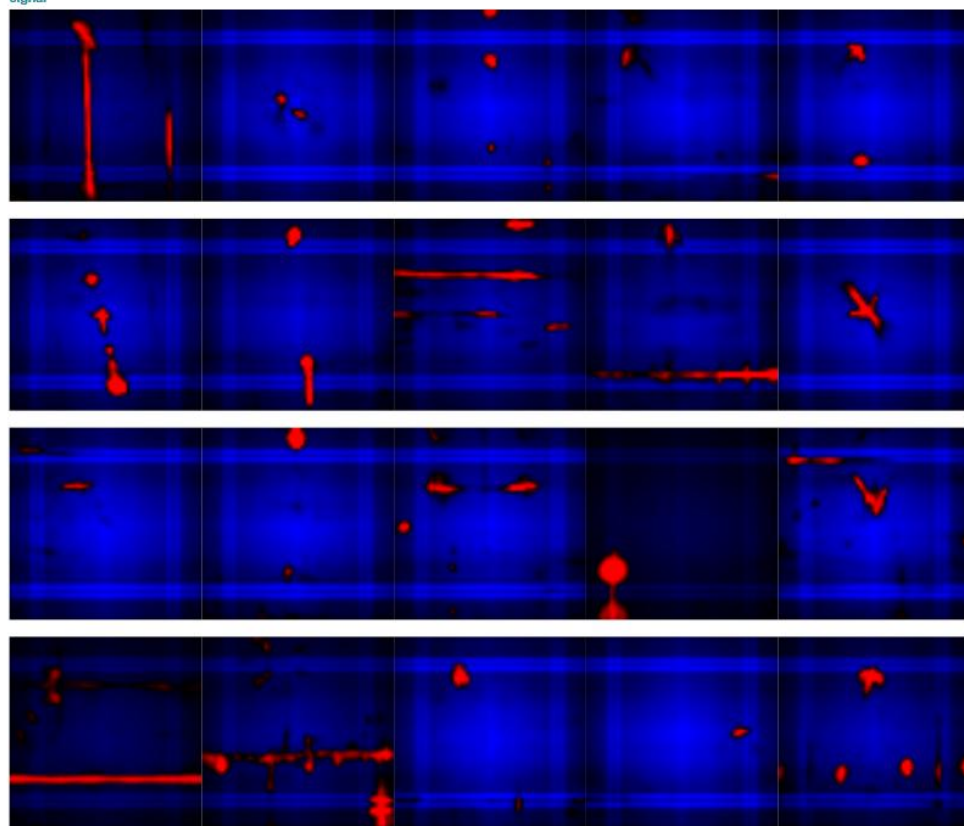
Top-20 Activating Images and Signal-Expression Maps of a hV4 Unit (BEGD, participant 01, replication 1, u169)

Top-20 activating images — BEGD/subj01/rep1/hv4 u169 (Symmetry)

raw



signal



— (blue)
 "A strong global horizontal balance signal with occasional vertical centerline fragments, most often expressed around the midline of open sky-and-horizon scenes rather than as a full symmetric object." (AIS = 1.000)

