FROM ACTIVATION TO SPECIFICITY: AUTOMATING COUNTERFACTUAL TESTING OF VISUAL REPRESENTATIONS IN THE HUMAN BRAIN

Anonymous authors
Paper under double-blind review

ABSTRACT

Identifying which brain regions represent a visual concept in the human brain is a central neuroscience challenge. Existing approaches have localized coarse functional regions (e.g., faces, places) through activation maximization, identifying regions that activate strongly for a target concept relative to others. Yet strong activation alone does not establish that a region represents the concept itself, as responses may instead be driven by correlated visual or semantic cues. We introduce BrainTRACE (Testing Representations through Counterfactual Evidence), an automated framework that combines generative and brain models to synthesize controlled stimuli and test candidate neural representations through targeted counterfactual-specificity testing. Given a query specifying a concept, our framework constructs targeted stimulus sets comprising concept images, counterfactual edits that remove the target concept while preserving other image content, and images with correlated distractors. It then uses an image-to-fMRI encoding model to predict brain responses and searches for representations that respond specifically to the target concept over correlated alternatives. BrainTRACE returns candidate representations supported by counterfactual-specificity evidence and proposes follow-up fMRI experiments to further test or extend these findings. Our approach recovers known functional localizations and identifies new candidate representations across dozens of concepts, supported by both predicted and measured fMRI data. Critically, we show that without counterfactual evaluation, a large fraction of localizations would be identified as false positives by our counterfactual-specificity testing, suggesting that activation alone is insufficient evidence of representation.

1 INTRODUCTION

The human brain organizes visual experience into representations of objects, scenes, actions, and semantic structures. Understanding how visual concepts are represented is therefore a central goal of neuroscience. Over decades, functional magnetic resonance imaging (fMRI) studies have revealed aspects of visual organization, from low-level visual maps (Sereno et al., 1995; Engel et al., 1997) to general category-selective regions (Kanwisher et al., 1997; Epstein & Kanwisher, 1998; Downing et al., 2001; Cohen et al., 2000), while recent computational approaches discover richer representations (Bao et al., 2025; Hwang et al., 2025; Luo et al., 2024). Central to this progress is category-contrast analysis (Kanwisher et al., 1997; Epstein & Kanwisher, 1998; Downing et al., 2001; Cohen et al., 2000), asking whether a brain region or voxel responds more strongly to a target category than to the average response across other categories. Such analyses revealed regions selective for concepts such as faces, bodies, and words. Image-to-fMRI encoders have since extended this paradigm by predicting brain activity for images without fMRI, allowing the use of more data and revealing more nuanced representations (Bao et al., 2025; Wasserman et al., 2025a).

Despite this progress, important limitations remain. Many methods do not test whether a strongly responding voxel or region truly represents the target concept, or whether its response is instead driven by correlated visual or semantic cues co-occurring with the concept, such as color, background, pose, or the presence of other objects. Without such counterfactual-specificity testing, many discovered localizations may in fact be false positives. Beyond identifying candidate representations, current

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

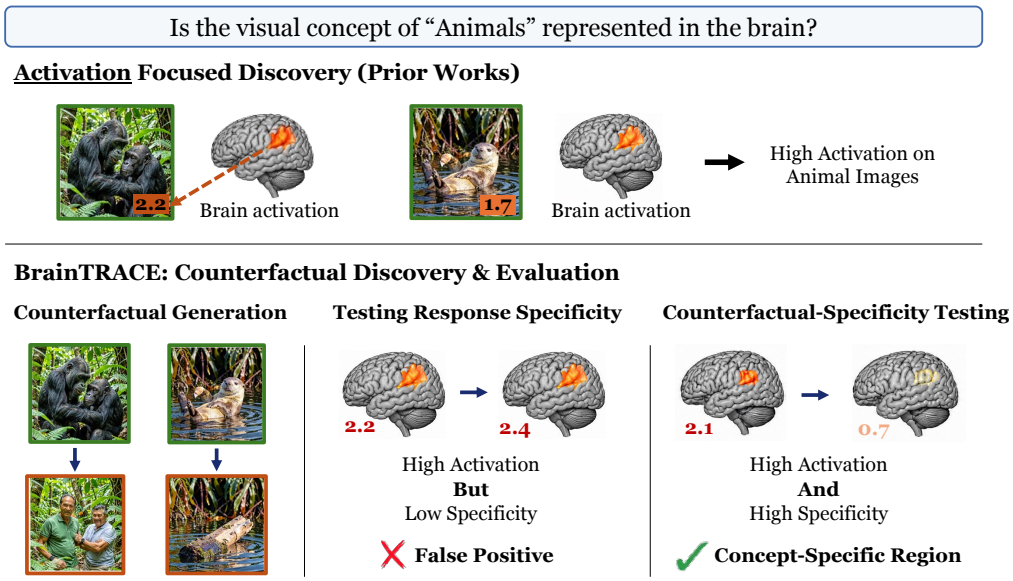

Figure 1: **Overview of our approach.** Activation-maximization based methods (top) identify regions with high responses to target concept, but cannot distinguish true concept representations from correlated cues. BrainTRACE (bottom) performs counterfactual-specificity evaluation using targeted counterfactual stimuli isolating the concept. Regions with high activation but low response difference between original and counterfactual images are identified as false positives, while regions with high activation and strong specificity response are identified as candidate concept-specific representations.

methods provide limited guidance about which follow-up fMRI experiments are worth running, which concepts are underrepresented in existing datasets, and which new stimuli would be most informative.

To address these limitations, we introduce *BrainTRACE*, an automated framework for discovering and evaluating visual concept representations through counterfactual-specificity testing. Given a target concept, BrainTRACE constructs targeted stimulus sets designed to isolate the concept from correlated visual and semantic factors. It then uses an image-to-fMRI model to predict brain responses and identifies voxels and regions selective to the concept rather than to co-occurring properties (Fig. 1). When evidence is insufficient, it proposes follow-up experiments to further test its discoveries.

More specifically, the targeted image–fMRI dataset contains three types of images: ***(i) Positive Images***, depicting the target concept, obtained through generation and retrieval from large image pools. ***(ii) Semantically Negative Images***, which depict correlated but distinct concepts proposed by a large language model that identifies concepts likely to co-occur with the target but not be the concept itself. For *human hands*, it may propose alternatives such as *human body*, *human legs*, or *robot hands*. ***(iii) Counterfactual Negative Images*** are created by editing positive images to remove or replace the target concept while preserving the rest of the image content, helping rule out explanations based on background, color, object position, or other co-occurring properties. Generation is particularly useful for providing clear and diverse instances, especially for concepts that are rare or underrepresented in existing datasets. Importantly, while discovery uses model-predicted responses to generated stimuli, when available, BrainTRACE independently evaluates discovered representations using measured fMRI responses to ***retrieved positive and semantically negative images***.

Together, this enables BrainTRACE to characterize representations in terms of whether a candidate representation responds specifically to the target concept under the tested alternatives, where it is represented, and with what level of confidence. It further assesses how well the concept is supported by the measured data and whether this evidence substantiates the finding. When evidence is insufficient, it proposes stimuli for follow-up fMRI experiments, including positive examples, semantic negatives, and counterfactual edits.

Using BrainTRACE, we study 260 visual concepts. We first show that many representations identified by activation-based methods are false positives under counterfactual-specificity evaluation, with over 70% failing to exhibit concept-specific responses under counterfactual-specificity evaluation,

despite achieving high activation scores. Using specificity scores, BrainTRACE then identifies which concepts are robustly represented and where, evaluated on both predicted and measured fMRI data. We validate the framework against known brain regions, showing it recovers areas selective for faces, bodies, places, and words, before turning to a broader range of representations supported by counterfactual-specificity evidence, including different body parts such as hands and legs; text-related concepts such as handwritten text, traffic signs, and logos; and additional concepts such as animal faces, food, tools, and social interactions.

By moving beyond activation-based analyses toward targeted counterfactual-specificity testing, BrainTRACE not only reveals that many previously identified representations may be driven by correlated factors rather than the concepts themselves, but provides a systematic approach for discovering robust, concept-specific representations. Through large-scale evaluation across hundreds of concepts, we demonstrate that counterfactual-specificity testing leads to more reliable and interpretable mappings between visual concepts and brain activity. Beyond analysis, BrainTRACE closes the loop with experimental design, automatically identifying gaps in existing data and proposing informative stimuli for follow-up studies. In doing so, it offers a unified framework for discovery, validation, and refinement of visual representations in the brain.

2 Related Works

A foundational goal of visual neuroscience is to understand where and how the brain represents the visual world. Early fMRI studies revealed that visual cortex is organized retinotopically and encodes basic image properties such as orientation, spatial frequency, color, and motion (Sereno et al., 1995; Kamitani & Tong, 2005; Henriksson et al., 2008; Conway et al., 2007; Tootell et al., 1995). While these low-level properties are easier to systematically manipulate and disentangle, the space of higher-level visual concepts is vastly larger and less structured. At this higher level, researchers presented subjects with images from a small set of carefully chosen, labeled categories and identified regions responding selectively to each, revealing broad category-selective regions in higher visual cortex for faces (Kanwisher & Yovel, 2006), scenes (Epstein & Baker, 2019), bodies (Downing et al., 2001), and written words (Cohen et al., 2000). While highly influential, this approach was inherently limited to a small, pre-specified set of broad concepts.

More recent work has sought to move beyond these narrow category sets. Advances in brain modeling (Wasserman et al., 2026; Beliy et al., 2025; Ratan Murty et al., 2021), and in particular encoding models that predict voxel-wise fMRI responses from image features (Kay et al., 2008; Beliy et al., 2024; Adeli et al., 2025), have enabled concept analysis at scale by predicting responses to images never measured in the scanner. One line of work uses such models to characterize already-defined regions by retrieving images from large datasets that most strongly activate a given region (Hwang et al., 2025), or by generating images that maximize its activation (Gu et al., 2022; Luo et al., 2023). Other recent approaches, such as MindSimulator (Bao et al., 2025), use concept-oriented image retrieval together with predicted fMRI responses to localize candidate concept-selective voxels. Yet a central limitation remains: a region that responds strongly to images of a concept may still be driven by correlated visual or semantic properties rather than by the concept itself.

Recent advances in brain modeling together with machine learning offer a new path forward. Modern generative models can synthesize and edit images with fine-grained semantic control (Rombach et al., 2022; Brooks et al., 2023; Wasserman et al., 2025b; Labs et al., 2025), while large language models can help propose correlated counter-hypotheses and reason about informative interventions (Kiciman et al., 2023; Alkan et al., 2025; Zhou et al., 2024). Combined with image-to-fMRI encoding models, these capabilities enable controlled stimulus sets that disentangle concept-driven activation from correlated factors, and to predict the corresponding fMRI responses. We propose BrainTRACE, which leverages these tools to address the specificity gap in activation-based visual concept representation discovery and to guide future experiments when existing data is insufficient for proper validation.

3 The BrainTRACE Framework

BrainTRACE takes as input a target concept together with a subject's image–fMRI dataset, and outputs a candidate voxel set representing that concept along with a confidence estimate. We automatically constructed a list of 260 concepts using ChatGPT (GPT-5) (OpenAI, 2025), selected to provide broad coverage of concepts, including established neuroscience concepts, concepts studied in recent representation-discovery work, and additional general and fine-grained concepts.

