BAR: Probing Brain Encoders with Concept-Based Explanations

Huadi Wang Weihao Xia Cengiz Oztireli
University of Cambridge

https://github.com/hwjustin/BAR

Abstract

*Brain encoders have demonstrated promising capabilities in extracting semantic features from brain activity. However, the internal computations of these models remain largely opaque, which limits their adoption in critical brain research and applications. To address this challenge, we propose the **Brain Activation Region (BAR)** framework to investigate human-interpretable concepts learned by brain encoders and input features contributing to this learning. Specifically, we train kernel-based probes in the latent spaces of MindEye and UMBRAE, two state-of-the-art models that interpret viewed images from fMRI signals. We further apply a feature attribution approach to concept density functions, evaluating specific brain voxels and regions sensitive to visual semantics. Our trained classifiers demonstrate high accuracy across diverse visual and semantic concepts, effectively explaining the predictions made by brain encoders. Additionally, the feature attribution reveals two regions of interest (ROIs) associated with visual concept processing in human brains, aligning with findings in recent neuroscience research.*

1. Introduction

Recent neuroscience studies have introduced effective brain encoders to interpret thoughts and perceptions from brain activity [8, 11, 12, 16, 17]. Based on fMRI signals collected when subjects viewed colour natural scenes, brain encoders extract conceptual and spatial features aligned with CLIP encoders [8, 11], variational autoencoders (VAE) [11, 12, 16], or multimodal large language model (MLLM) [15], enabling brain-to-image retrieval, reconstruction, captioning, and grounding. While these models demonstrate promising potential in brain-computer interfaces (BCI) and cognitive state analysis, their internal computations remain opaque and poorly understood. Without clear interpretability of brain encoders, it might not be safe to apply them in critical neuroscience applications.

To address this challenge, concept-based explanations could reveal how brain encoders extract features through the lens of user-specified concepts. One such approach is the Concept Activation Regions (CAR) [2], which trains a kernel-based probe in the model's representation space. By separating samples where a concept is present (concept positive) or absent (concept negative), this classifier could explain how specific concepts are represented by brain encoders. The CAR approach could be further combined with feature attribution methods, such as Integrated Gradient [13] and Gradient Shap [7], to identify input features that contribute to the model's concept learning.

In this work, we propose **Brain Activation Region (BAR)**, a unified concept-based explanation framework applicable to any brain encoder. Specifically, we apply CAR explanation to two state-of-the-art brain encoders—MindEye [11] and UMBRAE [15]—to investigate the concepts learned by these models and to identify brain voxels that contribute to specific concept learning. Our results show high classification accuracy across many visual and semantic concepts derived from the images presented to subjects, effectively explaining the predictions made by brain encoders. Furthermore, we identify regions of interest (ROIs) in human brains that are sensitive to certain visual concepts. By introducing concept-based explanations into brain encoding models, we provide a novel approach for researchers to explore the relationship between visual semantics and brain function.

2. BAR: Brain Activation Region

We propose the Brain Activation Region (BAR) framework to interpret both representative encoders through concept-based explanations. Two distinct pretrained brain encoders MindEye [11] and UMBRAE [15] are selected for our experiments based on their popularity and representativeness. They are trained on the Natural Scene Dataset (NSD) [1], which includes fMRI recordings in response to visual stimuli from COCO [6], using different objectives. MindEye produces embeddings trained with contrastive loss, while UMBRAE learns to reconstruct CLIP intermediate features using an element-wise reconstruction loss. The overview of BAR is illustrated in Fig. 1. First, we generate concept-positive and concept-negative voxel samples and input them

