Closed-loop hypothesis generation and testing for characterizing representations in visually-responsive brain regions

Oliver Contier*[,1,2], Shu Fujimori[1,3], N. Apurva Ratan Murty[4], Martin N. Hebart**[,2]

[1] Vision and Computational Cognition Group, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
[2] Department of Medicine, Justus Liebig University Giessen, Giessen, Germany
[3] Department of Mechanical and Intelligent Systems Engineering, Graduate School of Informatics and Engineering, The University of Electro-Communications, Tokyo, Japan
[4] School of Psychological and Brain Sciences, Georgia Institute of Technology, Atlanta, GA, USA

Lead author: Oliver Contier, contier@cbs.mpg.de

Corresponding authors:
*Oliver Contier, contier@cbs.mpg.de;
**Martin N. Hebart, hebart@cbs.mpg.de

Summary

A fundamental goal of visual neuroscience is to reveal the information represented in visual brain regions. Traditional targeted experiments provide strong tests of predefined hypotheses but scale poorly across the large space of possible representational features [1,2]. In contrast, large-scale, condition-rich studies allow broader model comparisons but may lack the specificity for discriminating among competing hypotheses [3,4]. Together, these constraints restrict the space of hypotheses that can be evaluated within any individual study. To address these challenges, we introduce a closed-loop approach that generates *in-silico* brain responses to natural images, derives candidate representational dimensions, and tests them *in vivo* with targeted neuroimaging experiments [5–9]. Applied to the fusiform face area (FFA) [10], this approach revealed 15 hypothesized representational dimensions from in-silico responses to 26,107 images, including distinct dimensions for human faces, animal faces, and other non-face dimensions. We assess the relevance of these dimensions for FFA responses in two new neuroimaging experiments. First, we showcase how these dimensions can be linked to readily available large-scale neuroimaging data, testing their relevance for predicting in-vivo responses. Second, we close the loop by translating these in-silico hypotheses into rigorous experimental validation, testing the degree to which they are represented in separable and generalizable activation patterns. The results confirm separable dimension-evoked responses for human faces, animal faces, and other non-face dimensions in human FFA. Together, these findings establish a broadly applicable framework for scalable hypothesis generation and experimental validation, enabling the systematic identification of interpretable representations in human visual cortex.

Keywords

fMRI, closed-loop, multidimensional representation, dimensions, visual cortex, fusiform face area, science automation

Results

Characterizing representations in visual cortex based on high-throughput sampling in-silico

To combine the complementary strengths of condition-rich approaches and controlled experiments, we introduce a closed-loop framework for broadly characterizing representations in human visual cortex (Figure 1A). This characterization is achieved by identifying interpretable dimensions underlying the representational structure of activation patterns in visually responsive brain regions. The framework draws on encoding models based on artificial neural networks (ANN) trained to predict neural responses to images [9,11,12]. Given their predictive accuracy, these models serve as a powerful computational proxy to generate predicted responses to arbitrary stimuli, allowing to sample in-silico brain responses at scales unattainable in vivo. These in-silico responses are then used to identify interpretable candidate representational dimensions, which are subsequently tested in targeted experiments.

We exemplify this closed-loop framework on the human fusiform face area (FFA), a region in higher-level visual cortex that responds highly selectively to images of faces [10]. Given its established role in visual face processing and the extensive work on characterizing its response profile [13,14], the FFA provides an ideal testbed for asking whether model-guided hypothesis generation can recover expected selectivity, and whether it can propose additional dimensions for experimental validation [15]. To generate candidate hypotheses, we used an existing encoding model of the FFA shown to predict functional MRI (fMRI) responses to visual objects with high accuracy [9], making it ideal for predicting responses to novel stimuli. To broadly sample the space of visual objects in our model, we used the THINGS image database, comprising 26,107 naturalistic images of 1,854 concrete, nameable objects [16]. We passed these images as inputs to the encoding model, yielding a high-throughput sample of in-silico FFA response patterns (Figure S1A).

In the next step, we derived hypothesized representational dimensions from the in-silico FFA response patterns (Figure S1B). To this end, we adapted a similarity embedding approach that has been used to uncover interpretable dimensions underlying behavioral similarity judgements in a triplet odd-one-out task [17,18]. The embedding represents each image as a sparse vector of non-negative weights along latent dimensions, thus yielding dimensions that can be meaningfully interpreted and translated into testable hypotheses about candidate features or other selectivities. Instead of using human odd-one-out choices, here we generated analogous choices from the in-silico FFA responses, treating it as a substitute for human behavior. Specifically, for a given set of three images, we defined the odd-one-out as the image with the most dissimilar in-silico FFA response. We repeated this process for 4 million random triplets, yielding an extensive set of simulated in-silico choices. Using these simulated choices, we then trained an embedding model that assigned each image a weight along every latent dimension and optimized weights such that similarities between the resulting image embeddings predicted the choices. After training, this yielded a set of latent dimensions that captured the similarity structure and predicted the odd-one-out choices of the in-silico FFA responses. We chose to

derive these dimensions not from the ANN model directly, but rather from the in-silico predictions of neural responses generated from this ANN, since this allowed us to specifically target the subspace of the underlying model representation that is realized in the in-silico patterns of brain activity. Importantly, these dimensions need not be specific to FFA, but may also be present in other parts of visual cortex (Figure S2)[19].



Figure 1. Closed-loop identification of representational dimensions in the fusiform face area (FFA). A: Overview of the approach. An encoding model trained on human neuroimaging data generates predicted in-silico responses to large stimulus samples, enabling extensive

characterization of the underlying representational dimensions, which are then tested in targeted in-vivo experiments. B: Representational dimensions identified from in-silico FFA responses. Each dimension is illustrated by its most highly weighted images. Word clouds show labels for each dimension provided by human raters (N=12).

The analysis revealed 15 representational dimensions underlying in-silico FFA responses (Figure 1B). At a coarse level, these dimensions could be classified into face- and body-related object properties, other high-level object categories, and visual-perceptual features. Among the face- and body-related dimensions, we found distinct dimensions encoding human faces, animal faces, and body parts. These results suggest distinct representations for human faces and animal faces within the FFA. Since the dimensions in the embedding were not constrained to be orthogonal, it is possible that the animal-face dimension merely reflected a scaled version of a more general face-selective dimension. However, we found that they were represented largely orthogonal from each other (r = -.05), indicating that responses to human and animal faces may indeed be related to separate representational axes.

Moreover, the recovered dimensions extended beyond face- and body-related information. Dimensions associated with other high-level object categories included scenes, food, and tools, suggesting that non-face related information is encoded in the FFA in a structured manner, consistent with broader multidimensional accounts of category-selective visual cortex [15,19–21]. Finally, some dimensions yielded images showing pronounced visual features such as shape, texture, and material properties, aligning with previously reported effects of lower- and mid-level visual features on FFA responses [22–28].

Together, these results indicate that in-silico FFA responses encode a broad set of object features that reproduce the region's canonical selectivity for faces while also revealing novel candidate dimensions. These findings demonstrate that the first stages of the proposed closed-loop framework were successful, yielding a broad set of candidate dimensions that serve as testable hypotheses about the representational space of the targeted brain region.

Validating candidate representational dimensions based on in-vivo FFA responses

Having generated candidate representational dimensions in silico, we tested whether they are also found in the human FFA in vivo. First, we showcase how this approach enables straightforward testing of generated hypotheses using available large-scale neuroimaging datasets [29–33]. A preliminary assessment of these dimensions on the raw in-silico FFA responses from which they were derived already indicated that they may elicit meaningful dimension-specific responses in FFA (Figure S3, S4). To assess the explanatory power of the hypothesized representation on independent data, we used the 15 candidate dimensions to predict responses to 8,740 broadly sampled object images obtained from the large-scale

dataset THINGS-fMRI [29] and quantified the encoding performance in held-out data. As a baseline for interpreting encoding performance, we also quantified the variance explained by a model reflecting discrete face selectivity, which was constructed from image labels indicating the presence or absence of human faces (Figure 2A).

To evaluate the capacity of the in-silico representation to serve as a model of the FFA, we first inspected the variance explained by all in-silico FFA dimensions combined. The results showed that, consistently across subjects, they explained substantially more variance in FFA responses than the baseline model (Figure 2B): Averaged across FFA voxels within each subject, the baseline model accounted for 4.02% to 8.55% of response variance, whereas the in-silico dimensions, which included continuous face-related dimensions, together explained 13.28% to 27.46% of variance. This difference was not driven by dimensionality (dimensionality-matched control with shuffled assignments, maximum 95% CI bound: -0.15% to 0.11%). These findings indicate that object responses in FFA were well characterized by the multidimensional in-silico representation beyond the selectivity for faces (Figure S5), thus validating the general capacity for testing individual in-silico hypotheses.

Next, we determined how sensitive FFA responses were to each hypothesized in-silico dimension. To this end, we quantified the response amplitude evoked by each individual dimension by contrasting FFA responses to images with the highest versus lowest dimension scores (Figure 2C). FFA responses were strongest for the human-faces dimension, replicating canonical face selectivity. Notably, this dimension yielded a larger FFA response than discrete face labels, indicating that it was particularly effective at eliciting strong FFA responses beyond the mere presence or absence of a face. Beyond human faces, the strongest responses were evoked by dimensions related to animal faces and human body parts, indicating that the model-derived dimensions captured structured response variation beyond the canonical human-face preference. Together, these findings demonstrate that the in-silico approach identified meaningful face- and body-related information represented in the FFA. Interestingly, several other dimensions were associated with reduced FFA responses, indicating that the in-silico FFA responses were not fully captured by dimensions that merely mirror increased FFA activity, but instead that representational dimensions can be reflected in both reduced and increased FFA responses [34–36].

Together, these results illustrate how the proposed closed-loop framework allows generating new hypotheses about the structure of FFA representations and directly test them using readily available neuroimaging datasets, thereby providing a scalable route from in-silico discovery to in-vivo validation.