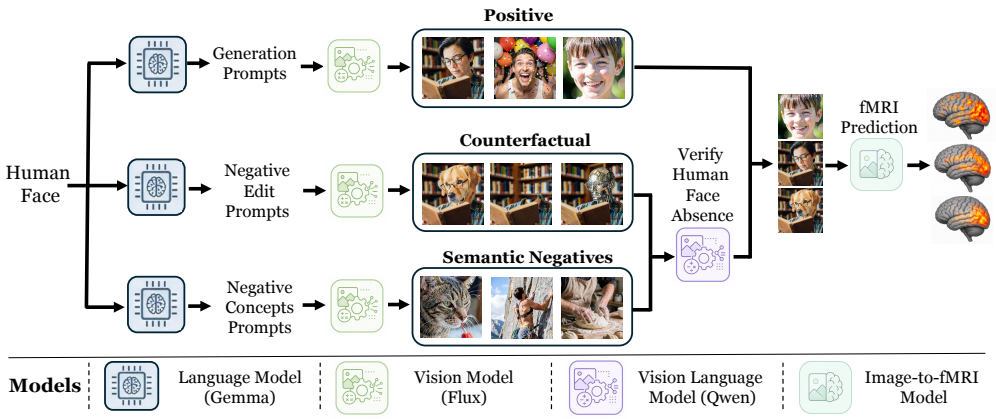

Figure 2: **Concept-Targeted Counterfactual Stimulus Generation.** Given a target concept (e.g., human face), BrainTRACE constructs a targeted stimulus set consisting of three types of stimuli: positive images, counterfactuals, and semantic negatives. A language model generates prompts for each type, which are used by a text-to-image model to synthesize images. Counterfactual and semantic negatives are designed to isolate the target concept from correlated visual and semantic factors. A vision-language model verifies the presence or absence of the target concept in each image. Finally, all images are passed through an image-to-fMRI model to obtain predicted neural responses.

Each image is associated with an fMRI response vector, where each entry corresponds to the activation of a voxel in the cortex ($\sim$ 40K voxels in our data). The pipeline proceeds in three stages: ***(i) targeted stimulus set generation***. Given a target concept, BrainTRACE constructs a targeted stimulus set designed to isolate the target concept from correlated visual and semantic factors. These include: positive images of the concept, semantic negatives depicting correlated but distinct concepts, and counterfactual negatives in which the target concept is removed or replaced while the rest of the image is preserved as much as possible. In addition, it retrieves positive and semantic-negative images from large image pools and, when available, from the measured image–fMRI dataset itself. Images without measured fMRI are passed through an image-to-fMRI encoder to obtain predicted responses. ***(ii) Concept-selective representation search.*** BrainTRACE then computes activation and specificity scores for each voxel based on the positive, semantic-negative, and counterfactual image sets, and uses these scores to identify a candidate voxel set with strong and concept-specific responses. ***(iii) Validation and Follow-Up Experiment Design.*** BrainTRACE then validates this candidate set using held-out predicted and measured fMRI data, quantifies how well the concept is represented in the measured dataset, and combines these signals into a final verdict. When measured evidence is insufficient, BrainTRACE analyzes which evidence is missing and proposes informative follow-up stimuli for future fMRI experiments.

3.1 Concept-Targeted Stimulus Set Generation

In this stage, BrainTRACE constructs a dedicated image–fMRI dataset for the target concept, designed to localize candidate representations while testing their specificity under counterfactual and semantic-negative comparisons. Starting from the target concept, it generates three types of image stimuli, highlighted in Fig. 2. ***(i) Positive Images:*** A large language model (Gemma-3-27B-IT (Gemma Team & Google DeepMind, 2025)) produces diverse prompts representing the target concept, aiming to capture a wide range of visual variations. These prompts are passed to a text-to-image model (FLUX.2 (Labs, 2025)) to synthesize one image per prompt. For each concept, BrainTRACE generates 200 positive images for training and 100 additional held-out images for validation. ***(ii) Semantic Negative Images:*** The language model also proposes a set of 10 *counter concepts* — that are semantically or visually related to the target concept, and may co-occur with it, but do not require the target concept itself to appear. For example, for the target concept *person surfing*, proposed counter concepts may include *a man fishing*, *beach*, or *wave*. For each counter concept, the language model generates 10 prompts while explicitly checking that the target concept is not mentioned in any of them. The text-to-image model synthesizes the corresponding images. An Image–Concept Verification step using a vision-language model (Qwen3-VL-8B (Team, 2025)) is then applied to verify that the target concept is indeed absent from the images. After filtering, this stage yields approximately 80–100 semantic negative images for training and a similar number for validation. ***(iii)***

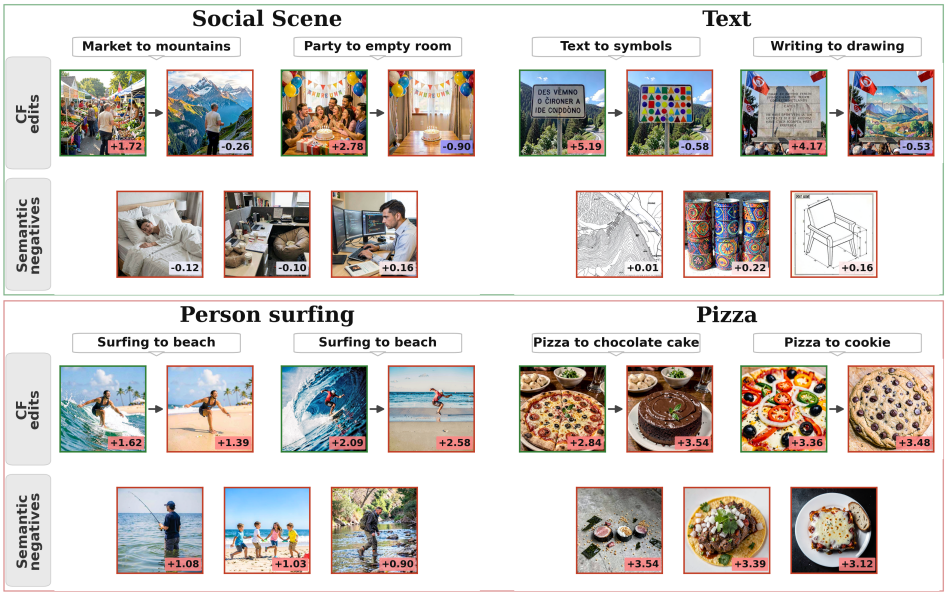

Figure 3: **Counterfactual-specificity evaluation of discovered regions.** Top: regions discovered by BrainTRACE, showing high activation for positives and lower activation for counterfactual edits and semantic negatives. Bottom: regions discovered by an activation-based method, which remain highly activated after edits or for related negatives, indicating false positives driven by correlated cues.

Counterfactual Negative Images: For each selected positive image, the language model proposes 10 edit instructions designed to minimally modify the image so that it no longer expresses the target concept. For example, for the concept *human face*, the edits may replace the human face with an animal face or remove the face entirely. These prompts are passed to an image editing model (FLUX.2), and the resulting images are verified. This process is applied to 50 training images and 20 validation images, yielding approximately 400–500 counterfactual negatives. Finally, all generated images are passed through an image-to-fMRI encoder to predict brain responses.

In addition to generated stimuli, BrainTRACE retrieves positive and semantic-negative images from the measured image–fMRI dataset using a CLIP-based retrieval strategy (Radford et al., 2021) inspired by Bao et al. (2025). Retrieved images are filtered to ensure the target concept is present in positives, absent from the negatives, and that negatives align with the intended alternative concept. This stage enables evaluation on measured data for concepts and counter concepts that are sufficiently represented in the dataset. Measured-data evaluation uses held-out images not used for voxel selection.

3.2 Concept-Selective Representation Search

Given the concept-targeted stimulus set, BrainTRACE assigns each voxel three scores with respect to the target concept. The first is an ***activation score***, measuring how strongly the voxel responds on average to the positive images. The second calculates the difference between the activation score and the hardest semantic negatives, defined as the 10 negative images that most strongly activate the voxel. The third calculates the difference between each positive image and its hardest edited counterpart, using the edit that produces the highest activation for that voxel. Intuitively, the activation score tests whether a voxel responds strongly to the concept, while the two ***specificity scores*** test whether this response remains higher than semantically related alternatives and controlled counterfactual edits.

BrainTRACE then uses the specificity scores for each voxel as the average of the semantic-negative and counterfactual scores (see F.4 for the detailed ranking score). For a given concept, the candidate representation is constructed as the set of all voxels with positive specificity scores or a predefined number of voxels ranked the highest. This voxel set serves as the candidate region representing the concept. BrainTRACE can then evaluate the region as a whole by averaging activation and specificity scores across all voxels in the set. This way, the discovered region is required not only to respond strongly to positive images, but also to respond more strongly to them than to semantic negatives and counterfactual edits. The specificity score of the discovered region on the training set is used to determine whether it is a strong candidate representation, while final quantitative evaluation is carried out on a separate evaluation split and on measured fMRI data.

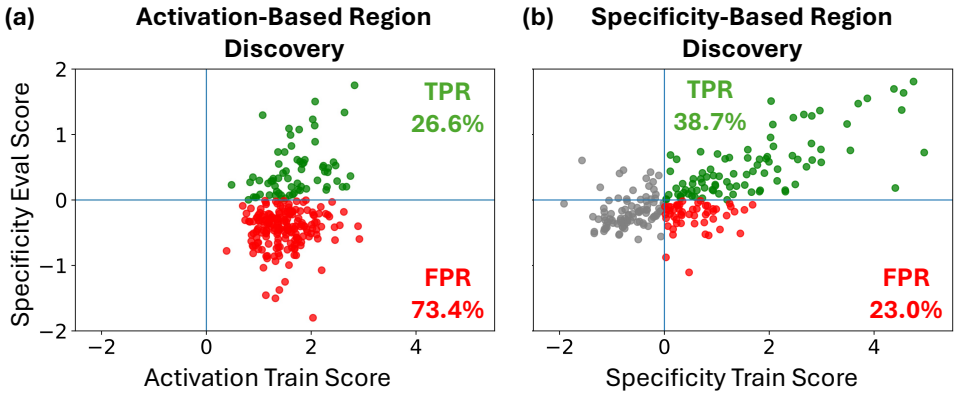

Figure 4: **Specificity-based ranking reduces false discoveries.** (a) Regions selected using activation-based ranking. (b) Regions selected using specificity-based ranking. Activation-based discovery frequently selects regions that respond strongly to the target concept but do not exhibit specificity under the tested alternatives, leading to many specificity false positives. By ranking candidates using specificity score, BrainTRACE suppresses correlation-driven discoveries, reduces false positives, and recovers more faithful concept representations.

3.3 Final Verdict and Follow-Up Experiment Design

The final verdict in BrainTRACE is determined by two main quantities. The first is the ***specificity evidence*** for the discovered candidate representation, measured by its specificity scores on the generated evaluation dataset and on the measured-data evaluation. High scores on both indicate stronger evidence that the voxel set responds selectively to the target concept rather than to correlated alternatives. ***Confidence Score*** is measured using statistical tests on our proposed scores, yielding for each concept representation a p-value for each score (see Appendix D.1).