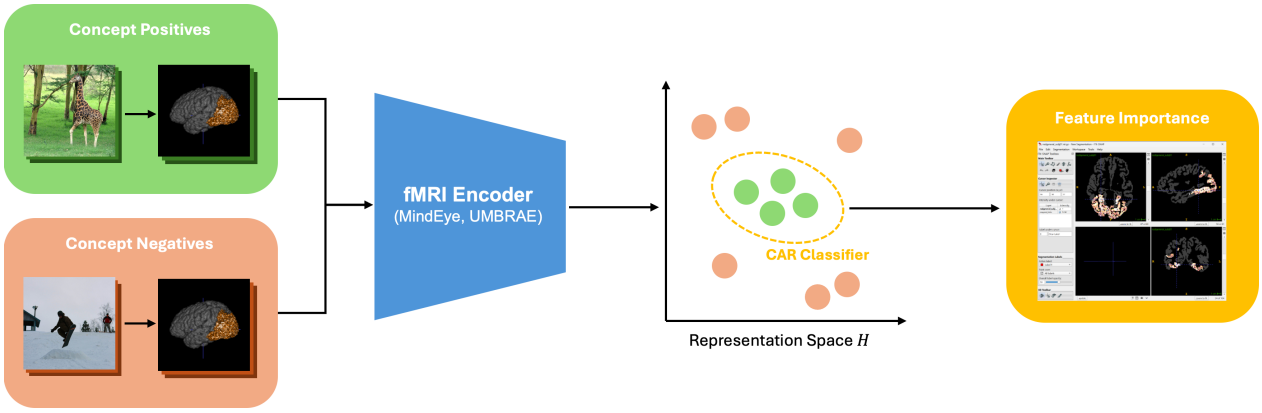

Figure 1. **Overview of Brain Activation Region (BAR).** BAR interprets pretrained brain encoders through concept-based explanations. BAR includes three steps: (a) Concept Generation (Sec. 2.1), (b) CAR Classifier Training (Sec. 2.2), (c) Voxel Visualisation (Sec. 2.3).

into brain encoders to obtain the corresponding embeddings (Sec. 2.1). Next, we train a CAR classifier to distinguish these embeddings and determine concepts learned by MindEye and UMBRAE models (Sec. 2.2). Last, we apply Integrated Gradient to evaluate voxel contributions to specific concept learning. The attributions and potential brain ROIs sensitive to visual semantics are visualised using a 3D medical imaging software ITK-SNAP [18] (Sec. 2.3).

2.1. Concepts and Embeddings Generation

To explain the conceptual information learned by brain encoders, we want to find ground truth concepts that capture the semantic content of individual fMRI inputs. Therefore, a natural and effective choice is to use the COCO image categories [6] of corresponding visual stimuli. These categories are manually annotated to describe the main objects in each image and have been widely adopted in object detection research. To obtain image category labels, we map each voxel-image pair in the NSD [1] back to its original COCO index, retrieving the annotation using the provided COCO API [6].

Using this concept definition, we construct probe datasets for CAR classifiers. For each concept and subject, we randomly select k positive samples containing the concept label and k negative samples without it. These voxel inputs are then passed through the brain encoders to obtain the corresponding embeddings. For the MindEye encoder [11], we load subject-specific pre-trained weights to account for individual differences in brain responses. For the UMBRAE encoder [15], we use a shared pre-trained model due to its cross-subject training strategy and subject-specific tokenizers. Finally, we assign the label `1` to concept-positive embeddings and `0` to concept-negative embeddings to formulate the binary classification task.

2.2. CAR Classifier Training

Following the procedure outlined in the CAR paper [2], we optimise a kernel-based Support Vector Classifier (SVC) to distinguish between concept positive and negative embeddings for each image category and subject. CAR assumes concept smoothness in the latent space $\mathcal{H}$, where concept positive and negative examples are scattered across distinct clusters. Under this assumption, we train an SVC $s_k^c : \mathcal{H} \rightarrow \{0, 1\}$ to partition $\mathcal{H}$ into concept activation regions $\mathcal{H}^c$ where concept c is mostly present and regions $\mathcal{H}^{\neg c}$ where concept c is mostly absent. In our case, if concept smoothness holds and the SVC captures a clear boundary between positive and negative clusters, this indicates that the brain encoder has learned the concept and made embeddings separable. To assess how well each concept is learned, we measure the classification accuracy of trained SVCs on their corresponding test sets.

2.3. Voxel Attribution and Visualisation

We further evaluate the relevance of individual voxels in identifying specific concepts. Specifically, we apply the Integrated Gradient approach to the concept density function defined in the CAR paper [2], computing the corresponding feature importance. Since our input voxels belong to `nsdgeneral`, a subset of voxels responsive to the NSD experiment in the posterior aspect of cortex, we map voxel contributions back to full-brain coordinates using the provided mask. This mapping enables us to localise the most sensitive voxels for various concepts. Finally, we visualise and qualitatively assess these contributions using the 3D medical imaging software ITK-SNAP, identifying potential brain ROIs associated with visual semantics.