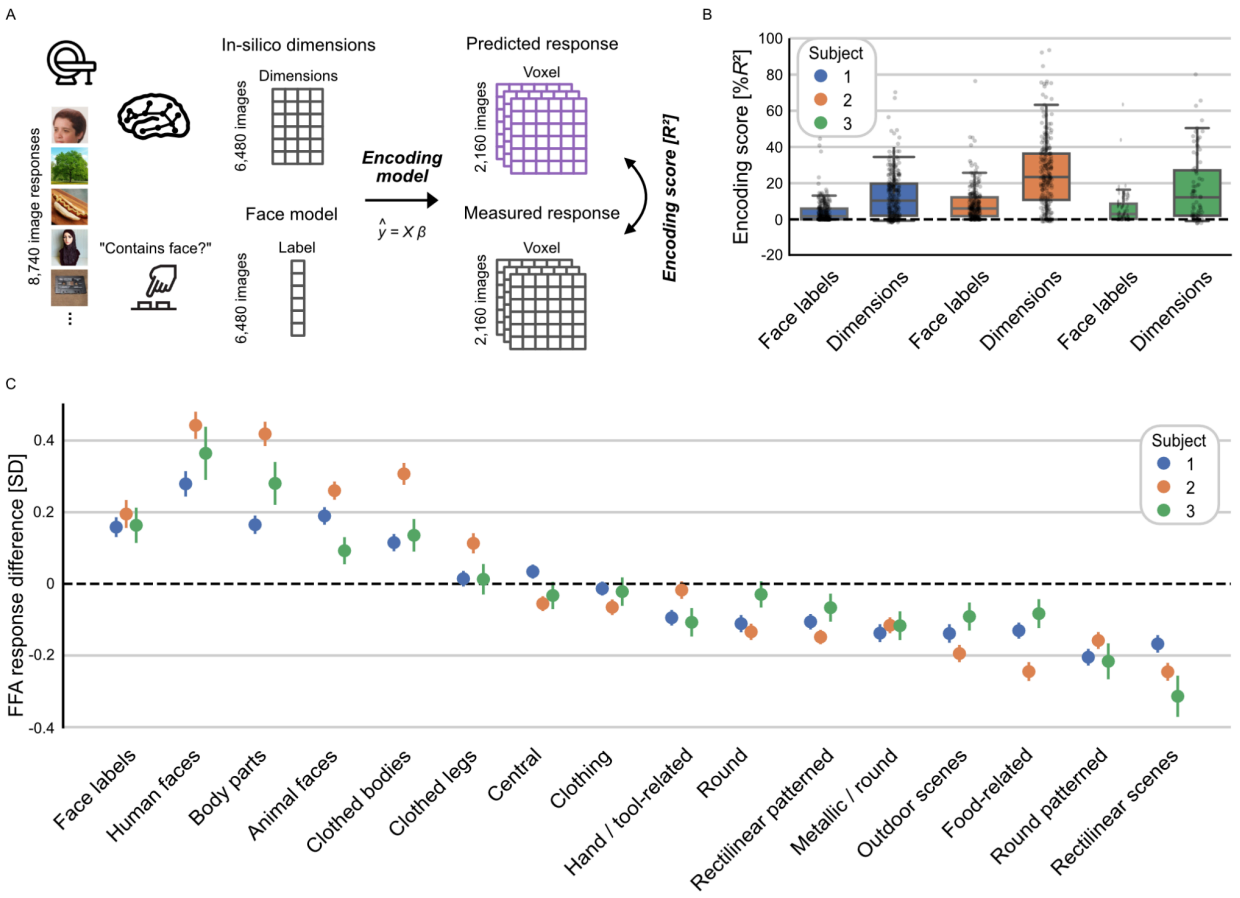


Figure 2. Validation of the hypothesized in-silico dimensions in an independent dataset of fMRI responses to broadly sampled objects. A: Illustration of the fMRI encoding model. The in-silico dimensions are used to construct an encoding model trained to fit voxel-wise FFA responses to 6,480 object images in three participants. A baseline model of face-selectivity was constructed from binary face labels. Encoding performance was scored by comparing predicted with true responses in 2,160 held-out image responses. B: Prediction performance of the in-silico dimensions and the face-selectivity model across voxels in FFA, measured in percentage of explained variance relative to the noise ceiling. C: FFA responses to each in-silico dimension, quantified as the difference in response to the 75 images with the highest versus lowest scores on each dimension (mean and 95% confidence intervals).

Targeted neuroimaging experiment demonstrates separable FFA response patterns for in-silico dimensions

While relating candidate dimensions to responses in existing neuroimaging datasets can be used to efficiently evaluate their relevance to in-vivo responses, this approach is constrained by the available stimuli sampled in such datasets, leading to two challenges. First, some candidate dimensions may be sampled sparsely, leading to a biased assessment of the presence of candidate representations. Second, the available sampled images may be insufficient for

targeted tests of a candidate dimension. Beyond these challenges, condition-rich datasets may provide less statistical power than targeted experiments to detect dimension-specific effects. Targeted experiments can address these challenges by deliberatively sampling each candidate dimension in an experimental design that ensures stable and separable dimension-specific responses that generalize across image content.

To provide a stricter evaluation of the hypothesized dimensions, we conducted a targeted neuroimaging experiment to specifically test whether these dimensions evoke stable and separable response patterns (Figure 3A). Participants viewed images most strongly associated with each dimension while fMRI recordings were acquired (N = 6, 4 female, 2 male, mean age = 26 years). Images were presented in a block design to elicit strong dimension-evoked responses, with one condition for each of the 15 hypothesized dimensions and an additional control condition containing images with low values across all dimensions (Figure S6). To attain stable response patterns, participants viewed blocks of each condition 16 times. Crucially, to ensure that responses were not driven by object-specific features but reflected a dimension as a whole, we constructed two stimulus sets for each dimension. The two sets consisted of independent images from different object categories while yielding similar distributions of dimension weights, allowing us to quantify the generalization of activation patterns across objects. We first obtained parameter estimates reflecting the activation evoked by each candidate dimension and computed a contrast with the control condition, thereby isolating dimension-specific responses from object-general, nonspecific visual activation. Based on these responses, we computed the within-dimension voxel pattern similarity, but evaluated across image sets (Figure 3B). Finally, as a measure of pattern selectivity, we computed the difference between this within-dimension similarity and the average of the between-dimension similarities with all other dimensions. Pattern selectivity therefore reflected the degree to which responses to a given dimension clustered together across images of different object categories, while remaining distinct from responses to other dimensions.

The results reveal dimension-evoked pattern selectivity in the human FFA (Figure 3C). The strongest and most consistent effects were observed for dimensions related to human faces, animal faces, and rectilinear scenes, each eliciting significant pattern selectivity in all participants. The strong selectivity for the animal faces dimension confirms the in-silico findings, which revealed largely distinct representational axes for human and animal faces ($r = -0.05$), thus indicating separate responses in human FFA to these dimensions. Interestingly, these responses overlapped in voxel responses, (Figure S7) mirrored also in spatial topographies (Figure S8), while animal-face responses correlated only weakly with independent face-localizer selectivity ($r = 0.14$ to 0.20; Figure S9). Together, these findings suggest that human- and animal-face information is encoded along distinct dimensions within overlapping voxel populations, rather than in spatially segregated clusters or as graded variation along a common face-selectivity dimension.

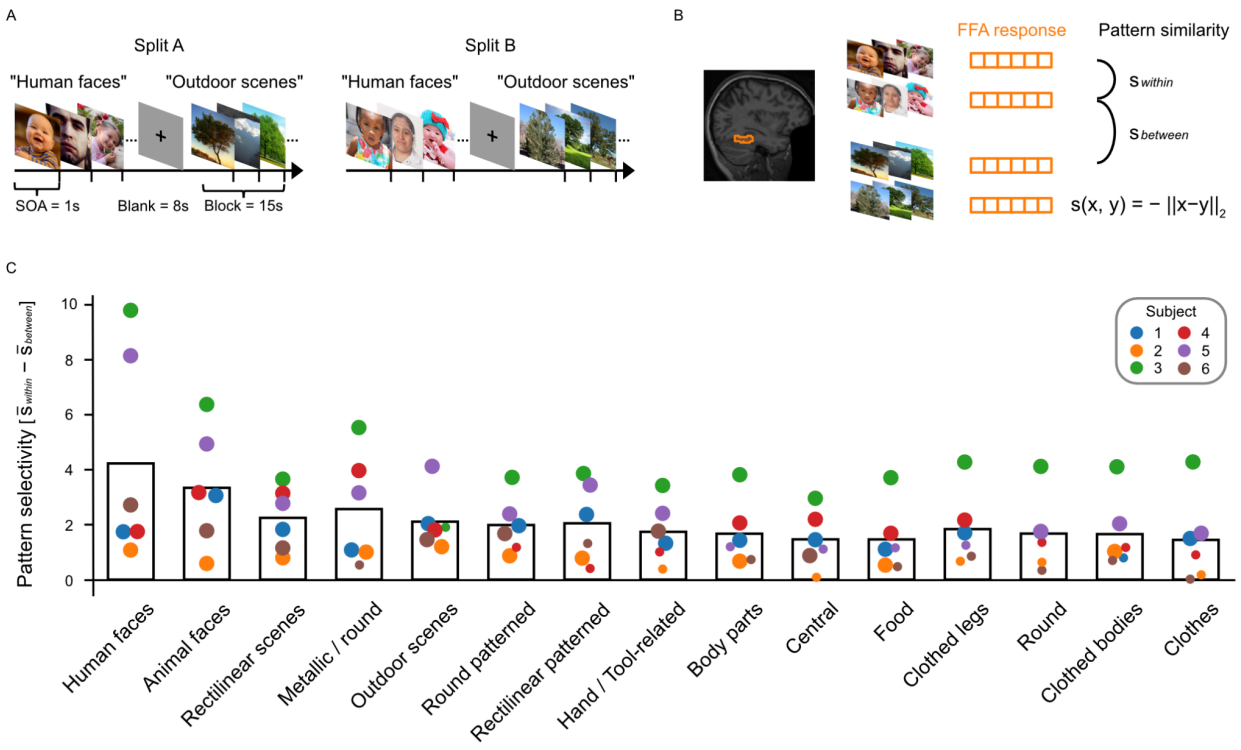


Figure 3. A targeted in-vivo fMRI experiment validating in-silico FFA dimensions. A: Experimental design. Participants (N=6) viewed images in a block design with 16 conditions (one for each dimension plus control condition). Conditions were split into two independent image sets spread across odd and even runs. B: fMRI response pattern similarity in the FFA determined within and between conditions. Within-condition pattern similarity was assessed across image sets. C: PResults showing the pattern selectivity in FFA for each dimension. Large dots indicate statistically significant effects (permutation test, $p < 0.05$; 10,000 permutations).

Beyond these dominant dimensions, the results revealed graded evidence for additional dimensions encoded in FFA responses. Dimensions such as “metallic / round”, “outdoor scenes”, and “round patterned” showed largely consistent selectivity across five out of six participants, suggesting that FFA response patterns also capture non-face structure linked to shape, texture, and scene features. Other dimensions, including “rectilinear patterned”, “hand / tool-related”, “body parts”, “central”, and “food”, showed consistent evidence in four out of six participants, suggesting that these dimensions may be present in FFA responses but are detected less reliably given their weaker responses. Finally, dimensions such as “clothed legs”, “round”, “clothed bodies”, and “clothes” showed more variable effects, suggesting that the experiment did not simply validate all candidate dimensions, but rather distinguished robust dimension-evoked responses from these less stable patterns. Note that statistical inference was performed within participants, and the recurrence of these effects indicates how consistently each dimension was expressed across participants, rather than similarity of response patterns between individuals. For completeness, we repeated the same analyses on the in-silico FFA responses, demonstrating that the experimental findings largely matched model-based

expectations (Figure S4, Figure S10, and Figure S11). In addition, we repeated the same pattern selectivity analysis on the existing THINGS-fMRI dataset used in our initial in-vivo analyses. As expected by the use of an event-related design, the results were weaker (Figure S12) yet they largely mirrored those found in the present experiment, with differences including less pronounced effects for “rectilinear scenes”.

Discussion

Progress in understanding the function of visual cortex strongly benefits from methods that can move efficiently from broad hypothesis generation to controlled experimental validation. While previous closed-loop studies primarily optimized images to modulate a given neural population, they collapsed the underlying representational structure onto individual axes [5–8,37–39]. Complementing these approaches, the present framework provides a multidimensional account of the representational structure in visual brain regions. This general framework is broadly applicable for the study of human vision and beyond, as it can be extended to different visual brain regions, stimulus domains, measurement modalities, and even other cognitive domains [40]. Importantly, large parts of this framework could potentially be automated. Open neuroimaging datasets [29–31,33] already provide a basis for encoding-model training and hypothesis screening [4,9,41], and targeted experiments can be derived with limited manual intervention. More broadly, this points towards a streamlined model-guided framework for scientific discovery, in which in-silico models generate and prioritize hypotheses at scale, while human researchers guide experimental design, validation, and interpretation.

The present results also inform our understanding of the representational structure of the FFA. While the strongest recovered dimensions were face-related, our results revealed additional interpretable representational structure beyond this face selectivity. Notably, many recovered dimensions reflected additional information, including other high-level categories and visual features, consistent with previous work showing that FFA responses contain non-face-related information [15,22]. Our results make this structure explicit, providing an interpretable account of the dimensions that may underlie both face-specific and object-general coding in the FFA [42]. Of note, the in-silico approach revealed largely orthogonal dimensions for human and animal faces, which evoked distinct response patterns in vivo. While previous work revealed differential FFA responses to human and nonhuman faces, our findings demonstrate that these differences do not merely reflect graded responses along a common dimension [43,44] but may reflect distinct representational dimensions.

These results also point to several directions for extending and refining the closed-loop framework. For example, the in-silico candidate dimensions necessarily depend on the representational properties of the underlying ANN model [45,46], which may leave the characterization of a region's dimensions incomplete. While the successful in-vivo validation of many in-silico dimensions indicates that they are capable of genuinely capturing neural responses, future work could broaden the space of candidate dimensions by using different or potentially more diverse training data. Future extensions of this framework may also determine the factors driving the divergence between in-silico and in-vivo dimension responses, for instance by examining the role of model architecture, training data, or in-silico sampling. In addition, while our choice of naturalistic images prioritizes ecological validity, future work may instead focus on more controlled stimuli. While the targeted fMRI experiment was optimized for robust within-participant inference, the sample size (N = 6) limits conclusions about the population-level generalizability of weaker or non-significant effects, which may reflect interindividual variability or subject-specific differences in signal-to-noise ratio (Figure S13).