The second is the ***concept coverage in the measured data***, estimated by analyzing what fraction of the desired positive and semantic-negative images can be successfully retrieved and verified in the measured image–fMRI dataset (see Appendix D.2). This quantity reflects how informative the measured data is for validating the concept. For example, if only a few measured images contain dogs, then even a low or high measured-data evaluation score is less informative, since the dataset provides only limited evidence about that concept. Based on the specificity evidence and the concept coverage in the measured data, the framework returns a final decision together with the candidate representation and, when relevant, a set of informative images for follow-up experiments. When concept coverage in the measured data is high, the measured-data evaluation is considered informative. In this case, strong and statistically significant specificity evidence leads to a high-confidence discovery, while weak specificity evidence leads to rejection. In contrast, when concept coverage in the measured data is low, the measured-data evaluation is less informative. In this case, strong specificity evidence on the generated dataset is treated as a promising but insufficiently validated finding and triggers follow-up experimental recommendations, whereas weak specificity evidence is treated as inconclusive.

4 Results

BrainTRACE supports both counterfactual-specificity evaluation of candidate regions proposed by existing methods and discovery of new regions that more faithfully represent target concepts. We evaluate on the *Natural Scenes Dataset* (NSD) (Allen et al., 2022), a large-scale 7-Tesla fMRI dataset consisting of 8 subjects each viewing approximately 10,000 natural images, reporting results on the 4 subjects who completed all sessions.

4.1 BrainTRACE Counterfactual-Specificity Testing

Throughout this section we compare four region discovery methods: (i) **Max Activation** ranks voxels by their average activation on generated positive images, following the standard localization approach. (ii) **MindSimulator** (Bao et al., 2025) retrieves images from COCO using CLIP scores and ranks voxels by predicted fMRI response. For a fair comparison we use the same image–fMRI encoder as in BrainTRACE and the same number of overall tested images. We also compare (iii) **MindSimulator+VLM**, which extends MindSimulator by using our filtering procedure with a

Table 1: **Quantitative comparison across discovery methods.** Average activation and specificity scores for regions discovered for the top 50 concepts, across four subjects. BrainTRACE preserves high activation while improving specificity.

	Activation		Specificity		
Method	**Gen.**	**Meas.**	**Gen.**	**Meas.**	**Edits**
Max Activation	**2.76**	0.70	0.08	0.18	0.44
MindSimulator	1.89	1.02	-0.44	0.27	0.23
MindSimulator+	2.13	**1.12**	-0.26	0.41	0.38
BrainTRACE	2.05	1.08	**0.62**	**0.71**	**0.98**

Table 2: **BrainTRACE aligns with known functional regions.** Fraction of top-ranked voxels fall within the expected functional regions for broad visual categories show high alignment with established functional organization.

	Voxel Alignment Accuracy		
Region	Top 100	Top 200	Top 500
Bodies	99%	99%	97%
Faces	90%	87%	84%
Places	74%	75%	74%
Words	99%	98%	97%

vision-language model, to filter retrieved images that do not clearly depict the target concept. Finally, our method (iv) **BrainTRACE** ranks voxels using a score combining both activation and specificity. For all methods, for the quantitative evaluation the discovered region per concept consists of the top 100 voxels ranked by the method score; results for other region sizes are provided in Appendix B.3.

BrainTRACE Evaluates Region Specificity. Fig. 3 shows examples of our method's counterfactual-specificity evaluation on two regions supported by specificity evaluation (top) and two regions that fail specificity evaluation (bottom). For Social Scene and Text, not only is the positive activation high, but the semantic negatives and counterfactual edits show much lower activation. The counterfactual edits are particularly strong: for example, for Text, signs containing text are replaced by the same sign and background but with shapes or paint instead of text, while semantic negatives include visually similar but distinct concepts such as sketches. In the bottom rows, regions discovered by activation-based methods show high activation for positives but similarly high activation for negatives. Surfing, for example, was presented as a discovered representation in MindSimulator, but counterfactual-specificity testing reveals that the region discovered by this method likely responds to correlated factors such as water, human presence, or body pose rather than to surfing itself.

Activation-Based Discovery Has High False Positive Rate. Fig. 4a shows, for each of 260 concepts and subject 1, the activation score and specificity score of the region discovered by MindSimulator approach. Each point represents one concept, with the activation score computed on the train set and the specificity score on the evaluation set. Concepts with a positive activation score are considered discovered by the method. We color in green those that are also supported by specificity evaluation (true positives, TPR) and in red those with high activation but negative specificity score (specificity false positives, FPR). Here, false positives are defined relative to our held-out counterfactual-specificity criterion, as no external ground-truth labels of concept-specific neural representations are available. Activation-based discovery yields a specificity false positive rate of nearly 70%, meaning that most discovered regions fail counterfactual-specificity evaluation, suggesting that their activation may be driven by correlated factors rather than the target concept.

Specificity-Based Ranking Discovers More Faithful Regions. Fig. 4b shows the same plot but with regions chosen according to the train specificity score. Concepts with a negative train specificity score are labeled as not supported by specificity evaluation and withheld from discovery (grey points), avoiding false positives. This reduces the false positive rate from 73.4% to 23%, while improving the true positive rate from 26.6% to 38.7%. Together, this shows that specificity-based ranking not only avoids spurious localizations but also recovers more genuinely concept-specific regions than activation-based methods.

Regions Identified by Specificity-Based Ranking Maintain High Activation While Improving Specificity Scores. Table 1 compares all four methods across activation and specificity scores. For each method we select the top 50 concepts by train score and report results on regions of 100 voxels (other region sizes in the appendix). Activation is measured as the average response to positive images, either generated by our pipeline and held-out (Gen.) or retrieved from the held-out measured image–fMRI pool (Meas.). Specificity is measured via semantic negative scores (on both measured and generated data) and counterfactual edit scores. Activation-based methods achieve high

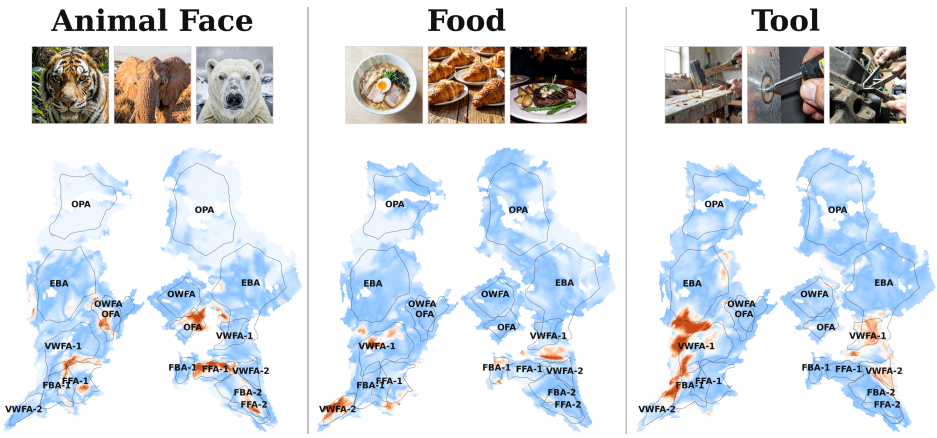

Figure 5: **Concepts discovered by counterfactual-specificity evaluation.** We show voxel-wise specificity scores on brain maps for three example concepts, with positive images above each map. Each panel shows a flatmap of high-level visual cortex for subject 1. Each voxel is colored by its specificity score, where warmer colors indicate higher concept-specific specificity evidence. Black outlines and labels mark NSD functional ROIs, allowing comparison with known visual regions.

activation scores but show notably low semantic negative scores, with other methods scoring below zero, indicating non-concept-specific regions. BrainTRACE achieves comparable activation scores while substantially improving specificity: for example, the semantic negative score on generated data improves from -0.44 (MindSimulator) to 0.62 (BrainTRACE), and on measured data from 0.27 to 0.71. This confirms that BrainTRACE discovers regions that are both strongly activated by the target concept and specific to it.

Specificity scores are reliable with respect to measured data and consistent across models used at different stages of the pipeline. We show that predicted responses closely match measured fMRI responses on the same NSD images ($r = 0.93{-}0.96$ for positive responses; $r = 0.72{-}0.80$ for semantic specificity). To show that our results are not driven by biases in any particular pipeline component, we further demonstrate that replacing the image-to-fMRI encoder, language model, or image generation/editing model yields highly consistent specificity maps ($r = 0.81{-}0.87$, cosine similarity ≈ 0.98). Full analyses are provided in Appendix C.

4.2 Fine-Grained Visual Concept Localization

Alignment with Known Functional Regions. To validate our method, we examine whether regions identified through counterfactual-specificity testing align with well-established functional areas, focusing on four broad categories: Bodies, Faces, Places, and Words. For each category, we select relevant concepts from our set, rank voxels by specificity score, and measure the fraction of top K voxels that fall within corresponding NSD-localized functional region. Table 2 shows results for voxel set sizes of 100, 200, and 500, demonstrating strong alignment across all four categories.

Discovery of Fine-Grained Concept Representations. Fig. 5 shows flatmaps of high-level visual cortex in which each voxel is colored by its specificity score, from low (blue) to high (red). Despite imposing no spatial constraints, the discovered regions are spatially localized, whereas activation-based methods produce many more high-scoring voxels (see Appendix B.1), reflecting the greater selectivity of specificity-based scoring. As examples, the representation for *tools* localizes near body-part and action-related regions such as EBA, while the representation for *animal faces* localizes within known face-selective areas such as FFA and OFA. More broadly, BrainTRACE discovers localized representations supported by specificity testing for many different concepts (see Appendix E for additional examples). Fig. 6 further shows binary maps of voxels with positive specificity scores for sets of related concepts, revealing fine-grained distinctions between them: body-part concepts such as *human face*, *human hands*, and *human legs* divide activity across face areas (FFA) and body-related regions (EBA and FBA), while text-related concepts such as *handwritten text*, *symbolic signs*, and *logos* show distinct patterns within VWFA- and OWFA-related regions.

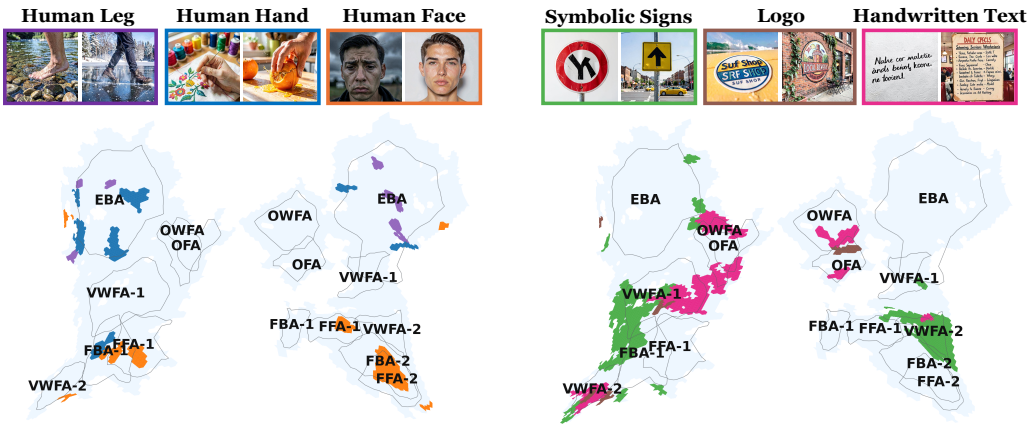

Figure 6: **Fine-grained organization of related concepts.** Left: body-related concepts, including human face, human hand, and human leg, show distinct voxel patterns across face- and body-selective regions. Right: text-related concepts, including handwritten text, symbolic signs, and logos, show distinct voxel patterns across word- and object-related visual areas. These results show that BrainTRACE discovers nearby semantic categories within high-level visual cortex.