3. Experiment

Applying our BAR framework to the MindEye and UMBRAE encoders, we design and organise our experiments

Table 1. **Concept Classification**. Mean accuracies are reported for the pretrained MindEye [11] and UMBRAE [15] encoders.

Concept Category	MindEye	UMBRAE
Creatures	88.8	87.9
Household Items	85.9	88.8
Transportation	85.8	89.7
Food	90.9	93.4
Everyday Objects	86.7	88.1

to answer two key questions about these models: (1) What semantic concepts do the two brain encoders learn for predictions, and how do they compare? (2) Which subset of voxels contribute to certain concept learning, and how do these voxels relate to specific ROIs or brain functions?

3.1. Concept Accuracies

Based on COCO image categories, we define 47 concepts to describe the semantic information from the input. With $k = 200$, we generate concept-positive and concept-negative embeddings to train each CAR classifier. To evaluate the effectiveness of concept learning, we group these concepts into five broad semantic categories and present the mean classification accuracies averaged over concepts and subjects, as shown in Tab. 1.

The results show that our trained CAR classifiers achieve high accuracy across most visual and semantic concepts, with an overall average performance of $(86.7 \pm 6.3)\%$ for MindEye embeddings and $(88.9 \pm 6.1)\%$ for UMBRAE embeddings. This classification performance suggests that both MindEye and UMBRAE encoders inherently capture these concepts from voxel input during their own training processes. We believe that this concept learning explains predictions from brain encoders and supports brain-to-image tasks such as captioning, grounding, retrieval, and reconstruction.

Despite relatively consistent performance across broad semantic categories, we observe notable accuracy variations between individual concepts for both encoders. For example, the accuracies for "sink", "skis", "toilet", "tennis racket", and "giraffe" exceed 95%, while the accuracies for "backpack", "bird", and "bicycle" are around 80%. We hypothesise two potential reasons for this pattern. First, high-accuracy concepts like "sink" and "giraffe" tend to have distinct shapes and colours in the visual stimuli, which may be clearly reflected in brain responses and learned by both encoders. In contrast, low-accuracy concepts like "backpack" and "bird" may exhibit greater variability in appearance, which are more challenging for brain encoders to capture. Second, the observed accuracy pattern aligns with the salience of individual concepts. It is known that certain stimuli are prioritised for attention and processing in the salience network of the human brain [3, 5, 9, 14]. Based on the BrainHub [15], our high-accuracy concepts "sink", "toilet", and "giraffe" fall into the "Salient" category, which may be strongly represented in the voxel input and contribute to higher CAR accuracy.

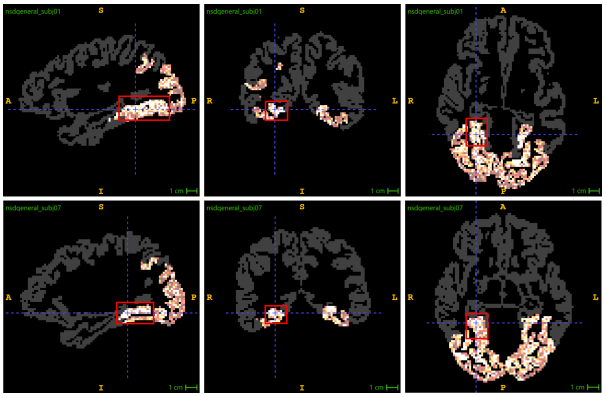

Figure 2. **Concept Attribution.** Feature attribution for the "person" concept in `sub01` (top) and `sub07` (bottom).

Comparing classification performance between CARs trained on MindEye and UMBRAE embeddings, we notice that UMBRAE yields slightly better results, particularly for concepts such as "bicycle", "couch", "potted plant", "spoon", and "book". This suggests that the UMBRAE encoder learns these semantic concepts more effectively than the MindEye encoder. We attribute this difference to the distinct training objectives of the two models. MindEye maps fMRI brain activity to CLIP image space through contrastive learning, which prioritises broad content associations. As a result, it may struggle to capture fine-grained details within complex scenes, such as "potted plant" or "spoon". In contrast, UMBRAE maps fMRI representations to image features via element-wise reconstruction, enabling more precise semantic and spatial alignments and capturing relevant concepts.

