Manuscript submission

The scope and limits of fine-grained image and category information in the ventral visual pathway

Markus W. Badwal[Y 1,2,3], Johanna Bergmann [1], Johannes Roth [2,4], Christian F. Doeller[* 1,5] & Martin N. Hebart[* 2,4,6]

[1] Department of Psychology, Max Planck Institute for Human Cognitive and Brain Sciences, 04103 Leipzig, Germany, [2] Vision & Computational Cognition Group, Max Planck Institute for Human Cognitive and Brain Sciences, 04103 Leipzig, Germany, [3] Department of Neurosurgery, University of Leipzig Medical Center, 04103 Leipzig, Germany, [4] Department of Medicine, Justus Liebig University, 35390 Giessen, Germany, [5] Kavli Institute for Systems Neuroscience, NTNU, 7030 Trondheim, Norway, [6] Center for Mind, Brain and Behavior, Universities of Marburg, Giessen and Darmstadt, 35390 Giessen, Germany

Y Corresponding author: badwal@cbs.mpg.de ** shared senior authorship*

Conflict of interest statement

The authors whose names are listed here certify that they have no affiliation or involvement in any organization or entity with any financial or other interest in this project.

Acknowledgement

We thank all the people who were so kind to read the manuscript and listen to project presentations, and whose constructive criticism and ideas have made this project possible.

The scope and limits of fine-grained image and category information in the ventral visual pathway

Markus W. Badwal[Y 1,2,3], Johanna Bergmann [1], Johannes Roth [2,4], Christian F. Doeller[* 1,5] & Martin N. Hebart[* 2,4,6]

[1] Department of Psychology, Max Planck Institute for Human Cognitive and Brain Sciences, 04103 Leipzig, Germany, [2] Vision & Computational Cognition Group, Max Planck Institute for Human Cognitive and Brain Sciences, 04103 Leipzig, Germany, [3] Department of Neurosurgery, University of Leipzig Medical Center, 04103 Leipzig, Germany, [4] Department of Medicine, Justus Liebig University, 35390 Giessen, Germany, [5] Kavli Institute for Systems Neuroscience, NTNU, 7030 Trondheim, Norway, [6] Center for Mind, Brain and Behavior, Universities of Marburg, Giessen and Darmstadt, 35390 Giessen, Germany

Y Corresponding author: badwal@cbs.mpg.de ** shared senior authorship*

Abstract

Humans can easily abstract incoming visual information into discrete semantic categories. Previous research employing functional MRI (fMRI) in humans has identified cortical organizing principles that allow not only for coarse-scale distinctions such as animate versus inanimate objects but also more fine-grained distinctions at the level of individual objects. This suggests that fMRI carries rather fine-grained information about individual objects. However, most previous work investigating fine-grained category representations either additionally included coarse-scale category comparisons of objects, which confounds fine-grained and coarse-scale distinctions, or only used a single exemplar of each object, which confounds visual and semantic information. To address these

challenges, here we used multisession human fMRI (female and male) paired with a broad yet homogenous stimulus class of 48 terrestrial mammals, with 2 exemplars per mammal. Multivariate decoding and representational similarity analysis (RSA) revealed high image-specific reliability in low- and high-level visual regions, indicating stable representational patterns at the image level. In contrast, analyses across exemplars of the same animal yielded only small effects in the lateral occipital complex (LOC), indicating rather subtle category effects in this region. Variance partitioning with a deep neural network and shape model showed that across exemplar effects in EVC were largely explained by low-level visual appearance, while representations in LOC appeared to also contain higher category-specific information. These results suggest that representations typically measured with fMRI are dominated by image-specific visual or coarse-grained category information but indicate that commonly employed fMRI protocols may reveal subtle yet reliable distinctions between individual objects.

Key words: concept representation; categories; basic objects; ventral visual cortex; fMRI; MVPA; RSA

Significance Statement

While it has been suggested that functional MRI (fMRI) responses in ventral visual cortex carry fine-grained information about individual objects, much previous research has confounded fine-grained with coarse-scale category information or only used individual visual exemplars, which potentially confounds semantic and visual object information. Here we address these challenges in a multisession fMRI study where participants viewed a highly homogenous stimulus set of 48 land mammals with 2 exemplars per animal. Our results

reveal a strong dominance of image-specific effects and additionally indicate subtle yet reliable category-specific effects in lateral occipital complex, underscoring the capacity of commonly employed fMRI protocols to uncover fine-grained visual information.