Consistency Across Subjects. Appendix A.1 shows brain maps for the same concepts across subjects with quantitative results in Appendix A.2. Despite the well-known variability of functional organization across individuals, we observe clear correspondence in the region locations where concepts are represented. This both strengthens the validity of our findings and suggests that the discovered concept-specific representations capture robust aspects of visual cortex organization.

4.3 Ablation and Analysis

Appendix B provides additional analyses of our method. **(i) Ranking scores.** We examine different activation- and specificity-based ranking scores (B.2). **(ii) Region size.** We analyze the effect of the number of voxels chosen to define a concept region, showing that specificity scores gradually decrease as region size increases (B.3). **(iii) Generated-stimuli reliability.** We evaluate the accuracy of the automated stimulus-verification procedure using human annotations, showing close agreement with human judgments, with 94.6% precision and 90.6% recall (B.4). **(iv) Low-level visual confounds.** We test whether low-level visual changes explain the counterfactual-specificity scores and find that they account for only 6.2% of their variance (B.5). **(v) Failure cases.** We additionally analyze BrainTRACE false positives and find that many arise from broad visual properties or failures in semantic-negative generation (B.6).

5 Conclusions

A central message of this work is the importance of testing concept specificity in visual representation discovery in the brain. While the distinction between responses driven by correlated factors and truly concept-specific representations has been widely discussed in other fields, and in some areas of neuroscience, it has received more limited attention in the study of visual concept representations from fMRI. BrainTRACE takes a step toward addressing this gap by supporting counterfactual-specificity evaluation and enabling discovery of more specific and reliable concept representations, including fine-grained concepts beyond broad category-level regions. At the same time, BrainTRACE depends on current language and vision models and is therefore not free of failure modes. Errors in image generation, editing, retrieval, verification, or in the scope of the proposed counterfactuals may still leave relevant correlated factors untested, meaning that some discovered representations may remain specific only with respect to the alternatives considered by the current pipeline. We expect this limitation to improve as language and vision models continue to advance. An important direction for future work is to make the pipeline more iterative, allowing current results and activation patterns to guide the proposal of new counterfactuals and semantic negatives for stronger validation. More broadly, BrainTRACE, together with its analysis of which concepts and negative controls are sufficiently represented in measured data, provides a step toward counterfactual-specificity testing of brain representations and tighter integration between computational and experimental neuroscience.

AI USE DISCLOSURE

Generative AI models formed part of our experimental methodology, including the generation and evaluation of visual stimuli, as detailed in the paper. We also used generative AI to support code implementation and revision of the manuscript's language. The authors reviewed all AI-assisted materials and accept full responsibility for the work's final text, analyses, and conclusions.

REFERENCES

Hossein Adeli, Sun Minni, and Nikolaus Kriegeskorte. Transformer brain encoders explain human high-level visual responses. *arXiv preprint arXiv:2505.17329*, 2025.

Atilla Kaan Alkan, Shashwat Sourav, Maja Jablonska, Simone Astarita, Rishabh Chakrabarty, Nikhil Garuda, Pranav Khetarpal, Maciej Pióro, Dimitrios Tanoglidis, Kartheik G Iyer, et al. A survey on hypothesis generation for scientific discovery in the era of large language models. *arXiv preprint arXiv:2504.05496*, 2025.

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

Guangyin Bao, Qi Zhang, Zixuan Gong, Zhuojia Wu, and Duoqian Miao. Mindsimulator: Exploring brain concept localization via synthetic fmri. *arXiv preprint arXiv:2503.02351*, 2025.

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

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

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

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

Bevil R. Conway, Sebastian Moeller, and Doris Y. Tsao. Specialized color modules in macaque extrastriate cortex. *Neuron*, 56(3):560–573, 2007. doi: 10.1016/j.neuron.2007.10.008.

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

Stephen A. Engel, Gary H. Glover, and Brian A. Wandell. Retinotopic organization in human visual cortex and the spatial precision of functional mri. *Cerebral Cortex*, 7(2):181–192, 1997. doi: 10.1093/cercor/7.2.181.

Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 1998.

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

Gemma Team and Google DeepMind. Gemma 3. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org.

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

Linda Henriksson, Lauri Nurminen, Aapo Hyvärinen, and Simo Vanni. Spatial frequency tuning in human retinotopic visual areas. *Journal of Vision*, 8(10):5:1–13, 2008. doi: 10.1167/8.10.5.

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

Yukiyasu Kamitani and Frank Tong. Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8(5):679–685, 2005. doi: 10.1038/nn1444.

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

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*, 1997.

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

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2023.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2`, 2025.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL `https://arxiv.org/abs/2506.15742`.

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

Andrew F Luo, Jacob Yeung, Rushikesh Zawar, Shaurya Dewan, Margaret M Henderson, Leila Wehbe, and Michael J Tarr. Brain mapping with dense features: Grounding cortical semantic selectivity in natural images with vision transformers. *arXiv preprint arXiv:2410.05266*, 2024.

OpenAI. GPT-5 System Card. `https://cdn.openai.com/gpt-5-system-card.pdf`, 2025. Accessed: 2026-05-06.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

N Apurva Ratan Murty, Pouya Bashivan, Alex Abate, James J DiCarlo, and Nancy Kanwisher. Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature communications*, 12(1):5540, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Roger B. H. Tootell, John B. Reppas, Anders M. Dale, Rodney B. Look, Martin I. Sereno, Rafael Malach, Thomas J. Brady, and Bruce R. Rosen. Visual motion aftereffect in human cortical area mt revealed by functional magnetic resonance imaging. *Nature*, 375(6527):139–141, 1995. doi: 10.1038/375139a0.

Navve Wasserman, Matias Cosarinsky, Yuval Golbari, Aude Oliva, Antonio Torralba, Tamar Rott Shaham, and Michal Irani. Brainexplore: Large-scale discovery of interpretable visual representations in the human brain. *arXiv preprint arXiv:2512.08560*, 2025a.

Navve Wasserman, Noam Rotstein, Roy Ganz, and Ron Kimmel. Paint by inpaint: Learning to add image objects by removing them first. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18313–18324, 2025b.

Navve Wasserman, Roman Beliy, Roy Urbach, and Michal Irani. Functional brain-to-brain transformation without shared stimuli. *NeuroImage*, 327:121741, 2026.

Xinjie Zhang, Peng Zhang, Shicheng Zheng, Jinghao Guo, Zhaoyang Jia, Yifei Shen, Xun Guo, Yuxuan Luo, Jiahao Li, Wenxuan Xie, et al. Mage-flow: An efficient native-resolution foundation model for image generation and editing. *arXiv preprint arXiv:2607.19064*, 2026.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. Hypothesis generation with large language models. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pp. 117–139, 2024.

Appendix

A Across Subjects Analysis

A.1 Consistency Across Subjects

While different subjects show substantial variability, even in their responses to the same image, observing correspondence in the localization of discovered representations across subjects provides strong validation of the method. It is also of independent interest, as it suggests that aspects of visual organization are shared across subjects despite individual variability. To examine this, we compare voxel-wise specificity maps for three concepts—*Animal*, *Human Interaction* and *Hands in Action*—across Subjects 1 and 2. Both the cortical flatmaps and the marked functional regions, taken from NSD, differ across subjects, and the specificity maps themselves are not identical. Nevertheless, as shown in Fig. S1, for all three concepts the main representation regions show clear correspondence across subjects. This further supports the validity of the discovered regions and suggests that visual concept representations identified through counterfactual-specificity testing capture shared aspects of cortical organization.

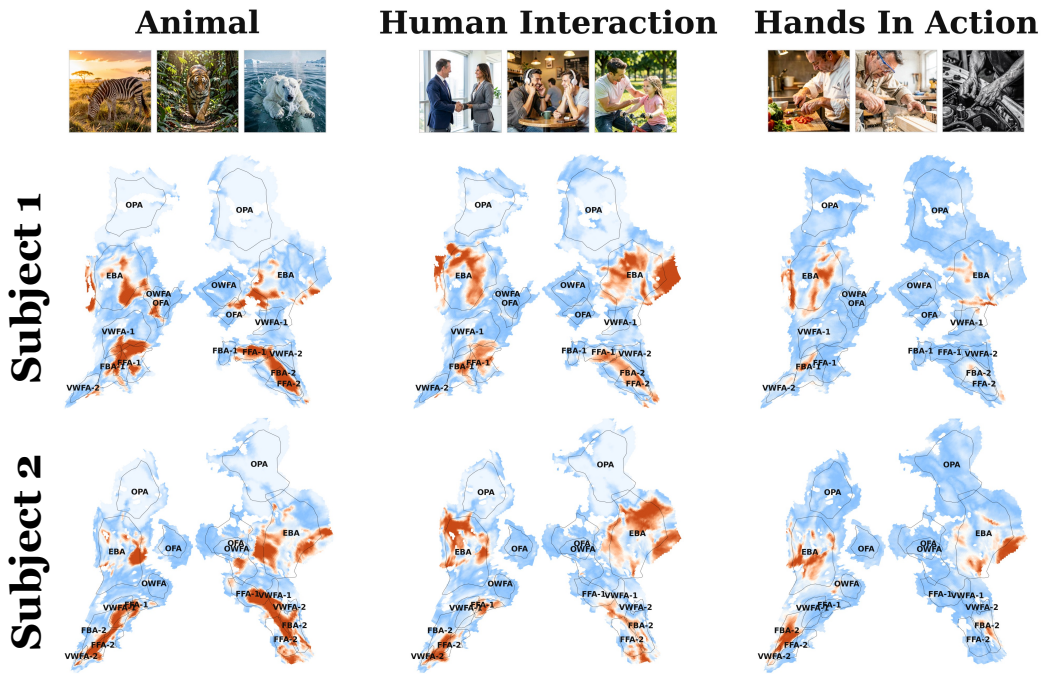

Figure S1: **Cross-subject consistency of concept representations identified through specificity testing.** Cortical flatmaps show BrainTRACE specificity scores for three representative concepts, Animal, Human Interaction and Hands in Action, across two NSD subjects. Example positive images for each concept are shown above the corresponding maps. Warmer colors indicate higher concept-specific specificity evidence, while cooler colors indicate lower scores. Across subjects, high-scoring voxels appear in similar high-level visual regions, demonstrating that BrainTRACE discovers spatially localized and reproducible representations despite individual variability in cortical organization.

A.2 Quantitative Results Across Subjects

Table T1 compares the quantitative results across the four NSD subjects used in our analysis (1, 2, 5, and 7), who completed all sessions, using regions of size 100 and the top 50 concepts selected by each method. The same overall pattern is consistent across subjects. Activation-based methods, especially Max Activation, achieve the strongest generated activation scores, while MindSimulator+ often gives the highest measured activation scores. However, BrainTRACE consistently yields the strongest specificity scores on generated negatives, measured negatives, and counterfactual edits for every subject. This suggests that the benefit of specificity-based ranking is not specific to a single subject, but generalizes across individuals.