Future work could target larger population samples, allowing for more fine-grained evaluation of the generalizability of individual dimensions. The encoding-model training data, in-silico sampling, and in-vivo studies all drew from the THINGS image database. Future work may test whether the dimensions recovered in-silico generalize to novel image sets. Finally, this general closed-loop framework is not tied to a particular method for identifying candidate dimensions. While the results validate the similarity embedding approach used here, future work could explore alternatives based on symmetric non-negative matrix factorization [47] or multidimensional scaling [48].

Together, these findings demonstrate an efficient closed-loop framework that allows for a broad characterization of representations in visual cortex by combining the breadth of condition-rich, data-driven approaches with the rigor of experimental tests. Here, this approach provided an interpretable characterization of representational dimensions underlying FFA responses, illustrating how linking discovery and validation can accelerate progress in visual neuroscience.

Acknowledgements

We would like to thank Florian Mahner for providing source code for the sparse, nonnegative similarity embeddings as well as the Members of Hebart lab for helpful discussions. This work was supported by a Max Planck Research Group Grant from the Max Planck Society, the ERC Starting Grant project COREDIM (ERC-StG-2021-101039712), a LOEWE Start Professorship from the Hessian Ministry of Higher Education, Research, Science and the Arts, and by the Germany Research Foundation (Deutsche Forschungsgemeinschaft, DFG) through the Cluster of Excellence “The Adaptive Mind” (EXC 3066/1, project number 533717223). The funders had no role in conducting this research or preparing the manuscript.

Author contributions

Conceptualization: O.C., N.A.R.M., M.H.; Methodology: O.C., N.A.R.M., S.F., M.H.; Data curation: O.C., N.A.R.M.; Formal analysis: O.C.; Supervision: M.H.; Visualization: O.C.; Writing - original draft: O.C., N.A.R.M., M.H.; Writing - review & editing: O.C., N.A.R.M., S.F., M.H.

Declaration of interests

The authors declare no competing interests.

Declaration of generative AI and AI-assisted technologies

During the revision of the initial manuscript draft, the authors used AI-assisted technologies to improve the clarity of formulations and check grammar and spelling. The authors reviewed and edited the suggested content as needed and take full responsibility for the content of this publication.

STAR Methods

Key Resources Table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
THINGS image database	Hebart et al., 2019	doi:10.17605/OSF.IO/JUM2F; https://osf.io/jum2f
THINGS-data fMRI dataset	Hebart et al., 2023	OpenNeuro: doi:10.18112/openneuro.ds004192.v1.0.5; Figshare: doi:10.25452/figshare.plus.c.6161151
Targeted fMRI validation dataset	This paper	OpenNeuro: ds007791; doi:10.18112/openneuro.ds007791; private reviewer link provided during submission
Final in-silico FFA dimensions	This paper	Zenodo: doi:10.5281/zenodo.20343141
Human ratings of dimension labels	This paper	Zenodo: doi:10.5281/zenodo.20343141
Experimental stimulus lists and task-design files	This paper	Zenodo: doi:10.5281/zenodo.20343141
FFA encoding model code	Ratan Murty et al., 2021	https://osf.io/5k4ds
Software and algorithms		
OpenAI CLIP-ResNet50 model	Radford et al., 2021; OpenAI	https://github.com/openai/CLIP
Sparse similarity embedding model / SPoSE	Hebart et al., 2020; Zheng et al., 2019	https://github.com/ViCCo-Group/SPoSE; https://github.com/florianmahner/object-dimensions
Custom analysis code	This paper	Zenodo: doi:10.5281/zenodo.20343141
fMRIPrep v23.0.2	Esteban et al., 2019	https://fmriprep.org
pydeface	Gulban et al., 2019	doi:10.5281/zenodo.3524401
PsychoPy v2023.1.x	Peirce et al., 2019	https://www.psychopy.org; RRID:SCR_006571
pyfLoc	Salo et al. / NBCLab	https://github.com/NBCLab/pyfLoc
Stanford VPNL fLoc paradigm	Stigliani et al., 2015	https://github.com/VPNL/fLoc; RRID:SCR_022454
fracridge	Rokem and Kay, 2020	doi:10.1093/gigascience/giaa133

Resource Availability

Lead contact

Further information and requests for code and resources should be directed to and will be fulfilled by the lead contact, Oliver Contier (contier@cbs.mpg.de).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The THINGS image database used to generate in-silico predicted FFA responses is publicly available on the Open Science Framework (https://osf.io/jum2f).
- The THINGS-data fMRI dataset used for validation analyses is publicly available via OpenNeuro (https://doi.org/10.18112/openneuro.ds004192.v1.0.5) and Figshare (https://doi.org/10.25452/figshare.plus.c.6161151) and preprocessing and analysis code is publicly available via GitHub (https://github.com/ViCCo-Group/THINGS-data).
- The newly collected fMRI data from the targeted validation experiment will be made publicly available upon publication on OpenNeuro (doi:10.18112/openneuro.ds007791).
- The final embedding of in-silico FFA dimensions are available via github alongside all custom code used for embedding estimation, the fMRI experimental design, preprocessing, analyses, statistical testing, and figure generation (https://doi.org/10.5281/zenodo.20343141).

Experimental Model and Study Participant Details

FFA encoding model training data

The in-silico FFA response predictions were generated using an encoding model trained on a previously collected fMRI dataset reported by Murty et al. [9]. In that study, four healthy adult participants (2 female) completed five neuroimaging sessions at the Athinoula A. Martinos Imaging Center at MIT on a 3 Tesla Siemens MAGNETOM Prisma scanner. The main event-related experiment was conducted across four additional sessions and measured responses to 185 naturalistic color images, sampled predominantly from the THINGS database and comprising face, body, scene, and object images. Each image was presented on average 20 times per participant, yielding high-reliability image-level response estimates. Category-selective regions of interest, including the FFA, were localized in each participant using a dynamic localizer with video clips of faces, bodies, scenes, objects, and scrambled objects ($p < 10^{-7}$). Full details on MRI image acquisition and data preprocessing are reported in the original manuscript [9].

THINGS-data fMRI dataset

We used the publicly available fMRI dataset part of the THINGS-data collection to validate the representational dimensions identified in silico. A detailed description of this dataset is provided in the original publication [29]. In short, this dataset contains fMRI responses from three participants (2 female, 1 male; mean age at the beginning of the study: 25.33 years). Across 12 scanning sessions, each participant viewed 8,740 unique images of 720 object concepts sampled from the THINGS database [16]. For each object concept, 12 different image exemplars were presented across sessions. Each of these images was presented only once. The additional test set of 100 repeatedly presented images was omitted for the present analyses. Participants viewed images while in an oddball detection task while maintaining central fixation (500 ms stimulus presentation, 4s of fixation, SOA=4.5s). For our analyses, we used the already preprocessed and publicly available single-trial fMRI response estimates. We also used the predefined regions of interest for the FFA, which were obtained in each participant based on an independent functional localizer contrasting responses to faces and objects.

fMRI experiment participants

Six adult participants took part in the targeted fMRI validation experiment. Sample size was determined to support robust within-participant statistical inference rather than group-level effects, with 16 task runs across two sessions for high single-subject reliability. Participants had normal or corrected-to-normal vision and no self-reported history of neurological or psychiatric disorders. Further demographic information was as follows: 4 female, 2 male, mean age 26 years, age range 21-33 years. All participants provided written informed consent before participation and received monetary compensation. The study was approved by the ethics committee of the Medical Faculty of Leipzig University under protocol 056/21-ek and was conducted in accordance with the Declaration of Helsinki. MRI data were acquired at the Max Planck Institute for Human Cognitive and Brain Sciences on a 3 Tesla Siemens Connectom MRI scanner.

Method details

THINGS images

All in-silico and in-vivo responses were collected based on the THINGS image database, which contains 26,107 naturalistic images spanning 1,854 concrete, nameable object concepts [16]. The THINGS images depict objects in natural backgrounds and were manually curated to provide high-quality exemplars of a broad range of everyday objects. The object concepts were selected to comprehensively sample concrete, nameable things in the American English language and each concept is represented by multiple images. To generate in-silico FFA responses, all 26,107 THINGS images were used as inputs to the encoding model. The same image database served as the stimulus pool for the targeted fMRI experiment (see Targeted fMRI experiment design).

Obtaining predicted responses from the in-silico encoding model

To generate in-silico FFA responses, each THINGS image was passed through an existing image-computable encoding model trained to predict voxel-wise fMRI responses in the FFA [9,49]. The encoding model was built using a OpenAI CLIP-ResNet50 backbone architecture. Specifically, feature activations were extracted from the block4-1-conv2 layer and mapped to voxel-wise FFA responses using regularized ridge regression. The model was trained on voxel-wise responses from all four subjects, and the ridge regularization parameter was fixed at 0.01. After fitting, all model parameters and hyperparameters were frozen. The original encoding model [9] yielded an average prediction performance of r = 0.92 and r = 0.91 within left and right FFA respectively. Because the resulting model is fully image-computable, we subsequently passed all THINGS images through the encoding model to obtain predicted voxel-wise FFA activations for each image. These predicted activations were then used as input to the embedding analysis to derive the inferred representational dimensions.

Generating simulated odd-one-out triplets

We converted the predicted FFA response patterns into simulated odd-one-out judgments as a basis for identifying latent representational dimensions. First, we standardized the in-silico FFA responses by converting each in-silico voxel responses to zero mean and unit variance to account for outlier voxels, and in-silico voxels were concatenated across participants. We then randomly sampled 4 million image triplets, computed pairwise similarities based on the euclidean distance between their response patterns. The image which was least similar to the other two was assigned as the odd one out.

Similarity embedding model

We fit a sparse, non-negative embedding model originally developed for identifying representational dimensions from behavioral odd-one-out choices to the simulated odd-one-out triplets [17,18]. The approach is conceptually related to other dimensionality reduction techniques such as symmetric non-negative matrix factorization [47] or multidimensional scaling [48], as the learned non-negative embedding implicitly factorizes an image-by-image similarity matrix However, rather than directly reconstructing numerical similarities, the embedding model predicts the probability of a given odd-one-out choice by applying a softmax nonlinearity to the pairwise dot-product similarities.

In this embedding model, each image is represented by non-negative weights along a set of latent dimensions. Similarity between two images is modeled as the dot product of their dimension-weight vectors, and the model is optimized to predict the odd-one-out choices derived from the in-silico FFA response patterns. Triplets were split into training and test sets, with 90% used for model fitting and 10% held out for evaluating embedding performance. The model was trained by minimizing cross-entropy loss on the simulated triplet choices with an L1 penalty to produce sparse dimension weights. Model weights were randomly initialized with 120 initial dimensions and near-zero dimensions were pruned after training. Hyperparameters including the L1 penalty and learning rate were selected via a grid search and evaluated based

on cross-validated prediction accuracy, resulting in a final configuration of β = 0.0003 and learning rate = 0.001. The final embedding consisted of one vector of dimension weights per THINGS image, yielding an image-by-dimension embedding matrix.

Reproducible group-level embedding

To identify dimensions that consistently occur across random initializations, we trained the model 20 times and retained only reproducible dimensions. Specifically, we quantified dimension-wise split-half reproducibility by matching each dimension to its maximally correlated dimension in another run, and discarded dimensions with an average Fisher-z corrected Pearson correlation below r = 0.6. We then obtained an aggregated embedding of group-level dimensions across the 4 in-silico subject response datasets. To that end, we performed agglomerative clustering on the z-scored set of pooled dimensions and selected the number of clusters based on a joint inspection of silhouette, Calinski-Harabasz, and Davies-Bouldin scores. The group level embedding was defined based on the resulting 17 median cluster centroids. Finally, we retained only dimensions that could be reliably predicted from the individual in-silico responses of all four subjects (R^2 > 0.6), resulting in the final set of 15 representational dimensions.