Introduction

Our visual system provides us with the remarkable ability to derive from our immediate sensory experience a deeper semantic understanding of objects, allowing us to recognize them and form higher-level categories. This ability is thought to be supported by the ventral visual pathway that comprises lateral and ventral occipitotemporal cortex (LOTC; VOTC). Beyond its characteristic responses to faces, scenes, bodies, objects and text (Kanwisher, 2010a), ventral visual cortex exhibits distinct responses differentiating between animate and inanimate objects (Chao et al., 1999; Kriegeskorte et al., 2008; Konkle & Caramazza, 2013; Hebart et al., 2023). Recent work suggests that VOTC and LOTC respond to dimensions such as animacy in a continuous, or even hierarchical, manner (Connolly et al., 2016; Sha et al., 2015; Thorat et al., 2019), indicating the presence of neural representations of increasingly complex categorical information in these brain regions that may only in part be explained by low-level visual features (Jozwik et al., 2016). Thus, while there is disagreement about the degree to which these effects reflect visual or semantic features (Andrews et al., 2010, 2015; Coggan et al., 2016; Long et al., 2018; Bracci et al., 2017, 2019; Jagadeesh & Gardner, 2022), there is a general consensus that ventral visual cortex carries representations that allow humans to distinguish objects at a coarse categorical level (Clarke, 2019), which can be seen as a first step towards unraveling the neural mechanisms underlying object recognition and understanding (DiCarlo et al., 2012; Bracci & Op de Beeck, 2023).

At the more fine-grained category level of individual, basic-level objects, numerous fMRI studies have addressed the question to what degree ventral visual cortex carries information distinguishing individual objects (Haxby et al., 2001; Spiridon & Kanwisher, 2002; O'Toole et al., 2005, Eger et al., 2008; Cichy et al., 2011; Clarke & Tyler, 2014; Iordan et al., 2015). These findings have been taken as evidence that ventral visual cortex indeed represents individual basic-level object categories beyond the immediate visual representations evoked by individual exemplars (Eger et al., 2008; Cichy et al., 2011; Clarke & Tyler, 2014). These cortical object representations appear to show a hierarchical organization paralleling the hierarchical organization of object categories, with more fine-grained category information found at higher spatial resolutions and more coarse-grained information at a more broadly-distributed spatial level (Brants et al., 2011; Bracci and Op de Beeck, 2023).

Despite the evidence for both coarse and fine-grained category representations in previous fMRI studies, most studies highlighting the generalization in ventral visual cortex across stimuli have included objects that are different not only at the basic but also at a more superordinate categorical level. For instance, some work has interweaved the difference between animate and inanimate stimuli with the difference between mammals and insects (e.g. Cichy et al., 2011; Clarke & Tyler, 2014), which are hierarchically different category levels and known to be represented quite differently (Kriegeskorte et al., 2008, Connolly et al., 2012; Hebart et al., 2020). Despite most studies being interested in object-specific effects, most of them do not distinguish between fine-grained and coarse-grained category information and run comparisons including all levels of comparisons. Thus, for studies including both coarse-scale and fine-grained category comparisons, seemingly fine-grained category effects at the level of basic objects may, alternatively, be explained by the presence of coarse categorical effects, since the reported average effects of all comparisons (e.g.,

decoding accuracy) also includes coarse-grained effects. Adding to this challenge, much research has not been controlling for low-level or mid-level visual features (e.g. Clarke & Tyler, 2014; Singer et al., 2023). Thus, at least part of the measured response in relation to individual categories in ventral visual cortex could be driven by such basic visual features (Andrews et al., 2010, 2015; Nasr & Tootell, 2012; Bracci & Op de Beeck, 2016; Proklova et al., 2016; Long et al., 2018). Given these two potential confounds – the mixture of coarse-level and fine-grained category information and the mix of visual and semantic information – there is currently a gap in our understanding of (1) the extent to which fMRI responses in ventral visual cortex carry fine-grained categorical information and (2) the regions that allow for such distinctions.