Table T1: **Scores on the top 50 concepts across subjects (region size 100).** Average activation and specificity validation scores for regions of size 100 discovered separately for each subject. For every subject, scores are averaged over the top 50 concepts selected by the corresponding method. The four subjects shown are the NSD subjects who completed all scanning sessions. Max Activation ranks voxels by average activation on generated positives; MindSimulator ranks voxels by predicted responses to CLIP-retrieved concept images; MindSimulator+ additionally filters retrieved images using VLM verification; and BrainTRACE ranks voxels using the combined specificity score. Across subjects, activation-based methods achieve the highest activation scores, while BrainTRACE consistently attains stronger specificity scores on generated negatives, measured negatives, and counterfactual edits.

		Activation		**Specificity**		
Subject	**Method**	**Gen.**	**Meas.**	**Gen.**	**Meas.**	**Edits**
1	Max Activation	**2.83**	0.74	0.10	0.19	0.43
	MindSimulator	1.96	1.04	-0.49	0.30	0.26
	MindSimulator+	2.20	**1.19**	-0.21	0.46	0.43
	BrainTRACE	2.09	1.16	**0.62**	**0.79**	**0.97**
2	Max Activation	**2.84**	0.77	0.04	0.15	0.45
	MindSimulator	1.89	0.97	-0.51	0.18	0.21
	MindSimulator+	2.22	**1.06**	-0.27	0.28	0.38
	BrainTRACE	2.17	0.99	**0.65**	**0.61**	**0.99**
5	Max Activation	**2.72**	0.80	0.08	0.26	0.48
	MindSimulator	1.95	1.24	-0.39	0.39	0.23
	MindSimulator+	2.12	**1.35**	-0.34	0.55	0.39
	BrainTRACE	2.04	1.25	**0.58**	**0.84**	**1.00**
7	Max Activation	**2.67**	0.50	0.11	0.13	0.39
	MindSimulator	1.76	0.83	-0.37	0.21	0.21
	MindSimulator+	1.97	**0.90**	-0.22	0.35	0.31
	BrainTRACE	1.90	0.89	**0.61**	**0.62**	**0.96**

B Ablation & Analysis

B.1 Comparison of Specificity and Activation Maps

We compare the regions discovered by BrainTRACE to those obtained by a standard activation-based localization method, in order to understand how specificity scoring changes the resulting concept maps. The main difference is that activation-based localization tends to assign high scores to broad sets of voxels, often producing large regions that likely reflect correlated visual or semantic factors rather than concept-specific representations. In contrast, BrainTRACE uses specificity scoring to favor voxels whose responses remain stronger for the target concept than for both semantic negatives and counterfactual edits, leading to much more selective candidate regions. This increased selectivity is important for two reasons: it reduces the false positive rate of concept discovery, and it also allows the method to conclude that some concepts are not reliably represented rather than forcing a broad activation pattern into a putative localization. Fig. S2 illustrates this contrast for several example concepts. Across all of them, the activation-based maps are substantially broader; for example, for *Child*, almost the entire EBA receives high activation scores, which is unlikely to reflect a representation specific to that concept alone. In contrast, the BrainTRACE maps isolate smaller and more plausible concept-specific regions.

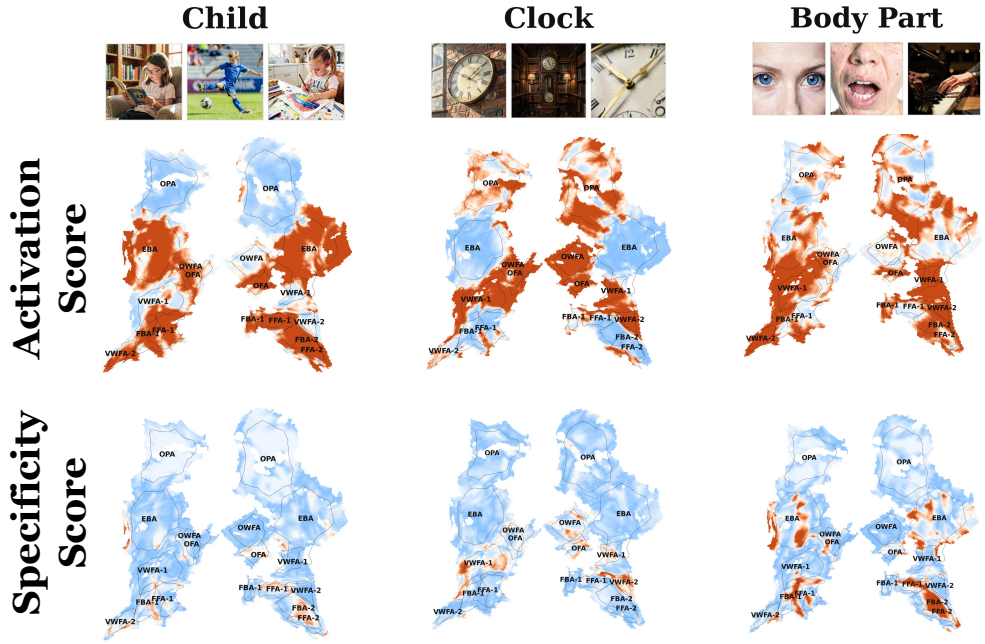

Figure S2: **Specificity versus activation-based localization across diverse concepts.** Maps compare BrainTRACE specificity scoring with activation-based localization for Child, Clock, and Body Part. Across these examples, activation-based localization produces broad high-response patterns, while specificity scoring yields more selective maps by suppressing responses driven by correlated visual or semantic cues.

B.2 Analysis of Ranking Scores

Table T2 analyzes how different ranking criteria affect the quality of the discovered regions. We analyze the following ranking criteria: Max Activation–Generated (MAG), Max Activation–Retrieved from Measured (MAM), Max Activation–Retrieved from Pool (MAL), Max Activation–Retrieved from Filtered Pool (MALF), Semantic Specificity-Generated (SSG), Semantic Specificity-Retrieved from Measured (SSM), Semantic Specificity-Retrieved from Pool (SSL), and Counterfactual Specificity-Generated (CFG).

Several trends emerge. First, ranking by **max activation** alone leads to the highest activation scores, especially when positives are generated (MAG) or retrieved from measured data (MAM), but performs substantially worse on the specificity criteria, particularly on generated semantic negatives. This indicates that activation-based ranking often identifies regions that respond strongly to the target concept images, yet are less selective when challenged with correlated alternatives. Retrieval from measured data improves measured activation and measured specificity scores relative to retrieval from the large image pool, while filtering the retrieved pool also improves specificity performance, suggesting that retrieval quality has a strong effect on the final region quality.

Second, ranking by the specificity criteria alone shifts the balance. **Semantic Specificity – Generated (SSG)** achieves the strongest generated specificity score, while **Semantic Specificity – Retrieved from Measured (SSM)** gives the strongest measured specificity score. **Counterfactual Specificity – Generated (CFG)** achieves by far the best edit score. These results support the intended role of each ranking signal: generated semantic negatives best enforce discrimination against correlated generated alternatives, measured retrieval strengthens performance on measured-data evaluation, and edit-based ranking most strongly captures sensitivity under controlled counterfactual edits.

Finally, combining the different signals provides the best overall balance. Adding semantic-specificity and edit-based scores already improves the average score over the single-score variants, and further adding activation-based signals increases activation performance without collapsing the specificity scores. The best overall result is obtained by **CFG+SSG+SSL+MALF+SSM**, which achieves the highest average score across all criteria. This suggests that no single ranking signal is sufficient on its own: strong region discovery requires combining activation, generated semantic negatives, measured semantic negatives, and edit-based counterfactual-specificity testing. Overall, the table supports the main design choice of BrainTRACE, namely that the most reliable regions are obtained by integrating multiple complementary signals rather than relying on activation alone.

Table T2: **Analysis of different ranking scores.** Scores are reported for the 50 top-ranked concepts, averaged across subjects, using regions of size 100. Rows compare different ranking criteria used to define the voxel region. Max Activation ranks voxels by their average activation across images, using positives generated by the model, retrieved from measured images, or retrieved from a large image pool, with and without filtering. Semantic Specificity subtracts the strongest negative response, using the top-10 negative images. Counterfactual Specificity compares each positive image to its corresponding edited negatives by subtracting the maximum edited response. Combination rows aggregate these signals; the final combination provides the best balance across activation and specificity criteria.

	Activation		**Specificity**			
Method	**Gen.**	**Meas.**	**Gen.**	**Meas.**	**Edits**	**Avg.**
Max Activation - Generated (MAG)	**2.76**	0.70	0.08	0.18	0.44	0.83
Max Activation - Retrieved from Measured (MAM)	1.71	**1.83**	-0.33	0.54	0.22	0.79
Max Activation - Retrieved from Pool (MAL)	1.89	1.02	-0.44	0.27	0.23	0.59
Max Activation - Retrieved from Filtered Pool (MALF)	2.13	1.12	-0.26	0.41	0.38	0.76
Semantic Specificity - Generated (SSG)	1.73	0.37	**0.94**	0.09	0.71	0.77
Semantic Specificity - Retrieved from Measured (SSM)	1.29	1.18	-0.10	**0.99**	0.39	0.75
Semantic Specificity - Retrieved from Pool (SSL)	1.43	0.74	0.16	0.50	0.63	0.69
Counterfactual Specificity - Generated (CFG)	1.67	0.56	0.45	0.21	**1.42**	0.86
CFG+SSG+SSL	1.90	0.65	0.82	0.37	1.12	0.97
CFG+SSG+SSL+MAG	2.29	0.73	0.81	0.39	1.08	1.06
CFG+SSG+SSL+MALF	2.10	0.84	0.70	0.45	1.05	1.03
CFG+SSG+SSL+MALF+SSM	2.05	1.08	0.62	0.71	0.98	**1.09**

B.3 Region Size Analysis

Table T3 analyzes how the quality of the discovered regions changes as the region size increases from 50 to 1000 voxels. A clear pattern emerges across all sizes. Activation-based methods, especially **Max Activation**, consistently achieve the strongest activation scores on generated positives, while **MindSimulator+** often achieves the highest measured activation scores. However, these gains come at the cost of substantially weaker specificity scores. In contrast, ***BrainTRACE*** consistently obtains the strongest specificity scores across all region sizes and across all three counterfactual-specificity evaluation criteria: generated negatives, measured negatives, and counterfactual edits. As the region size increases, scores gradually decrease for all methods, indicating that larger regions become less selective. Nevertheless, the relative advantage of ***BrainTRACE*** remains stable, showing that the proposed specificity ranking yields regions that are more faithful to the target concept across a broad range of region sizes.