Dimension labeling by human raters

To provide a basis for interpreting these representational dimensions, we asked human raters (N = 12) to provide labels for each dimension. Independent volunteer raters were recruited anonymously from the local academic community and were not involved in the analyses. This study was approved by the local ethics committee of the Medical Faculty of the University Medical Center Leipzig (157/20-ek). For each dimension, participants were shown the 64 top-ranked images for each dimension in a 8 by 8 grid and were instructed to provide three words that best describe what these images have in common. The resulting labels were corrected for spelling errors, and near-duplicates were resolved manually. These preprocessed labels were then used to produce word clouds where the size of each term is proportional to its frequency of occurrence. The final dimension names were then assigned by the authors based on these independent labels.

THINGS-data validation analysis

To validate the in-silico dimensions in-vivo, we used the publicly available single-trial response estimates of the THINGS-data fMRI dataset, restricting our analysis to FFA voxels based on the available regions of interest [29]. We constructed a linear encoding model in which image-wise voxel responses were predicted as a weighted linear combination of the corresponding 15 in-silico dimension weights. As a baseline reference, we constructed an analogous one-predictor encoding model from manually obtained, binary face labels indicating whether each image contained a face. Both the dimension-based and the face-label encoding models were fit using fractional ridge regression [50], selecting independent regularization parameters for each voxel (range from 0.05 to 1.0 in steps of 0.05). To optimize regularization parameters and estimate predictive performance, we used a leave-3-sessions-out cross-validation scheme.

Predictors and responses were z-scored independently in each training and test fold to avoid leakage. Optimal fractional ridge parameters were selected based on the pooled R^2 score within a given training fold. The final encoding score was the cross-validated R^2 pooled across test folds, normalized by the single-trial noise ceiling (see Quantification and Statistical Analysis). To quantify the encoding score of each dimension individually, we repeated the encoding analysis using each dimension separately as the sole predictor. In addition to the encoding score, we quantified the contribution of each individual dimension to FFA responses by computing dimension-wise response amplitudes: For each dimension, we identified the 75 images with the highest and the 75 images with the lowest weights along that dimension and computed the difference between their mean z-scored FFA responses. To generate model-expected response amplitudes in silico, we repeated the response-amplitude analysis on each model participant's predicted responses to the images used in THINGS-fMRI.

Targeted fMRI experiment design

The targeted experiment used 16 conditions: one condition for each of the in-silico dimensions plus a control condition composed of images with low values across all 15 dimensions (cf. Methods Details, THINGS images). To enable assessing robust effects independent category-specific effects, two independent image sets were created for each condition. To that end, images were rank-ordered by their dimension weights and assigned to the two sets in an alternating fashion, with the constraint that no object concept appeared in both sets. Image sets were not constrained to be mutually exclusive for the different dimensions. The control condition was constructed by drawing an equal number of images from the lower end of each dimension's weight distribution while ensuring that no image appeared among the top 5,000 images of any dimension. The two image sets of each experimental condition were assigned to odd and even runs respectively. The fMRI runs used a block design, with each run presenting each of the 16 condition blocks once. Each block contained 15 images presented for 0.8 s with a 0.2 s inter-stimulus interval (15 s per block; no inter-block blank). Additional blank blocks were presented to allow estimation of the signal baseline. The order of blocks within a run was pseudo-randomized under the constraint that baseline blocks could not occur in the first or last position and could not occur consecutively. Stimuli were presented with PsychoPy at 8.4° of visual angle on a rear-projection screen. Each participant completed 16 runs across two scanning sessions (8 runs per session) and the two image sets were assigned to odd and even runs respectively. Each run started and ended with 10 s of fixation, for a total run duration of ~5 min. Participants performed a simple fixation detection task in which they were asked to respond when the viewed fixation cross (0.35° visual angle) changed color from white to red for a duration of 0.5 s in uniform intervals from 3 to 10 s. In addition to the 16 task runs, two functional localizer runs were acquired in each session (4 runs total per participant; 4 min each) using the Stanford VPNL fLoc localizer paradigm [51]. Each run consisted of 36 mini-blocks of 6 s, with each block containing 12 grayscale images from one of five visual categories (adult faces, headless bodies, written words, cars, houses, and phase-scrambled objects) presented for 0.4 s with a 0.1 s inter-stimulus interval. In the localizer runs, participants performed a one-back task in which they pressed a button whenever an image was repeated on consecutive trials.

MRI acquisition

MRI data were acquired on a 3 T Siemens MAGNETOM Connectom scanner at the Max Planck Institute for Human Cognitive and Brain Sciences (Leipzig, Germany) using a 32-channel head coil. Functional images were acquired with a multiband T2*-weighted gradient-echo EPI sequence at 2 × 2 × 2 mm isotropic resolution (FOV 204 × 204 mm, matrix 102 × 102, 81 axial slices, multiband factor 3, TR = 2.0 s, TE = 24 ms, flip angle = 80°). Additional scanning included a T1-weighted anatomical scan (MPRAGE; 1 mm isotropic, TR = 2.3 s, TE = 2.91 ms, flip angle = 9°) acquired in each participant as well as gradient-echo field maps acquired in each session for susceptibility-distortion correction.

MRI data preprocessing

fMRI data were preprocessed with fMRIPrep v23.0.2 [52]. In short, functional volumes were motion-corrected, susceptibility-distortion corrected using the acquired field maps, slice-timing corrected, and co-registered and resampled to each participant's T1-weighted anatomical image. To remove identifiable facial features while preserving brain tissue, T1-weighted anatomical images were defaced using pydeface [53]. No spatial smoothing was applied to the main task data to preserve fine-grained spatial information for the pattern-similarity analyses. To estimate response to the in-silico dimensions, we fit a general linear model to each functional task run [54]. To that end, we constructed run-wise design matrices by convolving the block design with a canonical hemodynamic response function and including confound time series estimated during preprocessing (six rigid-body motion parameters, framewise displacement, plus the first six anatomical CompCor components; [55]). We then estimated the beta weights representing the response to each condition and obtained contrasts between responses to the dimension-evoked conditions and the control condition, providing an estimate of the selective response to each in-silico dimension. Regions of interest for the FFA were defined in each participant by analyzing the functional localizer runs with a general linear model: Condition-wise boxcar regressors were convolved with a canonical double-gamma HRF and fit to the preprocessed fMRI timeseries, yielding parameter estimates for each condition. We then aggregated these run-level estimates across runs in a fixed-effects second-level analysis and computed parameter estimate contrasts (faces > all other conditions). The resulting statistical maps were corrected for multiple comparisons based on a cluster-forming threshold of $z > 3.7$ ($p < 0.001$). The results then served as the basis for manually delineating face-selective clusters in the fusiform gyrus, providing regions of interest for the FFA. Note that this localizer protocol differs from the dynamic face localizer used in the encoding model training dataset, which may affect voxel selection. We additionally defined the extrastriate body area (EBA) and parahippocampal place area (PPA) using the localizer contrasts (bodies > all other conditions, places > all other conditions). Primary visual cortex (V1) was defined using the Juelich V1 atlas [56] intersected with visually responsive voxels (all conditions > blank). Within each participant, these ROIs were matched to the FFA in voxel count and reliability distribution across voxels.

Quantification and Statistical Analysis

Encoding-model performance in THINGS-data

Encoding-model performance was quantified per voxel as the coefficient of determination (R^2) of the cross-validated prediction pooled across the four leave-3-sessions-out folds. To control for differences in trial-wise response reliability across voxels, R^2 values were normalized by the single-trial noise ceiling provided in the THINGS-data collection, yielding a normalized measure of explained variance. Voxels with a noise ceiling below $R^2 = 0.01$ were excluded from the analysis. For each participant, we report the mean encoding scores across FFA voxels as well as the distribution of voxel-wise R^2 values. We compared the explanatory power of the full 15-dimension embedding model to that of a one-predictor face-label baseline model fit with the identical pipeline. Dimension-wise response amplitudes (see Method Details) were averaged across FFA voxels per participant, and 95% confidence intervals were obtained by bootstrap resampling across voxels. Per-subject contrasts were averaged across FFA voxels, and 95% confidence intervals were obtained by bootstrap resampling across voxels. To obtain a dimensionality-matched statistical baseline, we permuted image assignments in the 15-dimensional design matrix, refit the full pipeline 200 times, and defined the chance interval as the 2.5th–97.5th percentiles of median noise-ceiling-normalized R^2 across FFA voxels with noise ceiling >1%.

Pattern selectivity analysis

To assess the selectivity of FFA response patterns evoked by each in-silico dimension, we computed the within- and between-condition similarity between run-wise FFA activation patterns obtained from the first-level GLM analysis of the targeted fMRI experiment. To focus the analysis on dimension-evoked responses and exclude activation shared with generic object viewing, we based the pattern analysis on the contrast between dimension-wise betas and the control condition. FFA voxels were z-scored across conditions to account for outliers and voxels with a mean split-half correlation below $r = 0.2$ were excluded to mitigate the effect of noise on the analysis. Voxel-wise reliability scores were obtained by correlating each voxel's 15 dimension > control responses between averages of the eight odd and eight even runs. Pattern similarity was quantified using the Euclidean distance, which we sign-inverted so that higher values reflect higher similarity. Within-condition similarity for a given dimension was computed across the two independent image sets (i.e. odd vs. even runs), eliminating any contribution of stimulus-specific or object-specific features that did not generalize across these image sets. Between-condition similarity was computed analogously, as the average across-set similarity between activity patterns evoked by a given dimension and those evoked by every other dimension. The pattern selectivity of each dimension was then defined as the difference between the within-condition similarity and the mean between-condition similarity. Statistical significance was assessed at the single-subject level with a permutation test in which condition labels in odd runs were shuffled within one image set, breaking the correspondence between conditions while preserving overall response structure. For each participant and each dimension, pattern selectivity was repeatedly computed for 10,000 random permutations, yielding a null distribution from which we obtained right-tailed p-values. A dimension was

considered statistically significant in a given participant if $p < 0.05$. Statistical inference was performed within-participant, as the sample size does not support stable group-level tests. Bar plots show means with 95% confidence intervals alongside descriptive group means. We applied the same pattern selectivity analysis to EBA, PPA, and V1 (Figure S2). To obtain model-expected pattern selectivity scores, we repeated the pattern selectivity analysis to the in-silico predicted FFA response patterns to the images selected for the targeted experiment. Voxel responses were z-scored across images, averaged across the 15 images of each dimension image set, subtracted the corresponding control pattern. We then applied the same negative-Euclidean selectivity analysis. To define a random-stimulus baseline, we repeated the analysis for 10,000 pseudo-experiments containing nonexperimental images. We also quantified pattern selectivity for each dimension in the THINGS-fMRI responses. To that end, we uniquely assigned 12 training images to each dimension from images for which that dimension had the largest weight. Images were rank-alternated into two sets of six, and 12 control images were sampled from unused images outside the top 500 of every dimension. For FFA voxels with noise ceiling >1%, we then computed pattern selectivity as described above. To assess overlap between human- and animal-face responses, their dimension > control parameter estimates were averaged across the 16 runs and correlated across FFA voxels within participants. To assess topographic organization, we extracted fixed-effects z-statistics separately from left and right FFA, percentile-ranked voxel positions along the anterior-posterior, superior-inferior, and medial-lateral axes, z-scored responses within hemisphere, and averaged them in eight equal bins. Bins with fewer than five voxels were omitted. To compare dimension responses with independent face selectivity, we correlated each fixed-effects dimension > control z-map with the faces > all localizer z-map across bilateral FFA voxels within each participant. To visualize the selected stimuli in predicted FFA response space, we fit a PaCMAP embedding to the predicted responses to all 26,107 THINGS images. Responses were combined across the four encoding-model subjects with equal mean-squared-distance weighting, reduced to 100 principal components, and embedded in two dimensions using PaCMAP with Euclidean distance (30 neighbors) [57]. The two experimental image sets for each dimension were overlaid on this embedding. To directly compare in-silico predictions and in-vivo results, we compared observed response amplitudes from the THINGS-fMRI analysis and pattern selectivity from the targeted experiment with their respective model predictions described above (Figure S4). For each measure, values were averaged across participants, computed Pearson correlations across the 15 dimensions and fitted ordinary least-squares regression lines to visualize their relationship.