3.2. Feature Attribution

To localise brain voxels involved in concept learning, we apply the Integrated Gradient approach [13] to compute feature attributions and visualise the results using ITK-SNAP [18]. Based on empirical analysis, we identify two potential ROIs sensitive to specific visual semantics.

3.2.1. Right Fusiform Gyrus - "Person" Concept

Recognising the concept of "person" is one of the most fundamental cognitive functions in human perception and social interaction. To find specific brain regions sensitive to this concept, we analysed the corresponding voxel contributions, as shown in Fig. 2.

From the results, we discover a highlighted area (bounded by the red box) across all subjects, located in the right temporal and occipital lobes. With further research, we believe this brain region might belong to the

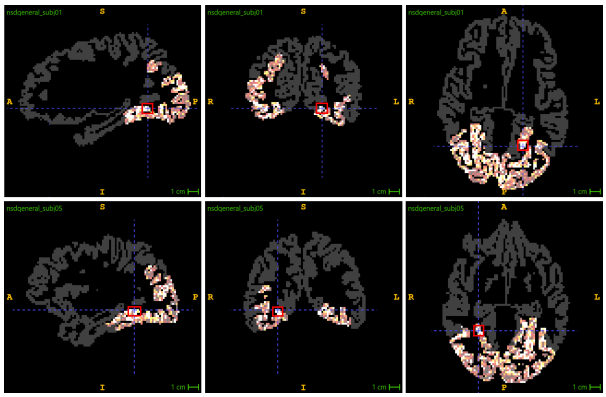

Figure 3. **Concept Attribution.** Feature attribution for the "giraffe" concept in `sub01` (top) and `sub05` (bottom).

right fusiform gyrus, which is known for face and body recognition. According to the dynamic visual stimulation experiment conducted by Jiang *et al*. [4], face categorisation may begin in the right fusiform face area (FFA). Notably, the individually localised right FFA identified in this study appears to correspond to the highlighted area we observed. Therefore, this region might be activated when subjects view human faces in the NSD experiment, which is subsequently leveraged by brain encoders to learn the "person" concept. In addition to the right FFA, we also observe some smaller highlighted regions in the voxel attributions, which may reflect other perceptual features relevant to the "person" concept. We leave the investigation of these additional areas to future neuroscience research.

3.2.2. Visual Area V4 - "Giraffe" Concept

In Sec. 3.1, we hypothesised that the "giraffe" concept may correspond to distinct shape and colour features in the visual stimuli. To explore potential brain regions involved in shape and colour processing, we examine the feature attributions for the "giraffe" concept, as illustrated in Fig. 3.

We find two highlighted spots (bounded by the red box) across all subjects, located in the left and right occipital lobes respectively. Based on our analysis, we speculate that these spots may correspond to the V4 area in visual cortex. Prior neuroscience studies have associated this brain region with colour processing and object recognition [10, 19], which may be specifically activated to process the distinct visual characteristics of "giraffe".

However, our voxel attribution results remain somewhat noisy, making it challenging to localise a well-defined brain region solely responsible for the "giraffe" concept. This could imply that multiple brain regions cooperate to process complex semantic information. Additionally, the NSD experiment was conducted on a general image dataset with diverse backgrounds, objects, and lighting conditions, which may have introduced confounding signals. More controlled experimental settings would be beneficial to precisely identify relevant brain regions.

4. Discussion

Our BAR framework has demonstrated effective capability to explain brain encoders and discover potential brain functions. Leveraging the CAR classifier, we achieve high concept accuracy by capturing the nonlinear concept distribution in the latent spaces. This approach provides accurate and fine-grained explanations for brain encoders, enhancing their transparency and reliability for future applications. Additionally, our framework could automatically identify ROIs in human brains by applying feature attribution on relevant concept density functions, offering novel perspectives for neuroscience research.

However, the BAR framework inherits certain limitations from concept-based explanations. It relies on ground truth concept labels in datasets, which may not always be available and often require significant annotation effort. While our BAR framework can reveal certain concepts learned by brain encoders, it does not comprehensively capture all meaningful features to fully explain model predictions. As a result, relying solely on BAR explanations may overlook critical factors in brain encoders and provide an incomplete picture.