Table T3: **Scores across region sizes for the top 50 discovered concepts.** Average activation and specificity validation scores for regions discovered by each method, evaluated at different region sizes. For each region size, scores are averaged over the top 50 concepts selected by the corresponding method. Max Activation ranks voxels by average activation on generated positives; MindSimulator ranks voxels by predicted responses to CLIP-retrieved concept images; MindSimulator+ additionally filters retrieved images using VLM verification; and BrainTRACE ranks voxels using the combined specificity score. As region size increases, scores gradually decrease for all methods, reflecting the reduced selectivity of larger regions. Nevertheless, BrainTRACE maintains relatively high activation while consistently achieving substantially stronger specificity scores on generated negatives, measured negatives, and counterfactual edits across all tested region sizes.

		Activation		**Specificity**		
ROI Size	**Method**	**Gen.**	**Meas.**	**Gen.**	**Meas.**	**Edits**
50	Max Activation	**2.91**	0.65	0.08	0.16	0.46
	MindSimulator	1.97	1.00	-0.45	0.26	0.24
	MindSimulator+	2.22	1.07	-0.26	0.39	0.40
	BrainTRACE	2.13	**1.08**	**0.68**	**0.71**	**1.06**
100	Max Activation	**2.76**	0.70	0.08	0.18	0.44
	MindSimulator	1.89	1.02	-0.44	0.27	0.23
	MindSimulator+	2.13	**1.12**	-0.26	0.41	0.38
	BrainTRACE	2.05	1.08	**0.62**	**0.71**	**0.98**
200	Max Activation	**2.61**	0.74	0.06	0.19	0.40
	MindSimulator	1.78	1.01	-0.47	0.25	0.19
	MindSimulator+	2.01	**1.12**	-0.30	0.39	0.32
	BrainTRACE	1.95	1.04	**0.55**	**0.69**	**0.90**
500	Max Activation	**2.35**	0.72	0.01	0.17	0.35
	MindSimulator	1.60	0.96	-0.49	0.21	0.14
	MindSimulator+	1.79	**1.05**	-0.32	0.34	0.28
	BrainTRACE	1.77	0.96	**0.45**	**0.62**	**0.80**
1000	Max Activation	**2.12**	0.64	0.00	0.16	0.31
	MindSimulator	1.42	0.90	-0.52	0.16	0.09
	MindSimulator+	1.62	**0.98**	-0.36	0.27	0.21
	BrainTRACE	1.59	0.87	**0.36**	**0.55**	**0.70**

B.4 Human Validation of Automated Stimulus Verification

We evaluate the accuracy of the automated stimulus-verification procedure using human annotations. Two annotators independently assessed 850 image–concept pairs sampled from 50 randomly selected concepts, including both positive and negative examples, and judged whether the target concept was present or absent in each image. We then compare these human judgments with the labels assigned by the automated verification stage used to construct the stimulus sets. Relative to the human annotations, the automated verification achieves a precision of 94.6% and a recall of 90.6%. Overall, the automated labels show close agreement with human judgments across the evaluated concepts.

B.5 Analysis of Low-Level Visual Confounds

We evaluate whether counterfactual-specificity scores can be explained by systematic low-level visual differences between the original and edited images. Across 67,517 original–edited image pairs, we quantify changes in nine image-level properties: mean luminance, RMS contrast, color-histogram distance, grayscale-histogram distance, edge density, edge energy, Laplacian energy, high-frequency power, and spectral centroid. For each concept, we average the change in each visual property across its image pairs and correlate these values with the corresponding counterfactual-specificity score across subject–concept pairs.

The associations between individual low-level features and counterfactual specificity are modest. Pearson correlations range from $r = 0.10$ for luminance to $r = 0.20$ for edge energy, with the corresponding Spearman correlations ranging from 0.13 to 0.24. Table T4 reports the correlations for all nine features. A joint linear regression using all nine standardized visual features explains 6.2% of the variance in counterfactual-specificity scores. We additionally residualize the counterfactual-specificity scores with respect to the variance predicted by the nine low-level features. The residualized scores remain highly similar to the original scores, with a Pearson correlation of $r = 0.97$ and a Spearman rank correlation of $\rho = 0.95$. Among subject–concept pairs in the top quartile of the original counterfactual-specificity scores, 92.3% remain in the top quartile after residualization. These analyses show that the observed variation in counterfactual specificity is only weakly associated with the measured low-level visual changes.

Table T4: **Association between low-level visual changes and counterfactual specificity.** For each visual property, we report the Pearson and Spearman correlations between the average change across original–edited image pairs for each concept and the corresponding counterfactual-specificity score across subject–concept pairs.

Visual Property	**Pearson** r	**Spearman** ρ
Edge energy	0.20	0.23
Laplacian energy	0.18	0.24
Edge density	0.17	0.21
Spectral centroid	0.17	0.20
High-frequency power	0.15	0.19
RMS contrast	0.15	0.19
Grayscale histogram	0.12	0.16
Color histogram	0.11	0.17
Mean luminance	0.10	0.13

B.6 Analyzing BrainTRACE False Positives

Although BrainTRACE yields a much lower number of false positives than activation-based methods, some false positives still remain. We analyzed these cases and found that many involve broad, generic image properties rather than more localized visual concepts. Representative examples include *lighting contrast*, *sky*, and *reflection*. In addition, while examining the pipeline outputs, we observed that a substantial fraction of the remaining failures are associated with semantic-negative generation failures. The semantic-negative generation pipeline is designed to produce images that exclude the target concept, using both LLM-guided prompt construction and VLM-based verification. In practice, however, failures occasionally occur. Fig. S3 shows representative examples for the target concepts *sky*, *reflection*, and *lighting contrast*, where the generated negatives still visibly contain the target concept. For example, images generated for counter concepts related to outdoor scenes may still contain large sky regions, while negatives intended to exclude reflections may nevertheless include reflective surfaces such as water or glass. These cases highlight an important limitation of semantic-negative generation for broad visual properties that are difficult to isolate cleanly. This challenge is especially pronounced when the goal is to generate semantically similar concepts that exclude the target concept itself. This remains a limitation of the current pipeline, and we expect future advances in language and vision models to improve it.



Figure S3: **Failure cases in semantic-negative generation.** Representative examples for the target concepts *sky*, *reflection*, and *lighting contrast*, where the generated semantic negatives still contain clear instances of the target concept.

C ALIGNMENT WITH MEASURED FMRI & CONSISTENCY ACROSS MODELS

C.1 ROBUSTNESS TO PIPELINE COMPONENTS

To assess whether BrainTRACE's conclusions depend on particular model choices within the pipeline, we perform a set of component-level robustness analyses. In each analysis, we replace a single component with an alternative model and rerun the corresponding evaluation, while leaving the remainder of the pipeline unchanged. We assess robustness by comparing the resulting voxelwise specificity maps using Pearson correlation and cosine similarity, and by measuring the median fraction of voxels with significantly positive specificity under the original pipeline that retain positive specificity after the component substitution. For the image-to-fMRI encoder, we additionally compare concept-level true-positive and false-positive rates.

For the language-model and image-generation/editing robustness analyses, we use Subject 1 and the same randomly selected subset of 20 concepts across all model substitutions, enabling controlled comparisons under a fixed concept set.

C.1.1 ROBUSTNESS TO THE IMAGE-TO-FMRI ENCODER

We replace the original image-to-fMRI encoder with the independently trained subject-specific Transformer encoder of (Adeli et al., 2025). Concept-level performance is similar across the two encoders: the TPR is 38.7% with the original encoder and 36.6% with the alternative encoder, while the FPR decreases from 23.0% to 18.3%. The resulting voxelwise specificity maps are also highly consistent, with a mean Pearson correlation of $r = 0.87$ and a mean cosine similarity of 0.98. Among voxels with significantly positive specificity under the original encoder, a median of 79.5% retain positive specificity under the alternative encoder. Together, these results suggest that both concept-level discovery and voxelwise localization are largely preserved under a substantially different image-to-fMRI prediction model.

C.1.2 ROBUSTNESS TO THE LANGUAGE FOUNDATION MODEL

We replace Gemma 3 with Mistral Small 3.2[1] for the language-model component of stimulus-set construction and repeat the analysis for Subject 1 across 20 concepts. The resulting specificity maps remain highly similar to those obtained with Gemma 3, with a mean Pearson correlation of $r = 0.86$ and a mean cosine similarity of 0.98. The median retention of voxels with positive specificity is 97.9%. This indicates that the identified specificity patterns are not strongly tied to the particular language model used to construct the stimulus sets.

C.1.3 ROBUSTNESS TO THE IMAGE GENERATION AND EDITING MODEL

We replace the FLUX-based image generation and editing component with Mage-Flow (Zhang et al., 2026) and repeat the analysis for Subject 1 across 20 concepts. Despite the use of a different generative model, the resulting specificity maps remain strongly aligned with those from the original pipeline, with a mean Pearson correlation of $r = 0.81$ and a mean cosine similarity of 0.98. The median retention of voxels with positive specificity is 93.8%. These results indicate that the voxelwise specificity patterns are largely preserved across alternative choices of image generation and editing model.

[1] https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506

Table T5: **Robustness of BrainTRACE to alternative pipeline components.** Each analysis replaces a single pipeline component with an alternative model. Pearson correlation and cosine similarity quantify agreement between the resulting voxelwise specificity maps, and retention denotes the median fraction of voxels with significantly positive specificity under the original pipeline that retain positive specificity after the substitution.

Component	Alternative Model	Pearson r	Cosine	Retention
Image-to-fMRI encoder	Adeli et al.	0.87	0.98	79.5%
Language model	Mistral Small 3.2	0.86	0.98	97.9%
Generation/editing model	Mage-Flow	0.81	0.98	93.8%

C.1.4 Robustness of FPR Across Evaluation Pipelines

We additionally test whether the false-positive-rate comparison between activation-based and specificity-based discovery depends on the particular models used to construct the evaluation stimuli. We perform this analysis on the same randomly selected 20-concept subset used in the Mistral and Mage-Flow robustness experiments above. Candidate regions are selected using the training split and evaluated on a fully disjoint validation split containing different positive and negative images. We repeat the false-positive evaluation using the original stimulus-construction pipeline, a Mage-Flow-based evaluation set in place of FLUX, and a Mistral-based evaluation set in place of Gemma. The activation-based FPR is 47.4% with the original evaluation set and 52.6% with both the Mage-Flow and Mistral evaluation sets. In contrast, the specificity-based FPR is 10.5% in all three settings. Thus, the reduction in false positives under specificity-based evaluation is preserved across the alternative stimulus-generation pipelines.

Table T6: **Concept-level robustness to the image-to-fMRI encoder.** True-positive rate (TPR) and false-positive rate (FPR) obtained with the original BrainTRACE encoder and the alternative image-to-fMRI encoder of Adeli et al.

	Original Encoder	Adeli et al. Encoder
TPR	38.7%	36.6%
FPR	23.0%	18.3%

Table T7: **False-positive-rate robustness across evaluation pipelines.** Candidate regions are selected on the training split and evaluated on disjoint validation stimuli constructed using the original pipeline or alternative generation/model components.