Supplemental Information

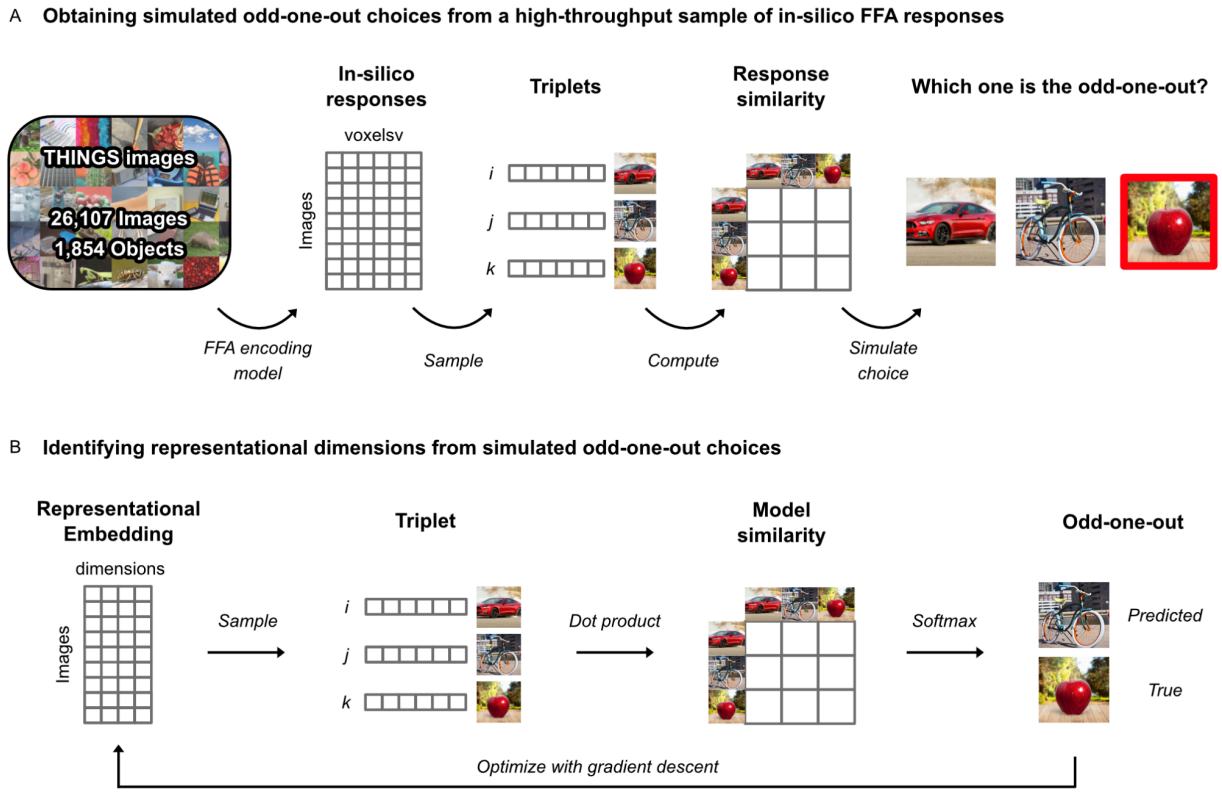


Figure S1. Identifying representational dimensions from in-silico brain responses A: High-throughput sampling of in-silico responses and simulated odd-one-out choices. The encoding model predicts voxel-wise FFA responses to 26,107 images of 1,854 objects taken from the THINGS database. For each of 4 million sampled image triplets, we compute response similarities using Euclidean distance and generate simulated odd-one-out choices. B: Identifying candidate representational dimensions. A sparse, non-negative embedding assigns each image weights along latent dimensions and is optimized to reproduce the odd-one-out choices determined from the in-silico responses.

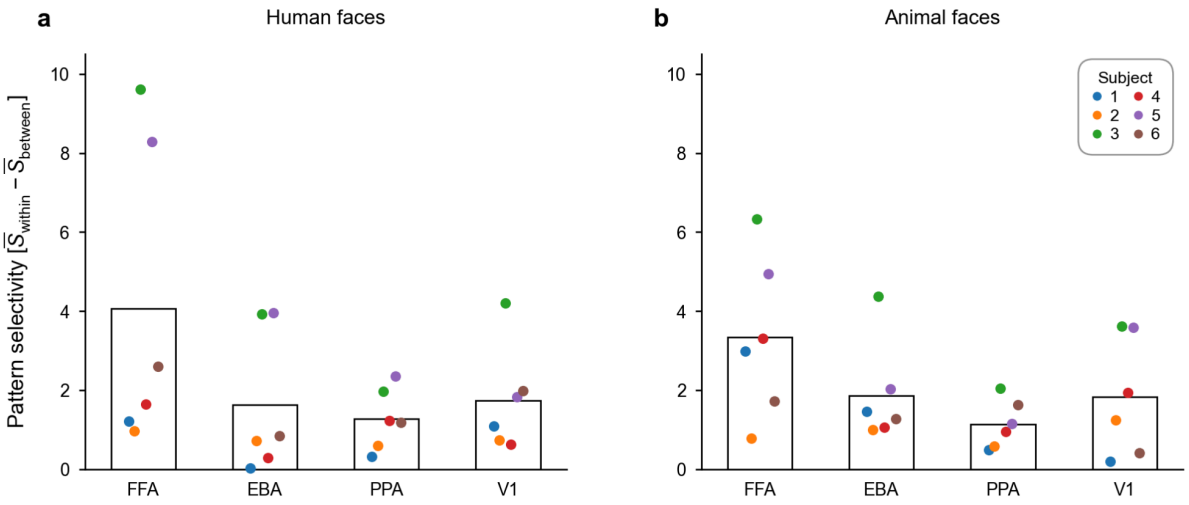


Figure S2. Pattern selectivity of the human- and animal-face dimensions across visual regions. Note that, due to differences in signal strength and anatomy between regions, only the presence of discriminative information can be interpreted, but not the degree to which a dimension is expressed in a region. A: Results for the “Human faces” dimension. B: Results for the “Animal faces” dimension. Pattern selectivity was assessed analogously to the main analysis (Figure 3). Regions of interest were matched to the FFA in voxel count and voxels were selected to approximate the reliability distribution in FFA within each participant.

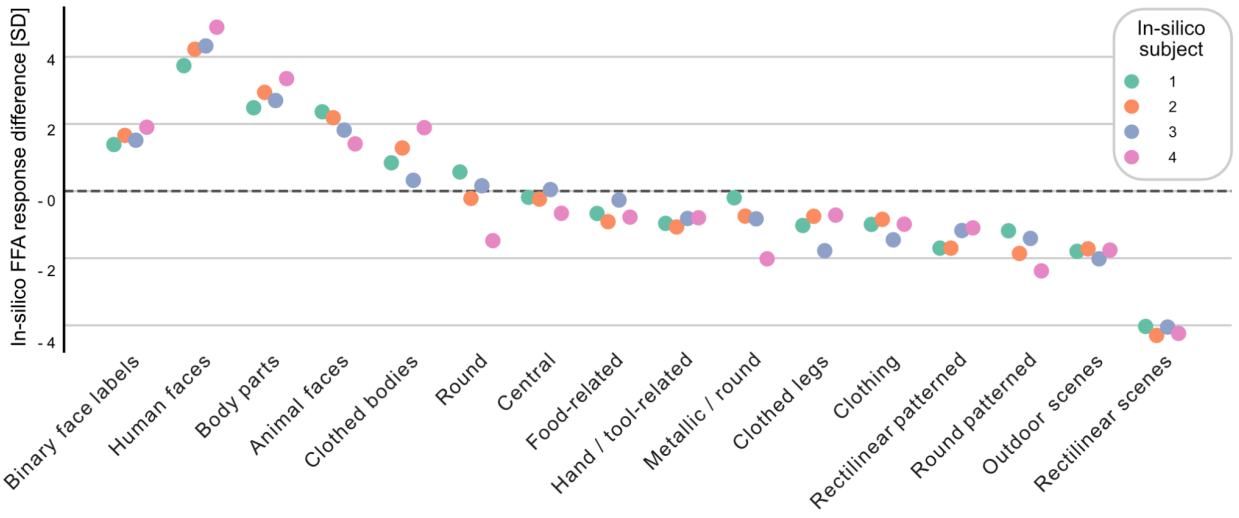


Figure S3. In-silico prediction of FFA responses to each dimension. Analogous to the analysis of the THINGS-fMRI dataset, we computed the difference between predicted in-silico responses to the 75 images with the highest vs. lowest weights on each dimension. Points indicate mean values across FFA for each individual subject.

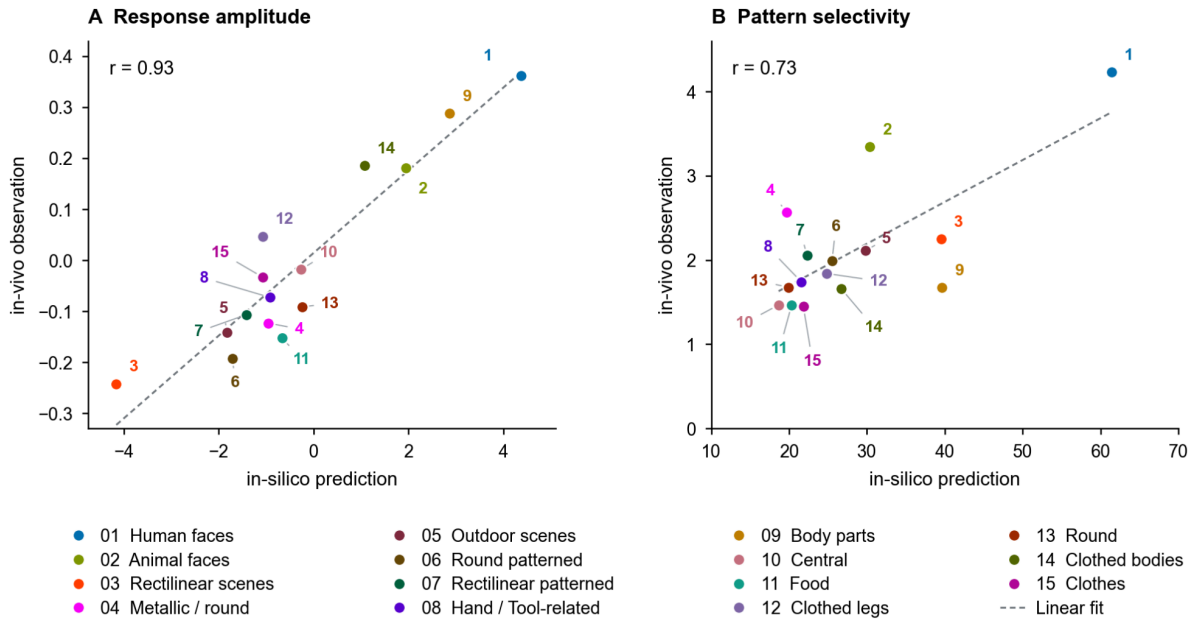


Figure S4. Comparison of in-silico predictions and in-vivo FFA responses. A. Response amplitudes to each dimension obtained from the THINGS-fMRI dataset (x-axis) compared to the in-silico FFA responses (y-axis). B. Analogous comparison of pattern selectivity per dimension. Each point shows one dimension. All effects were averaged across the respective dataset participants. Dashed lines show ordinary least-squares fit. However, note that the in-silico estimates of response amplitudes and pattern selectivity are circular, as they were quantified using the same data from which the candidate dimensions were derived.