We believe our contribution opens up multiple promising directions for future neuroscience research. Given the flexibility and efficiency of the BAR framework, it can be applied to a wider range of concepts and brain encoders to explain model predictions. To comprehensively study brain functions, future work could explore feature attributions beyond the posterior cortex (`nsdgeneral`), identifying additional ROIs sensitive to visual and semantic concepts. The BAR framework could also be extended to other brain signal modalities, such as electroencephalography (EEG) and magnetoencephalography (MEG), enabling broader applications in cognitive and perceptual studies.

5. Conclusion

In this paper, we propose the Brain Activation Region (BAR) framework to investigate concepts learned by brain encoders and discover potential brain functions. By training CAR classifiers in latent spaces of brain encoders, we achieve high concept accuracy across most COCO image categories, effectively explaining semantic representations learned by both encoders. To further explore brain voxels associated with concept learning, we apply Integrated Gradient on concept density function, identifying two potential brain regions contributing to visual concept processing. By introducing concept-based explanations for brain encoders, we believe our contribution could enhance the transparency of NeuroAI applications and advance our understanding in cognitive neuroscience.

References

[1] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, and others. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022. 1, 2

[2] Jonathan Crabbé and Mihaela van der Schaar. Concept activation regions: A generalized framework for concept-based explanations. In *NeurIPS*, pages 2590–2607, 2022. 1, 2

[3] Robert Desimone, Thomas D Albright, Charles G Gross, and Charles Bruce. Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, 4 (8):2051–2062, 1984. 3

[4] Fang Jiang, Laurence Dricot, Jochen Weber, Giulia Righi, Michael J Tarr, Rainer Goebel, and Bruno Rossion. Face categorization in visual scenes may start in a higher order area of the right fusiform gyrus: evidence from dynamic visual stimulation in neuroimaging. *Journal of Neurophysiology*, 106(5):2720–2736, 2011. Publisher: American Physiological Society Bethesda, MD. 4

[5] Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11):4302–4311, 1997. 3

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 1, 2

[7] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, 2017. 1

[8] Yongqiang Ma, Yulong Liu, Liangjun Chen, Guibo Zhu, Badong Chen, and Nanning Zheng. BrainCLIP: Brain Representation via CLIP for Generic Natural Visual Stimulus Decoding. *TMI*, 2025. 1

[9] Aina Puce, Truett Allison, Maryam Asgari, John C Gore, and Gregory McCarthy. Differential sensitivity of human visual cortex to faces, letterstrings, and textures: a functional magnetic resonance imaging study. *Journal of neuroscience*, 16 (16):5205–5215, 1996. 3

[10] Anna W Roe, Leonardo Chelazzi, Charles E Connor, Bevil R Conway, Ichiro Fujita, Jack L Gallant, Haidong Lu, and Wim Vanduffel. Toward a unified theory of visual area V4. *Neuron*, 74(1):12–29, 2012. 4

[11] Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, and others. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. In *NeurIPS*, pages 24705–24728, 2023. 1, 2, 3

[12] Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, and Tanishq Mathew Abraham. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. In *ICML*, 2024. 1

[13] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, pages 3319–3328, 2017. 1, 3

[14] Lucina Q Uddin. Salience processing and insular cortical function and dysfunction. *Nature reviews neuroscience*, 16 (1):55–61, 2015. Publisher: Nature Publishing Group UK London. 3

[15] Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. Umbrae: Unified multimodal brain decoding. In *ECCV*, pages 242–259, 2024. 1, 2, 3

[16] Weihao Xia, Raoul de Charette, Cengiz Öztireli, and Jing-Hao Xue. Dream: Visual decoding from reversing human visual system. In *WACV*, pages 8226–8235, 2024. 1

[17] Weihao Xia and Cengiz Öztireli. Mevox: Multi-task vision experts for brain captioning. In *CVPR Workshop*, 2025. 1

[18] Paul A Yushkevich, Yang Gao, and Guido Gerig. ITK-SNAP: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. In *EMBC*, pages 3342–3345, 2016. 2, 3

[19] Elisa Zamboni, Isaac Watson, Rüdiger Stirnberg, Laurentius Huber, Elia Formisano, Rainer Goebel, Aneurin J Kennerley, and Antony B Morland. Mapping curvature domains in human V4 using CBV-sensitive layer-fMRI at 3T. *Frontiers in Neuroscience*, 19:1537026, 2025. 4

BAR: Probing Brain Encoders with Concept-Based Explanations

Supplementary Material

S1. Experimental Setup

Dataset We conduct our experiments on the Natural Scene Dataset (NSD) [1], which consists of high-resolution, whole-brain 7T fMRI signals collected when human participants view colour natural scenes. These natural scene images are selected from the Microsoft Common Objects in Context (MS-COCO) image dataset [6].