Evaluation Set	Activation FPR	Specificity FPR
Original	47.4%	10.5%
Mage-Flow	52.6%	10.5%
Mistral	52.6%	10.5%

C.2 MEASURED–PREDICTED FMRI CONSISTENCY

We evaluate the consistency between measured and predicted fMRI responses on the same NSD images. All stimuli in this analysis were held out from encoder training. For each concept, we retrieve the corresponding positive and semantic-negative images from the measured NSD data and compute the encoder-predicted responses for those identical stimuli. For both measured and predicted fMRI, we construct two voxelwise maps: the mean response to positive images and the semantic-specificity map, defined as the positive response relative to the strongest semantic negatives. For each concept, we compute the voxelwise Pearson correlation between the measured and predicted maps and report the mean and standard deviation across concepts for each subject. Across subjects, the mean correlation between measured and predicted positive-response maps ranges from $r = 0.93$ to $r = 0.96$. For semantic-specificity maps, the mean correlation ranges from $r = 0.72$ to $r = 0.80$. These results show that the predicted responses closely reproduce the measured voxelwise activation structure and retain substantial agreement for the more selective semantic-specificity maps. Results for all four subjects are reported in Table T8.

Table T8: Consistency between measured and predicted fMRI responses. Mean $\pm$ standard deviation of the voxelwise Pearson correlation between measured and predicted maps across concepts for each subject. Positive-response maps contain the mean response to positive images, while semantic-specificity maps compare positive responses with the strongest semantic negatives. Measured and predicted responses are computed for the same NSD images.

Subject	**Positive Response**	**Semantic Specificity**
Subject 1	0.94 ± 0.05	0.76 ± 0.11
Subject 2	0.95 ± 0.03	0.77 ± 0.11
Subject 5	0.96 ± 0.03	0.80 ± 0.11
Subject 7	0.93 ± 0.05	0.72 ± 0.12

D FINAL VERDICT: CONFIDENCE SCORE & CONCEPT COVERAGE ANALYSIS

D.1 STATISTICAL TESTING OF BRAINTRACE DISCOVERED REGIONS

To assess whether the responses of BrainTRACE-discovered regions are statistically reliable, we perform image-level statistical tests on a held-out validation split and additionally quantify the corresponding effect sizes. The voxel region for each concept is selected using only the training split, while all statistical tests are performed on held-out validation images. We evaluate three criteria: activation, semantic specificity, and counterfactual specificity. To ensure that each statistical test is supported by sufficient data, we impose minimum image-coverage requirements. For the activation and semantic-specificity analyses, we require at least 20 positive images and at least 30 pooled semantic-negative images. For the counterfactual-specificity analysis, we require at least 10 positive–counterfactual image pairs.

Activation. We test whether the responses of the discovered region to positive images are significantly greater than zero. For each concept, we use a one-sided sign-flip test over the image-level region responses.

Semantic specificity. We test whether the discovered region responds more strongly to positive images than to semantically related negative images. For each concept, we compare the image-level positive responses against the corresponding strongest semantic-negative responses using a one-sided permutation test.

Counterfactual specificity. We test whether the discovered region responds more strongly to each positive image than to its edited counterfactuals. For each positive image, we select the edited counterfactual producing the strongest response in the discovered region and compute the paired response difference. We then test whether these positive-minus-counterfactual differences are significantly greater than zero using a paired one-sided sign-flip test.

For each predefined metric, we apply Benjamini–Hochberg false discovery rate (BH-FDR) correction across the 260 tested concepts, separately for each metric. Table T9 reports the number of concepts that pass the corrected significance threshold of $q \leq 0.05$, together with the number that additionally satisfy the corresponding minimum coverage requirement. Results in this analysis are reported for Subject 1. Generated activation and generated semantic specificity are significant for most tested concepts, while the measured-data criteria are satisfied by fewer concepts, reflecting the more limited coverage of positive and semantic-negative examples in NSD. Counterfactual specificity is more selective, with 57 concepts passing BH-FDR correction and 56 also satisfying the minimum coverage requirement.

Effect sizes. Because image-level statistical tests can achieve high power with many validation images, we additionally quantify the magnitude of each effect. For each validation criterion, we compute the same image-level contrast used in the corresponding statistical test: the mean positive response for activation, the positive-minus-semantic-negative response for semantic specificity, and the paired positive-minus-counterfactual response for counterfactual specificity. We report the median raw contrast and interquartile range (IQR) across concepts passing BH-FDR correction, together with a bootstrap 95% confidence interval for the median and a standardized effect size. For activation, the standardized effect is computed using one-sample standardization; for semantic specificity, using pooled standardization of the positive and semantic-negative responses; and for counterfactual specificity, using paired Cohen's d_z.

D.2 Retrieval Statistics and Missing Concept Coverage

A key part of BrainTRACE is understanding not only which concepts can be validated in the measured data, but also which concepts are *not* sufficiently represented there. This is important for two reasons. First, low measured-data scores should be interpreted differently when the target concept itself is rarely present in the measured dataset. In such cases, weak measured validation may reflect lack of coverage rather than lack of a true brain representation. Second, this analysis helps guide future experiments by identifying which concepts and which negative controls are missing from the current dataset, and therefore which additional stimuli would be most informative to collect.

We begin by analyzing the coverage of *positive* examples in the measured image–fMRI dataset. As shown in the histogram of positive examples per target concept (Fig. S4), the number of successfully retrieved and verified positive images varies substantially across concepts, ranging from very few examples to nearly 200. A considerable fraction of concepts have only a small number of measured positives, while others are much better represented. This variability is critical for interpreting measured-data validation: when a concept has only limited coverage, even strong or weak measured-data scores are less informative, since the available dataset provides only sparse evidence about that concept. In contrast, when many verified positive examples are available, measured-data validation can provide much stronger support for or against a discovered representation. This retrieval-coverage analysis therefore provides an estimate of how informative the measured dataset is for each concept. It is used both to interpret the reliability of discovered representations and to identify concepts for which additional data would be especially valuable. In practice, concepts with high specificity score but low positive coverage are treated as less well validated by measured-data evaluation alone, and are natural candidates for follow-up experiments with targeted stimuli.

In addition to positive coverage, BrainTRACE also analyzes retrieval of *semantic negatives*, which are important for validating specificity under counterfactual and semantic-negative comparisons on measured data. For each target concept, the language model proposes relevant negative concepts, and the retrieval pipeline searches for measured examples using CLIP-based scoring together with a double verification process: the target concept must be absent, and the intended negative concept must be present. The retrieval statistics further show substantial variability in the number of successfully retrieved semantic negatives, both per target concept and per target–negative pair. In particular, many target–negative pairs either yield no measured examples or only a few, while others reach the maximum retrieval count. This means that for some concepts, measured-data counterfactual evaluation is limited not only by missing positives but also by missing semantically informative negatives. Such cases are especially important for experimental planning, since they indicate which negative controls should be added in future fMRI studies to better distinguish true concept-specific responses from correlated alternatives.

Table T9: **Image-level statistical testing of BrainTRACE discovered regions.** For each criterion, we report the number of concepts passing BH-FDR correction at $q \leq 0.05$, and the number that pass both the statistical test and the corresponding minimum image-coverage requirement. Regions are discovered using only the training split and evaluated on held-out validation images.

Criterion	Pass BH-FDR	Pass BH-FDR & Coverage
Generated Activation	232	231
Measured Activation	160	96
Generated Semantic Specificity	234	234
Measured Semantic Specificity	147	87
Counterfactual Specificity	57	56

Table T10: **Effect sizes for the image-level statistical tests.** For each criterion, we report effect-size statistics across concepts passing BH-FDR correction at $q \leq 0.05$. Raw contrasts correspond to the image-level contrast used in the associated statistical test. Standardized effects use one-sample standardization for activation, pooled standardization for semantic specificity, and paired Cohen's d_z for counterfactual specificity.

Criterion	# Pass	Median Contrast [IQR]	95% CI	Std. Effect [IQR]
Generated Activation	232	1.15 [0.74, 1.60]	[1.03, 1.33]	1.69 [1.00, 2.64]
Measured Activation	160	0.50 [0.33, 0.81]	[0.42, 0.55]	0.97 [0.63, 1.54]
Generated Semantic Specificity	234	0.98 [0.71, 1.54]	[0.93, 1.08]	1.47 [1.07, 2.16]
Measured Semantic Specificity	147	0.54 [0.34, 0.82]	[0.44, 0.63]	1.00 [0.75, 1.47]
Counterfactual Specificity	57	0.76 [0.44, 1.25]	[0.57, 1.05]	0.99 [0.75, 1.42]

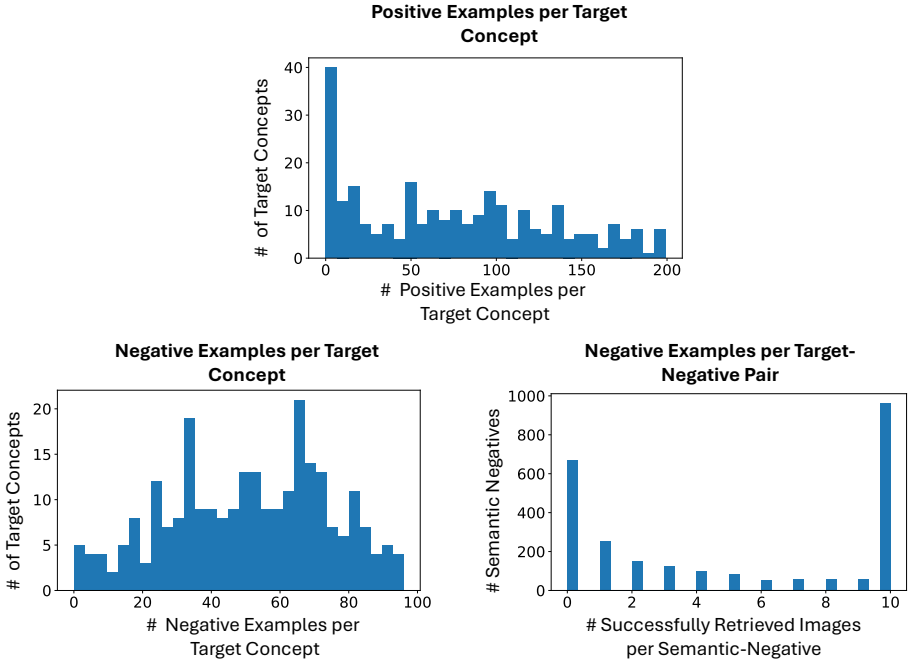

Figure S4: **Coverage of retrieved positives and semantic negatives in the measured dataset.** For each target concept, we count the number of successfully retrieved and verified positive examples (top) and semantic-negative examples (bottom left). We also report the number of successfully retrieved negatives for each target–negative pair (bottom right). This analysis is used to assess how informative the measured data is for validating discovered representations and for identifying missing evidence that may require follow-up experiments.