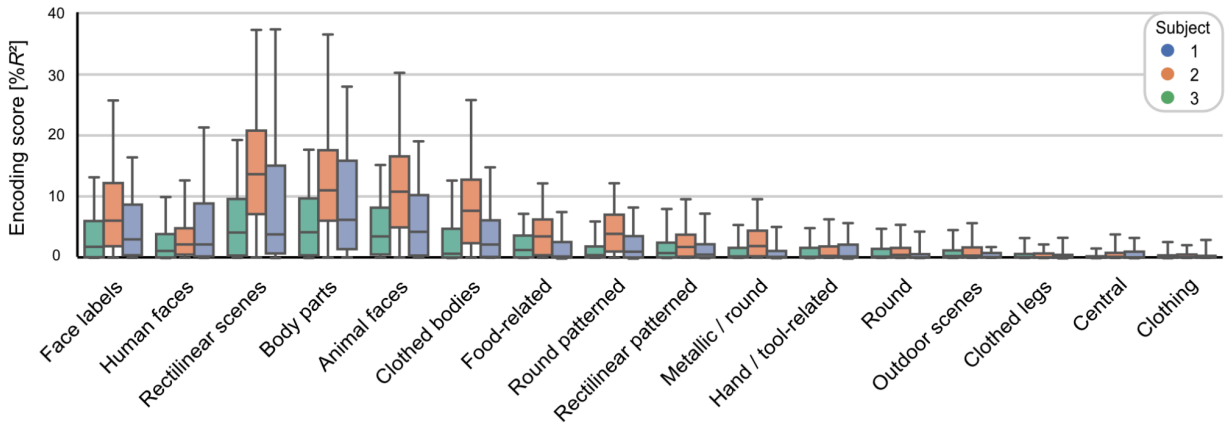


Figure S5. Encoding score of individual in-silico dimensions and binary face labels. Boxplots illustrate distribution across FFA voxels. Encoding scores are measured as percentage of explained variance relative to the noise ceiling.

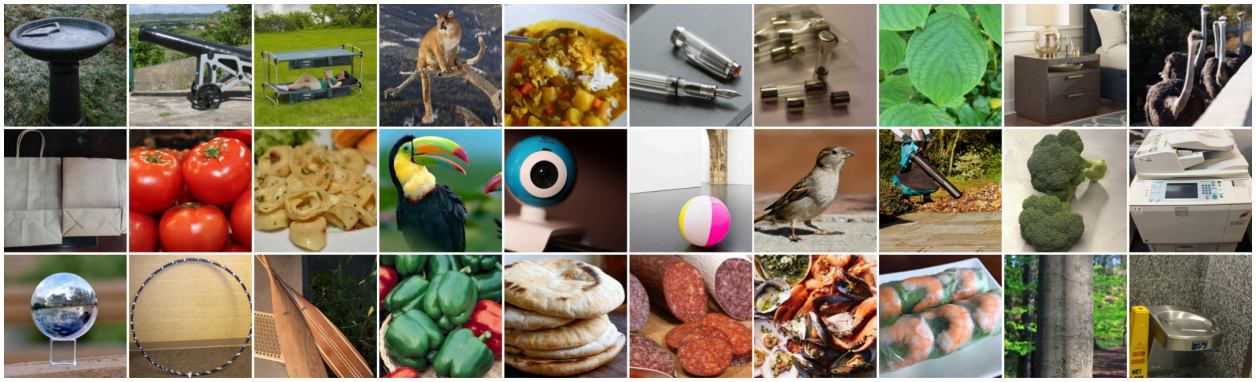


Figure S6. Example images used in the control condition of the fMRI experiment. Images were sampled from the lower end of each in-silico dimension's weight distribution, excluding images among the 5,000 top-ranked exemplars of any dimension. Images were assigned to the two image splits used in odd and even functional runs in an alternating fashion.

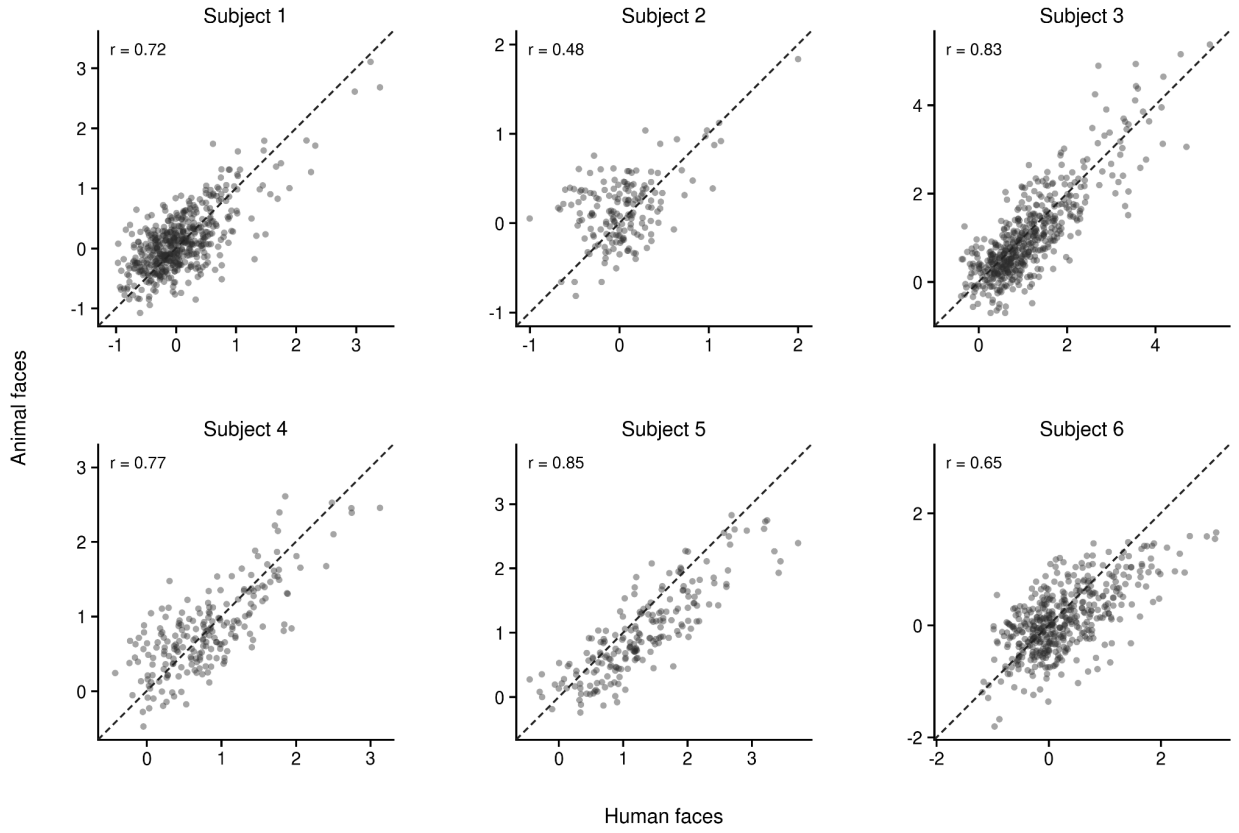


Figure S7. Voxelwise FFA responses to human and animal faces. Each point is one FFA voxel. X- and y-axes show mean responses to the human-face and animal-face dimensions respectively. The dashed line indicates the identity. Pearson's r is shown in each panel.

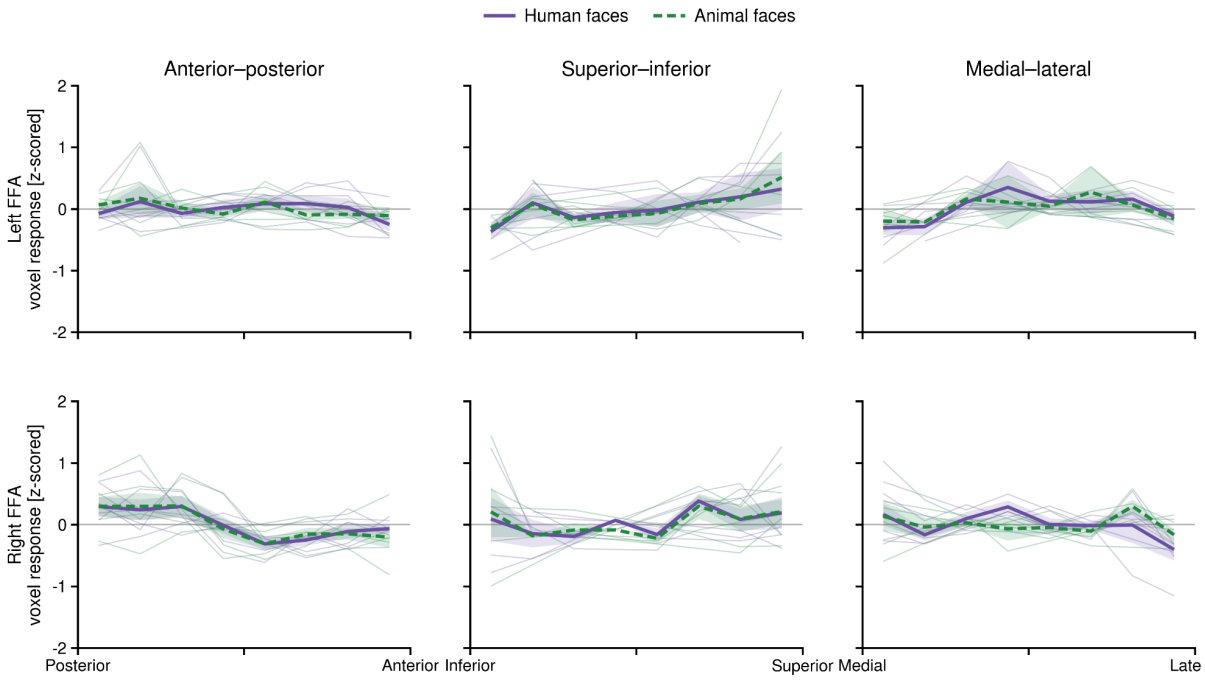


Figure S8. Topographic organization of human- and animal-face responses in FFA. Voxelwise responses to human and animal faces were extracted from individually defined left and right FFA and binned by relative anatomical position along the anterior-posterior, superior-inferior, and medial-lateral axes. Responses were z-scored within each hemisphere. Thin lines show individual participants. Thick lines and shading show mean ± SEM.

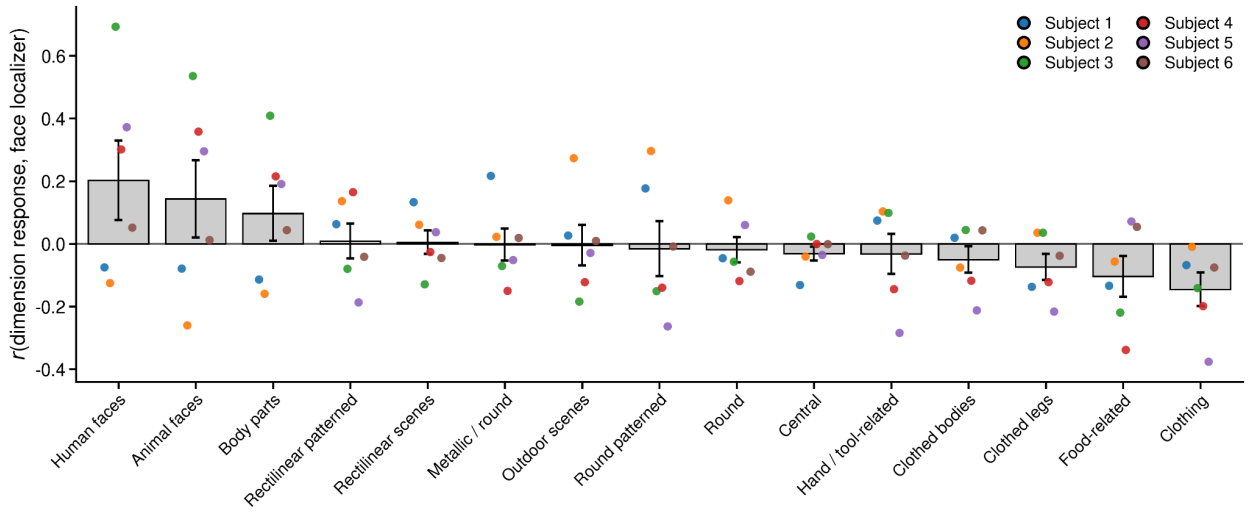


Figure S9. Voxel-wise correlation between dimension responses and the face localizer in FFA. For each subject, Pearson's r was computed across FFA voxels between the block-design dimension > control z-map and the faces > all localizer z-map. Bars show the mean correlation across subjects (n = 6), error bars indicate SEM, and colored dots show individual subjects.