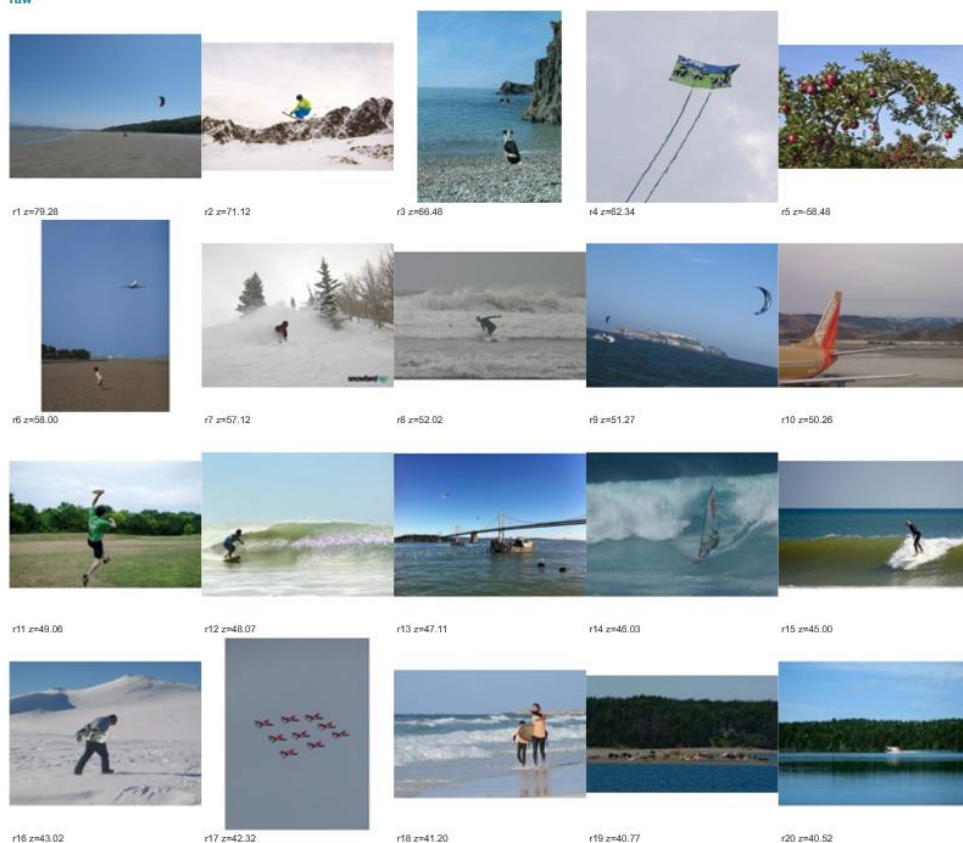
+ (red)

Figure C12

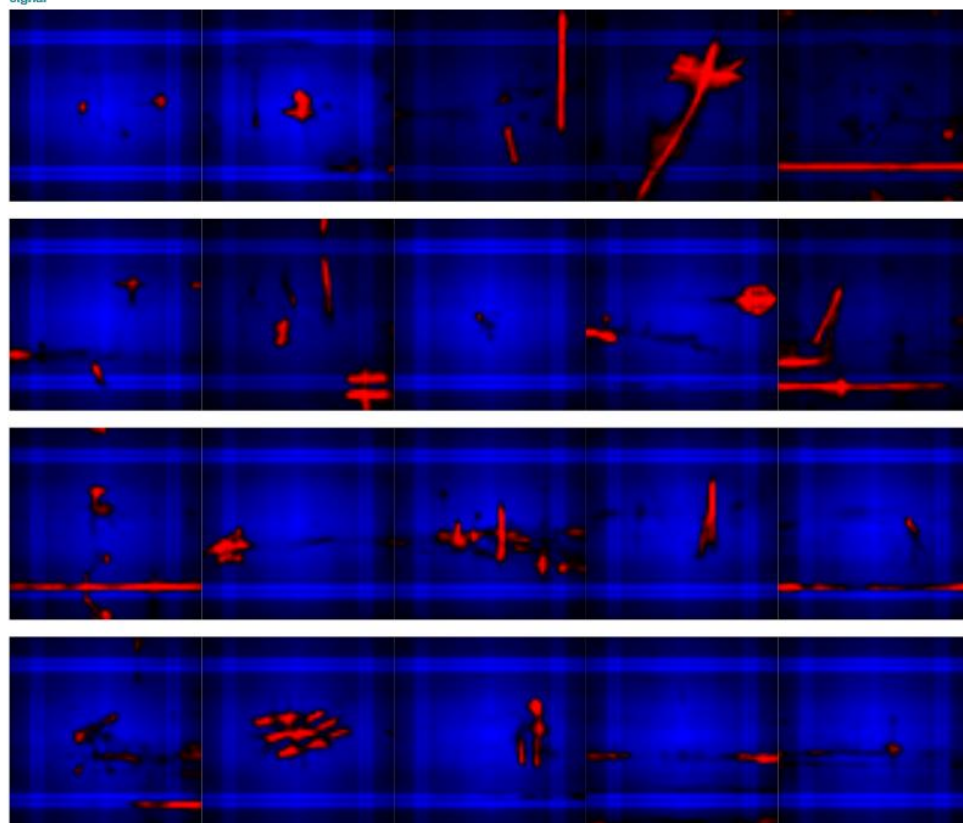
Top-20 Activating Images and Signal-Expression Maps of a V3d Unit (BEGD, participant 03, replication 1, u47)