E ADDITIONAL VISUAL RESULTS

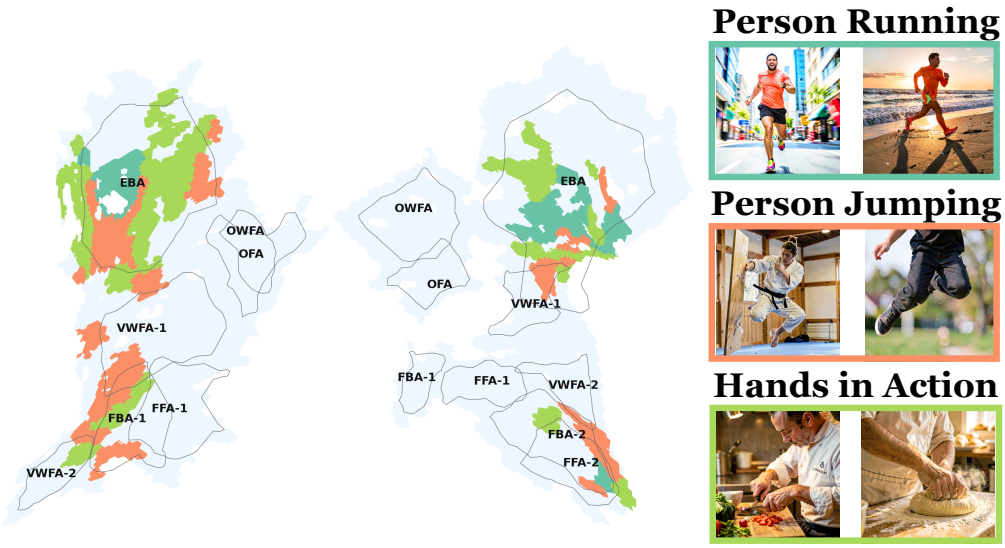

Figure S5: **Fine-grained organization of action- and body-related representations.** Binary maps show voxels with positive specificity evidence for Person Running, Person Jumping, and Hands in Action, with matching example images shown at right. BrainTRACE reveals partially distinct spatial patterns for these related body-action concepts across high-level visual cortex.

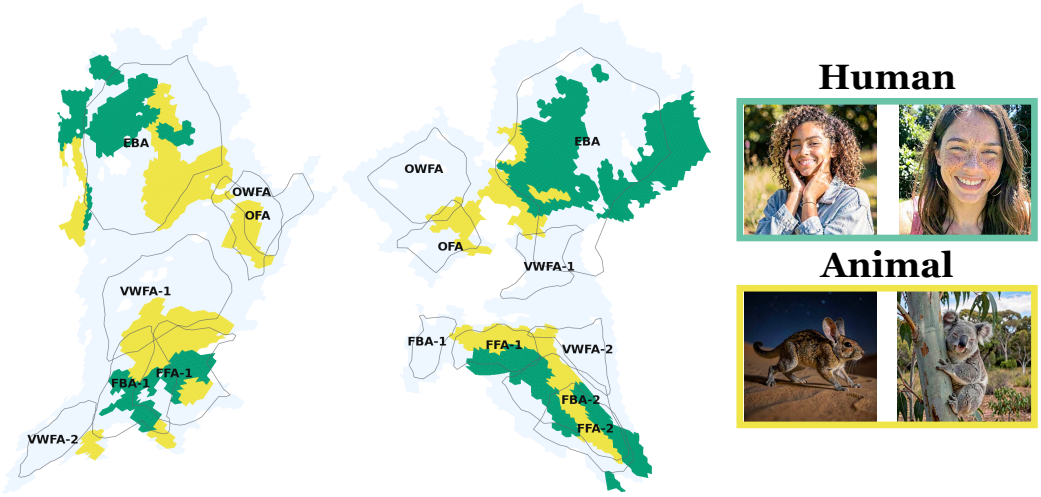

Figure S6: **Human and animal representations.** Binary maps show voxels with positive specificity evidence for Human and Animal, with matching example images shown at right. BrainTRACE reveals overlapping but distinct spatial patterns for these concepts across high-level visual cortex.

F Additional Technical Details

F.1 Brain Data

Functional magnetic resonance imaging (fMRI) measures changes in blood oxygenation that are related to neural activity over time. In visual fMRI experiments, subjects view images while whole-brain activity is recorded as a temporal sequence of 3D brain volumes. Following standard practice, this temporal signal is converted into image-level response estimates by fitting a response model, producing a beta value for each voxel and each presented image. Thus, after preprocessing, each image is associated with a single fMRI response vector, where each element corresponds to the estimated response of one voxel. We use the preprocessed fMRI responses provided by the Natural Scenes Dataset (NSD) (Allen et al., 2022), following the preprocessing and vectorization protocol used in prior work. The NSD contains 7T fMRI recordings from eight subjects viewing natural images from COCO, with each subject viewing approximately 10,000 images over repeated scanning sessions. We further normalize each voxel independently across all measured images to have mean 0 and standard deviation 1, so that a response above 0 indicates activation above that voxel's average response across images.

Large Predicted fMRI Pool. In addition to measured data, following the approach of Bao et al. (2025), we construct a large pool of predicted image–fMRI pairs. Specifically, we take 120K unlabeled images from COCO and use the image-to-fMRI model to predict the corresponding fMRI response for each subject. These predicted pairs are used to support retrieval over a large image pool and to implement the MindSimulator baseline.

F.2 Image & Brain Models

Image-Based Models For both image generation and image editing, we use FLUX.2-Klein-4B from Black Forest Labs[2]. For image generation, we use images of resolution 512×512, guidance scale 9.0, and 35 inference steps. For image editing, we use the same resolution and guidance scale, but reduce the number of inference steps to 15. For image–concept verification, we use Qwen3-VL-8B-Instruct[3]. Given an image and a target concept, the model is prompted to determine whether the concept is present in the image. We use temperature 0.0 and constrain the model to answer only `yes` or `no`.

Image-to-fMRI Prediction Model. Image-to-fMRI models are trained to predict a subject's fMRI response from a given image. We use the image-to-fMRI encoder from Beliy et al. (2024), which is trained jointly on image–fMRI pairs from all NSD subjects while still producing subject-specific predictions. This model allows us to take any image and predict the corresponding fMRI activation for each subject. We refer to these predicted responses as *predicted fMRI*. Since fMRI prediction remains noisy, we filter out voxels whose encoding performance on a held-out test set is below a Pearson correlation of 0.2, reducing noise in the representation discovery process.

F.3 Concept-Specific Image Retrieval

Following the MindSimulator approach, we use CLIP to retrieve images for a given concept. We first precompute CLIP image embeddings for all images in the NSD image–fMRI dataset, as well as for a large external pool of 120K images from COCO. Given a target concept, we encode it using the CLIP text encoder and compute the cosine similarity between the concept embedding and each image embedding, obtaining an alignment score for every image. Images are then ranked by this score, and the top-ranked images are retrieved. For the MindSimulator implementation used in our comparison, we use the same 120K external image pool, the same image-to-fMRI encoder, and retrieve the same number of images as generated by our method, in order to ensure a fair comparison. In practice, however, we found that many retrieved images do not clearly depict the desired concept, especially for more complex concepts where CLIP is less reliable and where COCO images often contain multiple objects and semantic elements. We therefore add a filtering step using a vision-language model (Qwen), which verifies whether the desired concept is indeed present in the retrieved image.

[2] `https://huggingface.co/black-forest-labs/FLUX.2-klein-4B`

[3] `https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct`

We also use retrieval from the measured image–fMRI dataset itself. For positive examples, retrieval is performed in the same way, followed by the same verification step to ensure that the target concept is clearly present. Retrieving semantic negatives is more challenging. For each target concept, the language model proposes a set of negative concepts, and for each such negative concept we seek measured images that satisfy two conditions simultaneously: the desired negative concept should be present, and the positive target concept should be absent. To enforce this, we use a two-stage CLIP-based retrieval procedure. First, we retrieve the top 100 images according to alignment with the negative concept. We then re-rank these candidates according to their distance from the positive concept, favoring images that are less aligned with the target concept. Finally, we apply a double verification step using the vision-language model: one check ensures that the negative concept is present, and the other ensures that the positive concept is absent. This retrieval pipeline allows us to obtain measured positive and semantic-negative examples that can be used for measured-data validation of the discovered representations.

F.4 Extended Details on Concept-Selective Representation Search

Given a target concept, BrainTRACE uses the concept-targeted stimulus set to assign each voxel both an activation score and specificity scores. Let $P = \{x_i^+\}_{i=1}^{N_P}$ denote the set of positive images, $N = \{x_j^-\}_{j=1}^{N_N}$ the set of semantic negative images, and for each positive image x_i^+, let $E_i = \{x_{i,k}^{\text{edit}}\}_{k=1}^{N_{E_i}}$ denote its corresponding counterfactual edited images. For a voxel v, let $a_v(x)$ denote its activation on the fMRI response associated with image x, whether measured or predicted.

We first define the ***Positive Score*** of voxel v as the average activation on the positive images,

$$S_{\text{pos}}(v) = \frac{1}{N_P} \sum_{i=1}^{N_P} a_v(x_i^+).$$

This measures how strongly the voxel responds, on average, to the target concept. To test specificity against semantically related alternatives, we define the ***Semantic-Negative Score*** by comparing the positive score to the hardest semantic negatives for that voxel. Specifically, let $N_v^{\text{hard}} \subseteq N$ be the set of the top 10 semantic negative images with the highest activations under voxel v. We then define

$$S_{\text{neg}}(v) = \frac{1}{N_P} \sum_{i=1}^{N_P} a_v(x_i^+) - \frac{1}{|N_v^{\text{hard}}|} \sum_{x \in N_v^{\text{hard}}} a_v(x).$$

A high value indicates that the voxel responds more strongly to the target concept than to the most confusable semantic alternatives. To test specificity under controlled edits, we define the ***Counterfactual Score*** by comparing each positive image to its hardest edited version. For each positive image x_i^+, let

$$e_i^*(v) = \max_{x \in E_i} a_v(x),$$

that is, the activation of voxel v on the single hardest edited negative associated with x_i^+. The counterfactual specificity score is then

$$S_{\text{edit}}(v) = \frac{1}{N_P} \sum_{i=1}^{N_P} \Big(a_v(x_i^+) - e_i^*(v) \Big).$$

A high value indicates that the voxel's response decreases when the target concept is specifically removed or replaced while the rest of the image is kept as similar as possible.

Using these scores, BrainTRACE constructs a candidate representation for the concept as a set of voxels whose average specificity score is positive. Concretely, for each voxel v, we define its specificity score as the average of the semantic-negative and counterfactual specificity scores,

$$S_{\text{spec}}(v) = \frac{1}{2} \Big(S_{\text{neg}}(v) + S_{\text{edit}}(v) \Big),$$

and select the voxel set

$$R = \{v : \ S_{\text{spec}}(v) > 0\}.$$

This set serves as the candidate region representing the concept. To evaluate the candidate region as a whole, BrainTRACE averages activations across voxels in R and computes the same scores at the region level. Specifically, for any image x, the region activation is

$$a_R(x) = \frac{1}{|R|} \sum_{v \in R} a_v(x),$$

and the positive, semantic-negative specificity, and counterfactual specificity scores are computed by replacing $a_v(\cdot)$ with $a_R(\cdot)$ in the definitions above. The specificity score of the discovered region on the training set is used to determine whether the region is a strong candidate representation for the concept. Final quantitative evaluation is then performed on a separate evaluation split and on measured fMRI data, as described in Sec. 3.3.

F.5 Computational Resources

Our pipeline uses large models during the data creation process. We ran all experiments on H200 GPUs, although smaller GPUs can also be used when accessing the open-source models through APIs. All models used in our work are open source. Given a target concept to query, the full process of positive and negative image generation, yielding approximately 1000 images per concept, score computation, and region proposal takes about 2 hours on a single H200 GPU.

F.6 Code Availability

Anonymized code for reproducing the main BrainTRACE pipeline and analyses is available at: `https://anonymous.4open.science/r/braintrace-repro-1C8C/`.