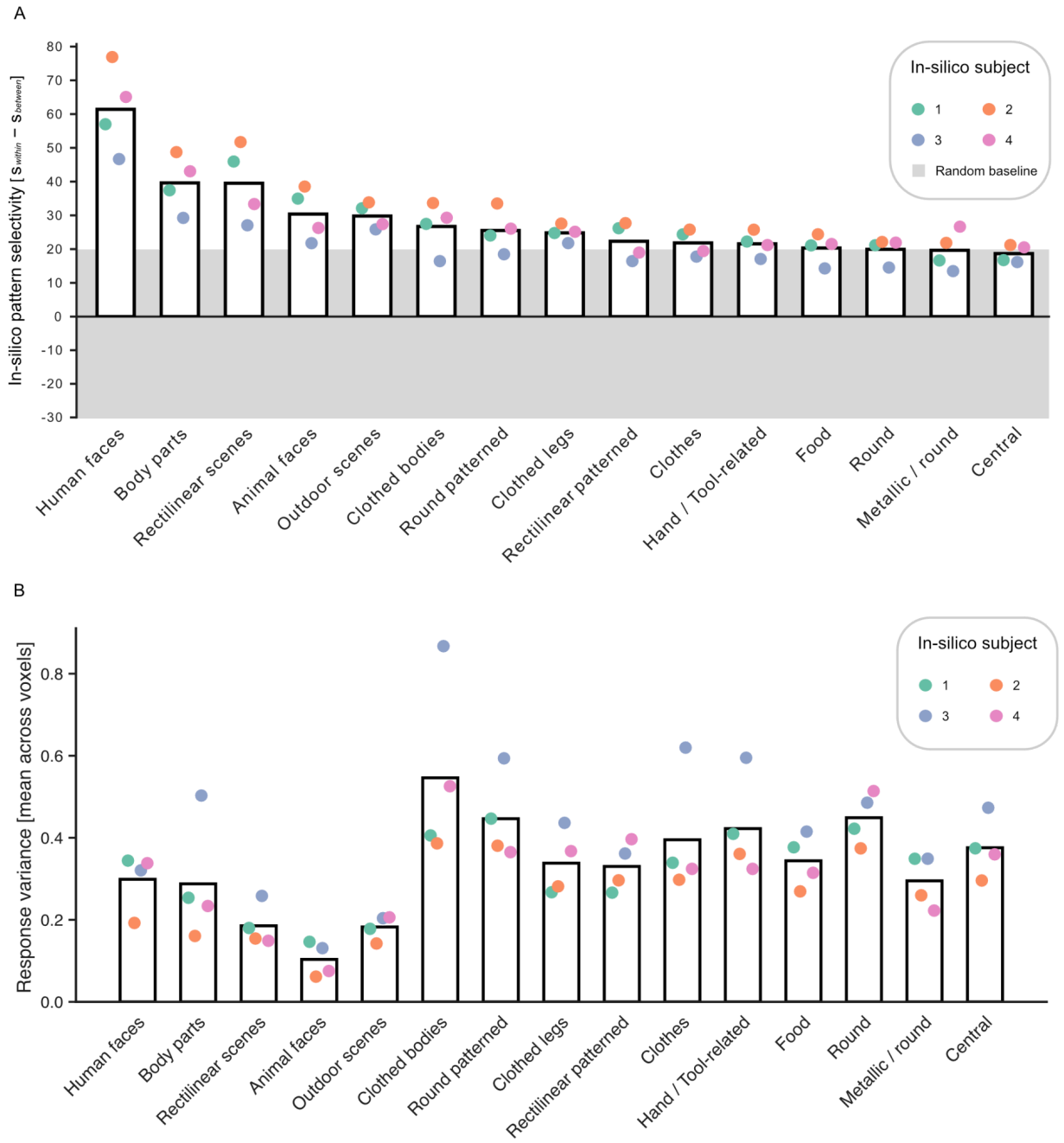


Figure S10. In-silico prediction of pattern selectivity for the targeted fMRI experiment. A: FFA response patterns were predicted using subject-specific encoding models fitted to four subjects. Pattern selectivity reflects the degree to which images of a given candidate dimension evoke similar response patterns, compared to response similarity between dimensions. Analogous to the main experiment, within-dimension pattern similarity was computed across two independent image sets. Bars show the mean across subjects and colored dots reflect individual subjects. The shaded area shows the central 95% of a random-stimulus baseline (10,000 repetitions). B: Variance across in-silico responses to the 30 images selected for each

dimension in the experiment. Dimensions are rank-ordered based on the pattern selectivity presented in panel A to allow for direct comparison

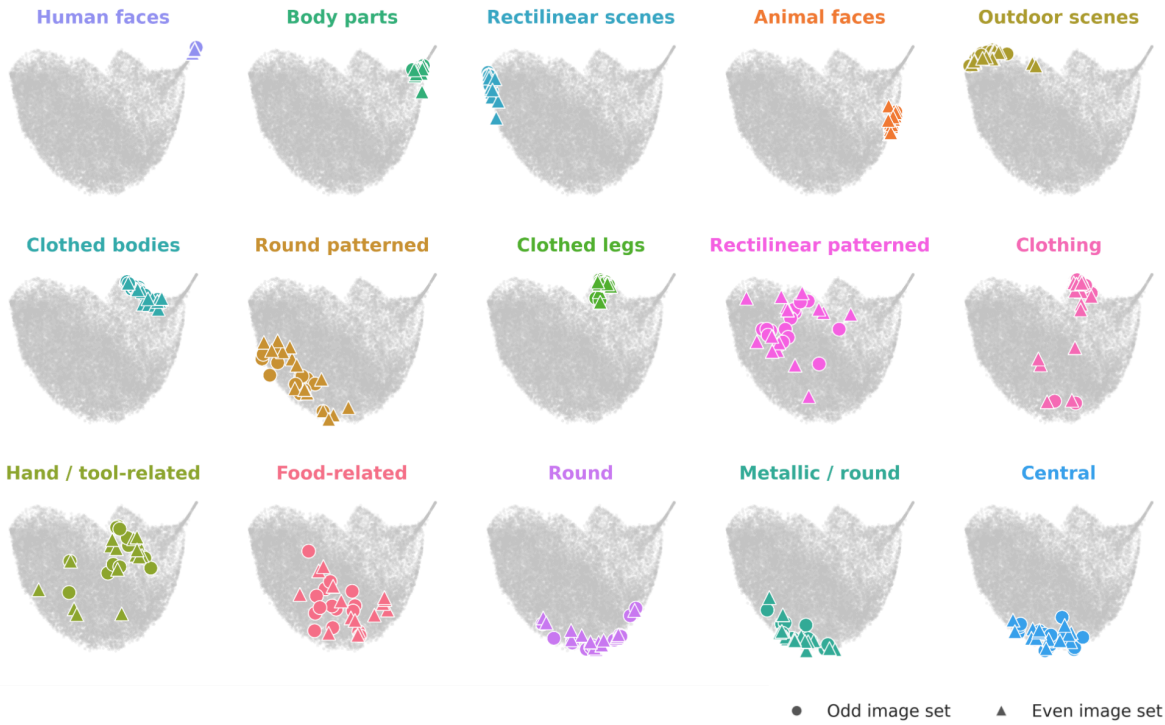


Figure S11. Clustering of dimension-selected images in predicted FFA response space. PaCMAP visualization of predicted FFA responses to all THINGS images (gray). Colored points show the experimental images selected for each dimension in the targeted fMRI experiment. Circles and triangles indicate the odd and even image sets, respectively. Panels are ordered by decreasing mean in-silico pattern selectivity to allow direct comparison.

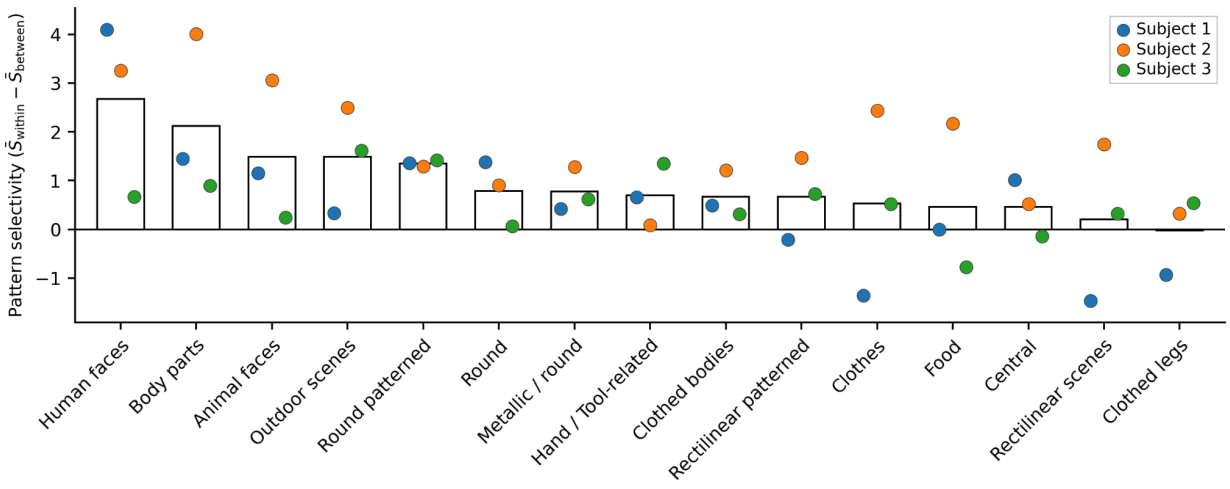


Figure S12. Pattern selectivity of FFA response patterns evoked in the THINGS-fMRI dataset by each hypothesized dimension. For each dimension, responses to 12 uniquely assigned images were split into two halves and averaged within each half. Pattern selectivity was computed as the difference between average within-dimension and between-dimension similarity. Samples show subject-wise pattern selectivity, and bars the mean across subjects.

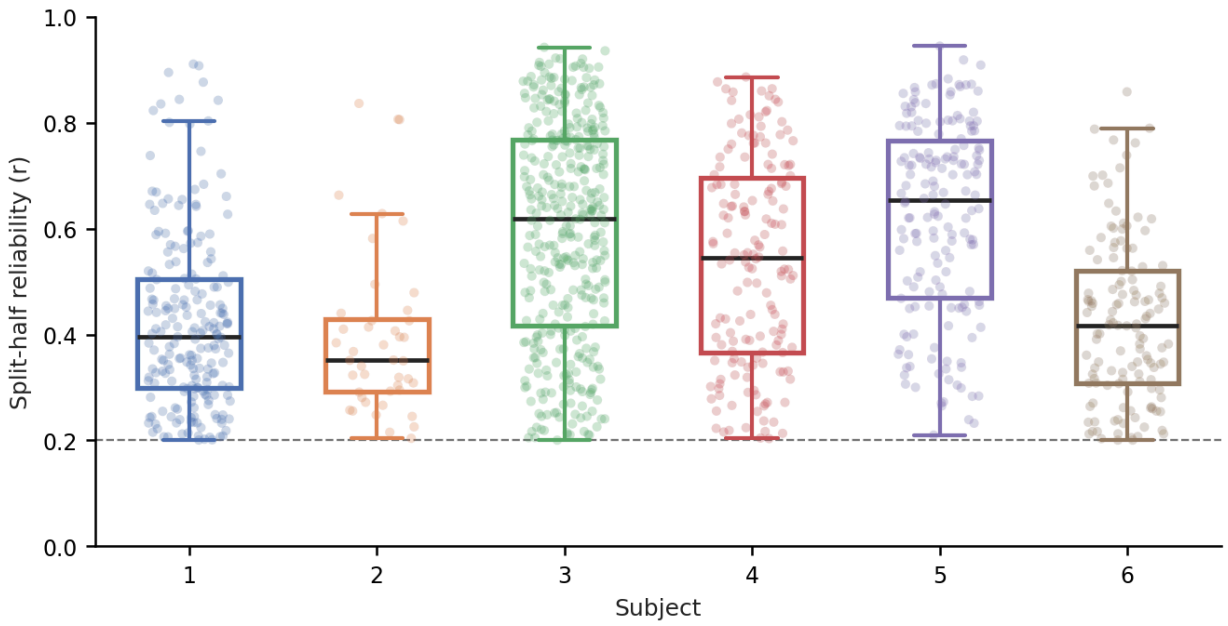


Figure S13. Subject-wise reliability of FFA responses in the targeted fMRI experiment. Reliability was quantified separately for each voxel as the Pearson correlation between response profiles estimated from independent halves of the data. Points represent individual voxels, boxplots show the median, interquartile range, and 1.5 × interquartile-range. The dashed line marks the reliability threshold of (r = 0.2) applied to the pattern selectivity analysis.