To address these outstanding issues more specifically, here we focused on basic-level within-category distinctions in land mammals, which constitute a large homogenous class of objects (Fig 1A). This approach allowed us to control for superordinate between-category effects (e.g. animacy) or other known, coarse categorical distinctions (e.g. mammals, insects, birds, or fish). By focusing only on land mammals with faces, we implicitly controlled also for the face-body distinction that could in part be driving overall responses (Kriegeskorte et al., 2008; Ritchie et al., 2021; Proklova & Goodale, 2022). To control for basic visual features, we used two exemplars per object category, which allowed us to compare brain responses for individual images with responses across images of the same category. Finally, to maximize statistical power, we conducted our study across three scanning sessions.

Materials and Methods

Participants

158 For our fMRI experiment, we collected data from a total of 21 adult volunteers. Of these,
159 one volunteer was excluded due to strong head motion and five volunteers due to poor
160 behavioral performance (see below). Thus, our final results are based on the behavioral and
161 fMRI data of 15 volunteers (N = 15, 7 female, 8 male, mean age: 27.8 years, age range: 22 -
162 38 years). The final sample size was limited by the extensive resources required for the
163 acquisition of multisession fMRI and the limited availability of scanning slots.

164 All volunteers had normal or corrected-to-normal vision including normal color vision and
165 provided written informed consent. The study was approved by the local Ethics Committee
166 of the University of Leipzig, Germany (481/19-ek). Participants were remunerated for the
167 online training and MRI scanning sessions.

168 **Stimulus set**

169 A total of 96 object images were used in the current study, where all images were from the
170 animal subcategory of terrestrial, non-flying mammals, i.e. mammals that live on the land
171 (Fig. 1A). The images covered a total of 48 mammals, where each mammal was presented as
172 two different exemplars. Each of the two exemplars had previously been assigned to one of
173 two color contexts (green or red background), resulting in 48 mammal images in the green
174 color context (= context 1) and 48 mammal images in the red color context (= context 2). The
175 assignment of exemplars to color context was consistent across participants. Mammal
176 images were collected via web search with public domain licenses. Given the comparably
177 small number of freely available public domain images at the time of data collection, we did
178 not employ specific criteria for finding images that were visually particularly dissimilar.
179 Images were next edited to remove the background. For simplicity, in the following we may
180 refer to terrestrial, non-flying mammals as animals.

181 As has been shown previously, models that aim to explain distinct neural representations of
182 animate objects appear to be driven significantly by effects such as face-body distinctions,
183 degree of animacy, and taxonomy (Connolly et al., 2012, 2016; Ritchie et al., 2021; Proklova
184 & Goodale, 2022). In order to avoid these effects and to create a stimulus set with a high
185 degree of homogeneity, we only included animals of the biologically distinct class of
186 mammals. Moreover, we also excluded purely aquatic mammals or flying mammals. Thus, all
187 animals were predominantly terrestrial, belonged to the same taxonomic class of mammals,
188 and all images showed both faces and bodies of animals. Thus, rather than constructing a
189 maximally diverse dataset (e.g. Hebart et al., 2020; Proklova & Goodale, 2022; Jozwik et al.,
190 2022), we chose to control for multiple factors and created a large, highly homogenous
191 stimulus set. This was a deliberate decision, as we were interested in the disentanglement of
192 fine-grained representations in the ventral visual pathway.

193 Our choice of terrestrial mammals was driven in part by the results from Hebart et al. (2020),
194 who showed that, based on behavioral similarity data, mammals strongly cluster together
195 within the kingdom of animals (Hebart et al., 2020), and in part by evidence showing that
196 highly animate objects (e.g. mammals) elicit strong and distinct neural responses (Sha et al.,
197 2015; Proklova & Goodale, 2022).

198 In the MRI scanner, participants were shown the colored objects in the center of a uniform
199 mid-gray screen through a mirror attached to the head coil, while the images were projected
200 centrally measuring 12° × 12° visual angle onto a rear-projection screen located at the back
201 of the scanner.

202 **Experimental design**

The experiment consisted of two parts, a memorization and a retrieval task, performed on four subsequent days. On day 1, participants learned the stimuli and the experimental design in a one-hour behavioral online training performed at home. On days 2 to 4, participants performed a short 15 min pre-scan behavioral training and subsequently entered the scanner for data collection.

The color context memory component was included in the task as one motivation of the study was to maximize the chance of attaining relevant signal not only in posterior ventral visual cortex but also in entorhinal and hippocampal cortex, which are known to be involved in context retrieval (Libby et al., 2019; Julian & Doeller, 2021). Additionally, the design made the task sufficiently challenging to be performed over multiple days and kept participants engaged.