Top-20 activating images — BEGD/subj03/rep1/V3d u47 (Symmetry)

raw



signal



"A broad horizontal horizon-like balance signal is expressed across the upper and lower image regions, with occasional localized vertical or diagonal symmetry fragments around centered upright structures." (AIS = 1.000)

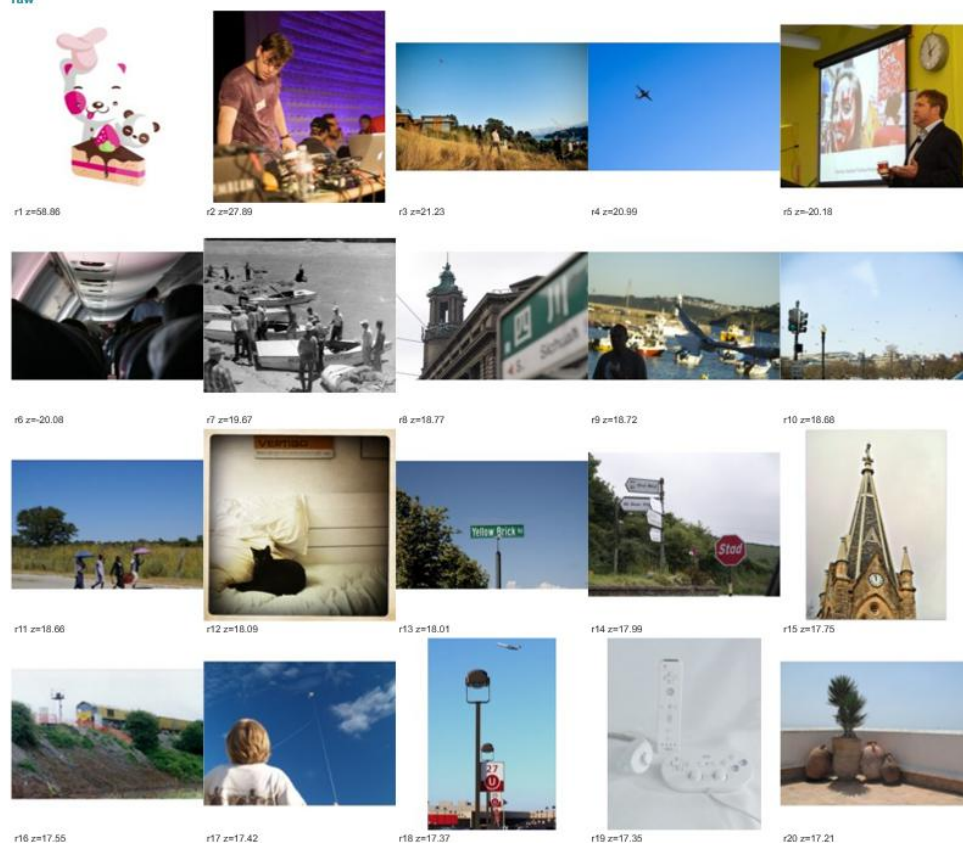
+ (red)