References

1. Naselaris, T., Allen, E., and Kay, K. (2021). Extensive sampling for complete models of individual brains. Current Opinion in Behavioral Sciences *40*, 45–51.

2. Chen, X., Beck, J.M., and Pearson, J.M. (2017). Neuron's eye view: Inferring features of complex stimuli from neural responses. PLoS Comput. Biol. *13*, e1005645.

3. Golan, T., Raju, P.C., and Kriegeskorte, N. (2020). Controversial stimuli: Pitting neural networks against each other as models of human cognition. Proc. Natl. Acad. Sci. U. S. A. *117*, 29330–29337.

4. Conwell, C., Prince, J.S., Kay, K.N., Alvarez, G.A., and Konkle, T. (2024). A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. Nat. Commun. *15*, 9383.

5. Bashivan, P., Kar, K., and DiCarlo, J.J. (2019). Neural population control via deep image synthesis. Science *364*. https://doi.org/10.1126/science.aav9436.

6. Ponce, C.R., Xiao, W., Schade, P.F., Hartmann, T.S., Kreiman, G., and Livingstone, M.S. (2019). Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. Cell *177*, 999-1009.e10.

7. Walker, E.Y., Sinz, F.H., Cobos, E., Muhammad, T., Froudarakis, E., Fahey, P.G., Ecker, A.S., Reimer, J., Pitkow, X., and Tolias, A.S. (2019). Inception loops discover what excites neurons most using deep predictive models. Nat. Neurosci. *22*, 2060–2065.

8. Gu, Z., Jamison, K.W., Khosla, M., Allen, E.J., Wu, Y., St-Yves, G., Naselaris, T., Kay, K., Sabuncu, M.R., and Kuceyeski, A. (2022). NeuroGen: Activation optimized image synthesis for discovery neuroscience. Neuroimage *247*, 118812.

9. Ratan Murty, N.A., Bashivan, P., Abate, A., DiCarlo, J.J., and Kanwisher, N. (2021). Computational models of category-selective brain regions enable high-throughput tests of selectivity. Nat. Commun. *12*, 5540.

10. Kanwisher, N., McDermott, J., and Chun, M.M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. J. Neurosci. *17*, 4302–4311.

11. Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., and DiCarlo, J.J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. U. S. A. *111*, 8619–8624.

12. Wang, A.Y., Kay, K., Naselaris, T., Tarr, M.J., and Wehbe, L. (2023). Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. Nat. Mach. Intell., 1–12.

13. Kanwisher, N., and Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. Philos. Trans. R. Soc. Lond. B Biol. Sci. *361*, 2109–2128.

14. Grill-Spector, K., Knouf, N., and Kanwisher, N. (2004). The fusiform face area subserves face perception, not generic within-category identification. Nat. Neurosci. *7*, 555–562.

15. Vinken, K., Prince, J.S., Konkle, T., and Livingstone, M.S. (2023). The neural code for “face cells” is not face-specific. Sci. Adv. *9*, eadg1736.

16. Hebart, M.N., Dickter, A.H., Kidder, A., Kwok, W.Y., Corriveau, A., Van Wicklin, C., and Baker, C.I. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. PLoS One *14*, e0223792.

17. Hebart, M.N., Zheng, C.Y., Pereira, F., and Baker, C.I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. Nature Human Behaviour *4*, 1173–1185.

18. Zheng, C.Y., Pereira, F., Baker, C.I., and Hebart, M.N. (2019). Revealing interpretable object representations from human behavior. arXiv [stat.ML], 1–16. https://doi.org/10.48550/arXiv.1901.02915.

19. van Dyck, L.E., Hebart, M.N., and Dobs, K. (2026). Multidimensional feature tuning in category selective areas of human visual cortex. J. Neurosci. *46*, e0038262026.

20. Contier, O., Baker, C.I., and Hebart, M.N. (2024). Distributed representations of behaviour-derived object dimensions in the human visual system. Nat. Hum. Behav., 1–15.

21. van Bree, S., and Hebart, M.N. (2026). Shared and distinct high-dimensional object spaces in human and macaque inferotemporal cortex. bioRxiv. https://doi.org/10.64898/2026.05.20.724014.

22. Long, B., Yu, C.-P., and Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. Proc. Natl. Acad. Sci. U. S. A. *115*, E9015–E9024.

23. Bracci, S., Ritchie, J.B., and de Beeck, H.O. (2017). On the partnership between neural representations of object categories and visual features in the ventral visual pathway. Neuropsychologia *105*, 153–164.

24. Bracci, S., and Op de Beeck, H. (2016). Dissociations and Associations between Shape and Category Representations in the Two Visual Pathways. J. Neurosci. *36*, 432–444.

25. Coggan, D.D., Liu, W., Baker, D.H., and Andrews, T.J. (2016). Category-selective patterns of neural response in the ventral visual pathway in the absence of categorical information. Neuroimage *135*, 107–114.

26. Coggan, D.D., Baker, D.H., and Andrews, T.J. (2019). Selectivity for mid-level properties of faces and places in the fusiform face area and parahippocampal place area. Eur. J. Neurosci. *49*, 1587–1596.

27. Rice, G.E., Watson, D.M., Hartley, T., and Andrews, T.J. (2014). Low-level image properties of visual objects predict patterns of neural response across category-selective regions of the ventral visual pathway. J. Neurosci. *34*, 8837–8844.

28. Martin, C.B., Douglas, D., Newsome, R.N., Man, L.L., and Barense, M.D. (2018). Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. Elife *7*. https://doi.org/10.7554/eLife.31873.

29. Hebart, M.N., Contier, O., Teichmann, L., Rockter, A.H., Zheng, C.Y., Kidder, A., Corriveau, A., Vaziri-Pashkam, M., and Baker, C.I. (2023). THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. Elife *12*, e82580.

30. Allen, E.J., St-Yves, G., Wu, Y., Breedlove, J.L., Prince, J.S., Dowdle, L.T., Nau, M., Caron, B., Pestilli, F., Charest, I., et al. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. Nat. Neurosci. *25*, 116–126.

31. Chang, N., Pyles, J.A., Marcus, A., Gupta, A., Tarr, M.J., and Aminoff, E.M. (2019). BOLD5000, a public fMRI dataset while viewing 5000 visual images. Scientific Data *6*, 49.

32. Horikawa, T., and Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. Nat. Commun. *8*, 15037.

33. Gong, Z., Zhou, M., Dai, Y., Wen, Y., Liu, Y., and Zhen, Z. (2023). A large-scale fMRI dataset for the visual processing of naturalistic scenes. Sci. Data *10*, 559.

34. Karantzas, N., Franke, K., Willeke, K., Diamantaki, M., Ramakrishnan, K., Bedel, H.A., Elumalai, P., Restivo, K., Fahey, P., Nealley, C., et al. (2026). Dual-feature selectivity enables bidirectional coding in visual cortical neurons. eLife. https://doi.org/10.7554/elife.109861.2.

35. Gondur, R., Stan, P.L., Smith, M.A., and Cowley, B.R. (2026). A tale of two tails: Preferred and anti-preferred natural stimuli in visual cortex. Int Conf Learn Represent *2026*, 98177–98204.

36. Epp, S.M., Castrillón, G., Yuan, B., Andrews-Hanna, J., Preibisch, C., and Riedl, V. (2026). BOLD signal changes can oppose oxygen metabolism across the human cortex. Nat. Neurosci. *29*, 1225–1236.

37. Henderson, M.M., Luo, A.F., Park, S., Tarr, M.J., and Wehbe, L. (2026). Diffusion-based stimulus optimization reveals functional organization across higher visual cortex. bioRxiv. https://doi.org/10.64898/2026.05.12.724119.

38. Prince, J.S., Wang, B., Fel, T., Jagadeesh, A.V., Vaziri, P.A., Alvarez, G.A., Livingstone, M.S., and Konkle, T. (2026). Parametric neural control differentiates top neural network models of primate visual cortex. bioRxiv. https://doi.org/10.64898/2026.08.16.745063.

39. Gaziv, G., Goulding, S., Ayvazian-Hancock, A., Bai, Y., and DiCarlo, J.J. (2025). Noninvasive precision modulation of high-level neural population activity via natural vision perturbations. arXiv [q-bio.NC]. https://doi.org/10.48550/arXiv.2506.05633.

40. Tuckute, G., Sathe, A., Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., Kay, K., and Fedorenko, E. (2023). Driving and suppressing the human language network using large language models. bioRxivorg. https://doi.org/10.1101/2023.04.16.537080.

41. Gifford, A.T., Jastrzębowska, M.A., Singer, J.J.D., and Cichy, R.M. (2025). In silico discovery of representational relationships across visual cortex. Nat. Hum. Behav. *9*, 2079–2098.

42. Shi, Y., Bi, D., Hesse, J.K., Lanfranchi, F.F., Chen, S., and Tsao, D.Y. (2026). Rapid concerted switching of the neural code in the inferotemporal cortex. Nature *653*, 146–155.

43. Kanwisher, N., Stanley, D., and Harris, A. (1999). The fusiform face area is selective for faces not animals. Neuroreport *10*, 183–187.

44. Tong, F., Nakayama, K., Moscovitch, M., Weinrib, O., and Kanwisher, N. (2000). Response properties of the human fusiform face area. Cogn. Neuropsychol. *17*, 257–280.

45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning Proceedings of Machine Learning Research., M. Meila and T. Zhang, eds. (PMLR), pp. 8748–8763.

46. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., and Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv [cs.CV]. https://doi.org/10.48550/arXiv.1811.12231.

47. Mahner, F.P., Lam, K.C., Pereira, F., and Hebart, M.N. (2026). Similarity-based representation factorization for revealing interpretable dimensions in representational data. arXiv [cs.CV]. https://doi.org/10.48550/arXiv.2605.26921.

48. Agarwal, S., Wills, J., Cayton, L., Lanckriet, G.R.G., Kriegman, D., and Belongie, S.J. (21--24 Mar 2007). Generalized non-metric multidimensional scaling. AISTATS *2*, 11–18.

49. Khosla, M., Ngo, G.H., Jamison, K., Kuceyeski, A., and Sabuncu, M.R. (2021). Cortical response to naturalistic stimuli is largely predictable with deep neural networks. Sci Adv *7*. https://doi.org/10.1126/sciadv.abe7547.

50. Rokem, A., and Kay, K. (2020). Fractional ridge regression: a fast, interpretable reparameterization of ridge regression. Gigascience *9*. https://doi.org/10.1093/gigascience/giaa133.

51. Stigliani, A., Weiner, K.S., and Grill-Spector, K. (2015). Temporal processing capacity in High-Level visual cortex is domain specific. J. Neurosci. *35*, 12412–12424.

52. Esteban, O., Markiewicz, C.J., Blair, R.W., Moodie, C.A., Isik, A.I., Erramuzpe, A., Kent, J.D., Gonçalves, M., DuPre, E., Snyder, M., et al. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat. Methods *16*, 111–116.

53. Gulban, O.F., Nielson, D., Poldrack, R., John lee, Gorgolewski, C., Vanessasaurus, and Ghosh, S. (2019). poldracklab/pydeface: v2.0.0. Preprint at Zenodo, https://doi.org/10.5281/zenodo.3524401 https://doi.org/10.5281/zenodo.3524401.

54. Woolrich, M.W., Ripley, B.D., Brady, M., and Smith, S.M. (2001). Temporal autocorrelation in univariate linear modeling of FMRI data. Neuroimage *14*, 1370–1386.

55. Behzadi, Y., Restom, K., Liau, J., and Liu, T.T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. Neuroimage *37*, 90–101.

56. Amunts, K., Mohlberg, H., Bludau, S., and Zilles, K. (2020). Julich-Brain: A 3D probabilistic atlas of the human brain's cytoarchitecture. Science *369*, 988–992.

57. Wang, Y., Huang, H., Rudin, C., and Shaposhnik, Y. (2020). Understanding how dimension reduction tools work: An empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization. arXiv [cs.LG], 1–73. https://doi.org/10.48550/arXiv.2012.04456.