The online and pre-scan behavioral training consisted of the memorization task and the retrieval task (Fig. 1B) performed in alternation. During the memorization task (Fig. 1B, left), stimuli within their color context were presented at rapid succession (stimulus duration: 1000 ms; interstimulus interval: 1000 ms) while subjects were asked to memorize each image exemplar in its respective context. This part of the experiment was a passive viewing task. Each of the 96 stimuli was presented once per run. During the retrieval task (Fig. 1B, right), stimuli were again presented in fast succession (stimulus duration: 1000 ms, interstimulus interval: 3000 ms), but now displayed on a uniform gray background. During the image presentation, a black fixation crosshair (Thaler et al., 2013) appeared at the center of the screen. Within the interstimulus interval, subjects had the first 2000 ms to assign the previously displayed animal stimulus to its correct color context by clicking the right or left button depending on the response mapping indicated by the color of the fixation crosshair (Fig. 1B, right). For example, when the fixation crosshair was green on the left and red was

227 the correct response, participants were supposed to push the right button. The side of each
228 color on the fixation cross was randomized. The purpose of the response mapping was to
229 decorrelate choices and button presses. In the online and pre-scan training, memorization
230 and retrieval tasks alternated. The online training consisted of eight runs, the pre-scan
231 training of two runs.

232 During the three fMRI scanning sessions, participants performed only the retrieval task, i.e.
233 without a colored background (Fig. 1B, right). For improved motivation, participants received
234 feedback after each run on the percentage of correct trials. Each fMRI session consisted of
235 eight experimental runs, where each run consisted of 96 trials. Each stimulus was presented
236 once per run. In order to improve model fitting, the event-related sequence was optimized
237 across the three scanning sessions so that the order was maximally different between all 24
238 runs. All in all, there were a total of 96 trials per run, all of which were target trials, eight
239 runs per session, and three sessions over three consecutive days, resulting in 2,304 trials for
240 each participant and 24 repeats of each individual stimulus.

241 For both the memorization and the retrieval task, stimulus presentation and control were
242 performed via a PC computer running PsychoPy2 (Peirce et al., 2019), built upon the open-
243 source programming language Python. The online behavioral training was performed using
244 the *Pavlovia* website (https://pavlovia.org/) build upon PsychoPy.

245 To minimize the chance that results were adversely affected by behavioral performance, we
246 decided to only include participants in our analyses that reached a certain behavioral
247 threshold. On day one, participants needed to reach at least 80% correct responses in a
248 majority of the runs. On day two, we set the threshold at 90% and above. All participants
249 who reached those thresholds on the first two scanning days achieved above 95% correct

trials on day 3, while most were able to get at least one perfect run with 100% correct (see Results).

MRI data acquisition

All MRI data were acquired using a 32-channel head coil on a research-dedicated Siemens Magnetom Prisma 3T MRI scanner (Siemens, Erlangen, Germany) located at the Max-Planck-Institute for Human Cognitive and Brain Sciences in Leipzig, Germany. The scanning procedure was the same for all three scanning days, with the exception of an anatomical T1-weighted scan on the first day. For the functional scans, whole-brain images were acquired using a segmented k-space and steady-state T2*-weighted multi-band (MB) echo-planar imaging (EPI) single-echo gradient sequence that is sensitive to the BOLD contrast (60 slices in interleaved ascending order; anterior-to-posterior (A–P) phase-encoding direction; TR = 1500 ms; 283 TRs per run; echo time (TE) = 24.2 ms; voxel size = 2 × 2 × 2 mm; matrix size = 96 × 96; field of view (FOV) = 204 × 204 mm; flip angle (FA) = 80°; MB acceleration factor 2, GRAPPA factor 2, partial Fourier 6/8). Each MRI session included eight functional task runs. Each functional task run was 7 min 4.5 s in length, during which 283 functional volumes were acquired. To perform distortion correction, for each functional run we recorded phase-encoding-reversed pairs. These images were incorporated in the preprocessing flow. At the end of the first scanning session, high-resolution T1-weighted anatomical sequences (MP2RAGE) were obtained from each participant to allow co-registration and brain surface reconstruction (176 slices; TR = 2300 ms; echo time (TE) = 2.98 ms; flip angle (FA) = 9 degrees; inversion time (TI) = 900 ms; matrix size = 192 × 256; FOV = 192 × 256 mm; voxel size = 1 × 1 × 1 mm).