Skeleton**Figure C13**

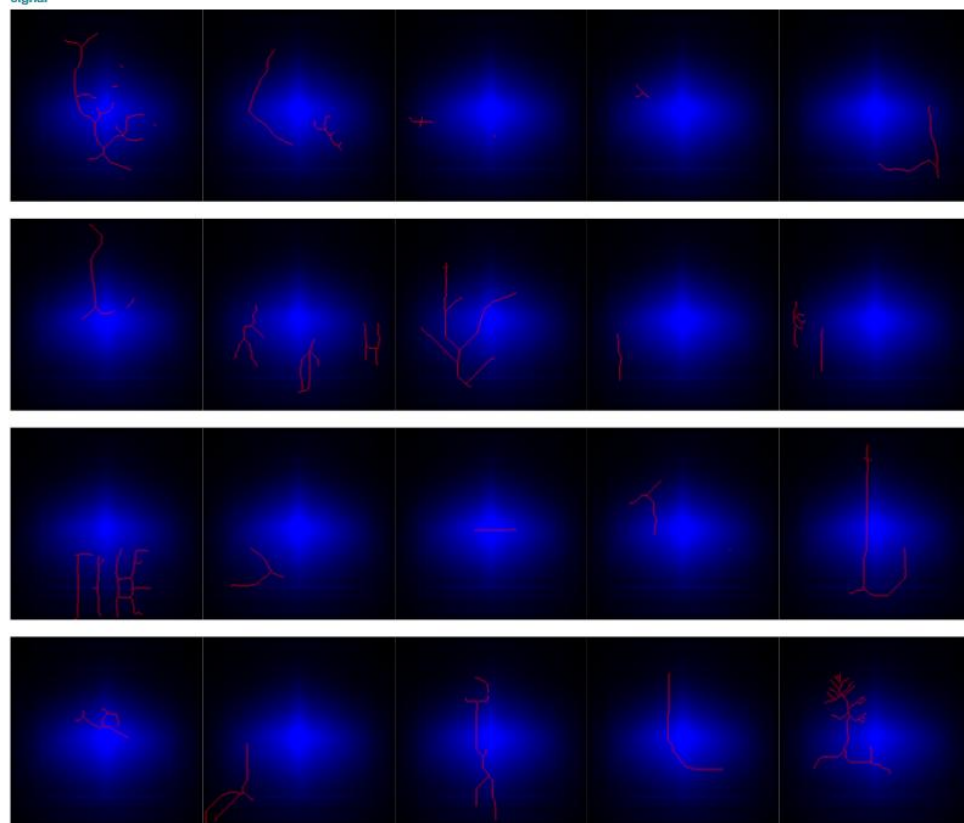
Top-20 Activating Images and Signal-Expression Maps of a V3d Unit (BEGD, participant 01, replication 1, u156)

Top-20 activating images — BEGD/subj01/rep1/V3d u156 (Skeleton)

raw



signal



— (blue)
 "A sparse skeleton signal concentrated on thin, elongated or branching structural cores, often with a few connected appendage-like axes and forked topological outlines rather than filled regions." (AIS = 1.000)

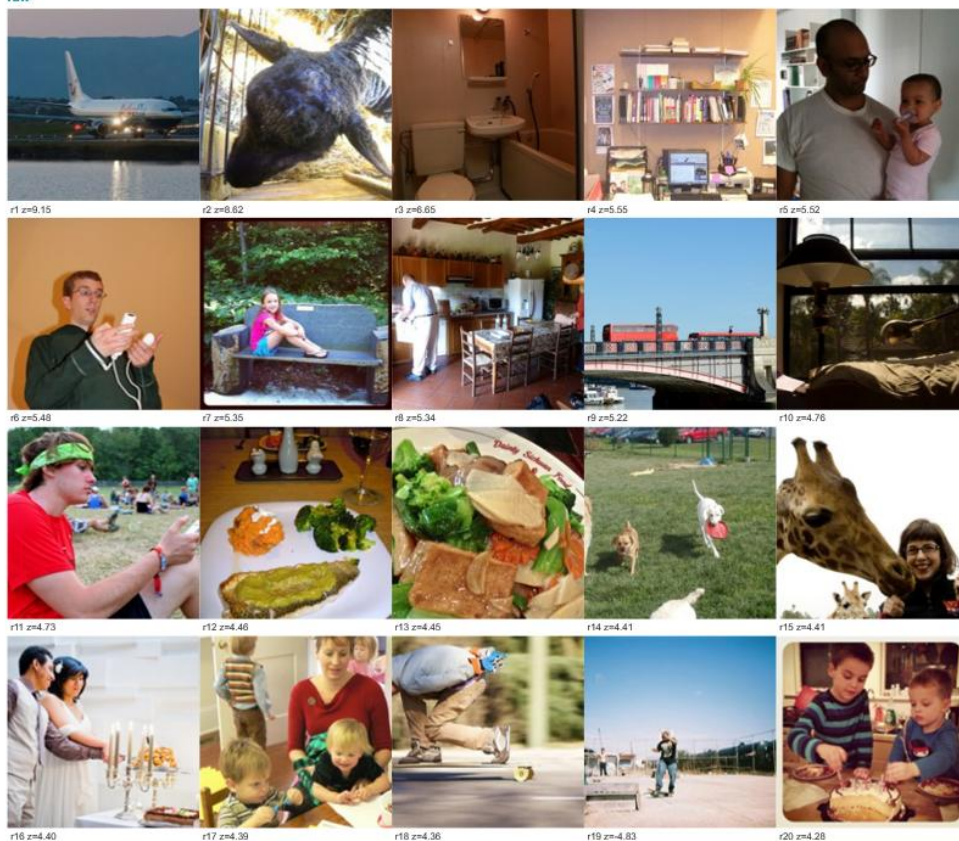
+ (red)