Brain fMRI Signal To align with MindEye [11] and UMBRAE [15] encoders, we use brain responses from the `nsdgeneral` regions of subjects 1, 2, 5, and 7. These regions contain voxels located in the posterior cortex that are particularly responsive to the NSD stimuli [1].

Brain Encoders We apply Concept Activation Regions (CAR) [2] to interpret two representative pretrained brain encoders: the `voxel2clip_cls` retrieval model from MindEye [11] and the `brainx-v1.4` model from UMBRAE [15]. For MindEye, we load subject-specific pretrained weights to obtain embeddings for concept-positive and concept-negative samples. For UMBRAE, we utilize cross-subject training weights, which are designed to generalize across multiple individuals. To ensure compatibility with CAR, we average the first 256 dimensions of each embedding, thus standardizing the input for further analysis.

Concept Generation For each subject, we define concepts based on MS-COCO categories that have at least 200 unique image-voxel pairs. As a result, we create 47 concepts to describe the semantic information from the input. With $k = 200$, we generate positive and negative voxel samples for each concept and feed them into the brain encoders.

CAR Training Details Following the official CAR implementation [2], we train a support vector classifier with Gaussian RBF kernel for each concept and subject. All models are trained using an NVIDIA RTX 4070 Ti SUPER GPU with default hyperparameters from `scikit-learn`.

S2. Concept Accuracies

We group the 47 concepts into five semantic categories, as shown in Tab. S1. The classification accuracies for individual concepts are presented in Fig. S1 for MindEye embeddings [11] and in Fig. S2 for UMBRAE embeddings [15].

Table S1. **Concept-Category Correspondence.** The categories correspond to the 80 classes from COCO, illustrating the mapping between extracted concepts and their respective object categories.

Category	Concepts
Creatures	bird, cat, dog, giraffe, horse, person
Household Items	bed, bench, bowl, chair, clock, couch, cup, dining table, fork, knife, oven, potted plant, sink, spoon, toilet, tv, vase
Transportation	airplane, bicycle, boat, bus, car, motorcycle, traffic light, train, truck
Food	cake, pizza
Everyday Objects	backpack, bottle, book, cell phone, handbag, laptop, skateboard, skis, sports ball, surfboard, tennis racket, tie, umbrella

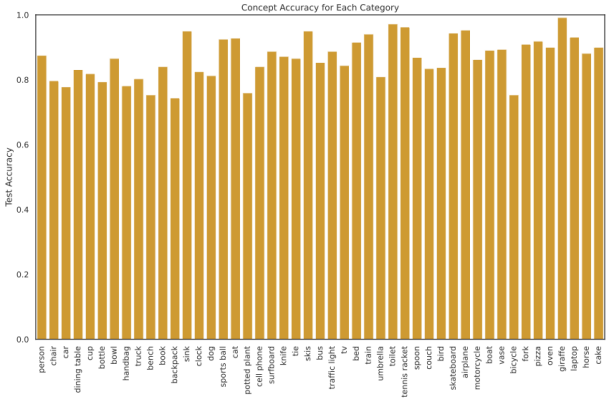

Figure S1. **Concept Accuracies for MindEye Embeddings.** Zoom in for details.

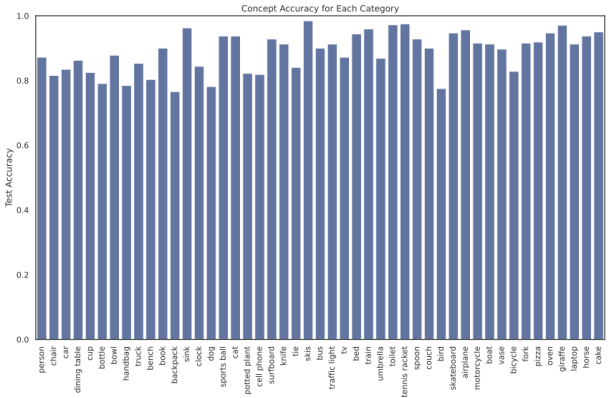

Figure S2. **Concept Accuracies for UMBRAE Embeddings.** Zoom in for details.