Figure C14

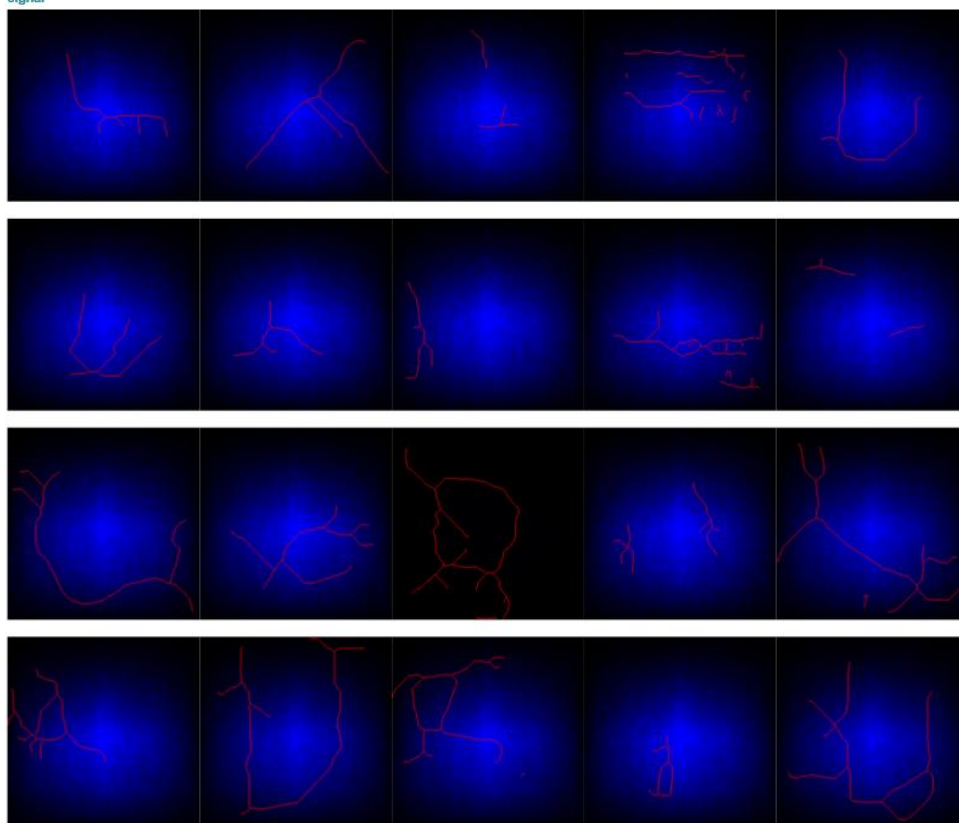
Top-20 Activating Images and Signal-Expression Maps of a V3v Unit (NSD, participant 04, replication 2, u983)

Top-20 activating images — GT/subj04/rep2/V3v u983 (Skeleton)

raw



signal



- (blue)

"A skeleton-like signal that emphasizes elongated connected contours and branching appendage structures, often tracing the main axis or limb-like topology of a foreground shape rather than its interior fill." (AIS = 0.950)

+ (red)