MRI data preparation and preprocessing

273 We arranged our data according to the Brain Imaging Data Structure (BIDS; Gorgolewski et
274 al., 2016) specification in order to allow for reproducibility of the task-based fMRI study, to
275 comply with standardization, and to use *fMRIPrep* for preprocessing.

276 Results included in this manuscript come from preprocessing performed using *fMRIPrep*
277 version 21.2.0 (Esteban et al., 2019), a *Nipype* (Gorgolewski et al., 2011; Gorgolewski et al.,
278 2017) based tool. Many internal operations of *fMRIPrep* use *Nilearn* (Abraham et al., 2014),
279 principally within the BOLD-processing workflow. For more details of the general *fMRIPrep*
280 pipeline and the detailed steps performed by *fMRIPrep* 21.2.0 on the anatomical and
281 functional data of each participant see https://fmriprep.org/en/stable/workflows.html.
282 Functional images were not smoothed for any analysis performed in this study.

283 **GLM fitting**

284 The blood-oxygen-level-dependent (BOLD) signal for each stimulus for each trial, at each
285 voxel, was modeled separately using the least squares separate (LS-S; Mumford et al., 2012)
286 approach. Using this approach, the general linear model (GLM) includes a single regressor
287 for the stimulus to be estimated while all other stimuli are combined into a single nuisance
288 regressor. Thus, for each run there are as many separate linear models fitted as there are
289 trials (96 GLMs per run). We used LS-S on the grounds that our experiment was a rapid
290 event-related task design with strongly overlapping evoked BOLD signals of adjacent trials,
291 which would result in highly autocorrelated beta estimates, as well as the fact that we
292 presented each stimulus only once per run. Mumford et al. (2012) showed that LS-S model
293 fitting is an improved approach in comparison to a regular GLM for obtaining reliable
294 activation patterns for rapid event-related designs like ours. Each onset regressor was
295 convolved with a canonical hemodynamic response function. For the experimental runs, we

additionally included six head movement regressors derived from estimated head motion parameters (translation and rotation along the x-, y-, and z-axes) in our GLMs. After fitting the GLM, our analysis yielded voxelwise beta weights for each stimulus for each run. For all analyses reported below, the estimated beta weights for each stimulus within each session were averaged across all 8 runs, resulting in a single beta weight for each stimulus for each session in each voxel. Our rationale behind averaging the beta weights within sessions was that every stimulus was presented only once per run, leading to noisy fitting. Thus, the averaged session-wise beta weight provided a more stable and reliable estimate. Thus, we ended up with a total of 96 beta weights for each of the three scanning sessions, one for each animal stimulus, for each subject. These voxelwise beta weights were used as input for all subsequent voxel pattern analyses and model comparisons.

fMRI data model fitting was done using custom code in Python partly based on the *Nilearn* toolbox (https://github.com/nilearn/nilearn). We performed model fitting in native space as many of our analyses were carried out within ROIs in each individual participant.

ROI definition

To quantify BOLD effects at the regional level, we defined regions of the ventral visual pathway by using ROIs from the probability brain maps of Wang et al. (2015), selected masks from n*eurosynth* (https://neurosynth.org/) as well as *neurovault* (https://neurovault.org/), and selected masks available in *FSL* (Fig. 2). We included Hippocampus (HC) and perirhinal cortex (PRC) as an extension of the ventral visual pathway, as recent studies have shown that they are implicated in higher conceptual representations (Clarke & Tyler, 2014; Theves et al., 2020). All ROIs are bilateral clusters including the left and right hemisphere.

319 We merged V1, V2 and V3 from Wang et al. (2015) to construct an early visual cortex (EVC)
320 mask (27,805 mm^3). Splitting the lateral occipital complex (LOC), in two subregions, the
321 lateral occipital region (LO; 9,263 mm^3) as well as the posterior fusiform region (pFus; 11,273
322 mm^3) were constructed using the object association regions from *neurosynth* and selecting
323 the appropriate subregions. The HC mask (8,719 mm^3) was taken from the Harvard-Oxford
324 brain atlas included in *FSL,* and the PRC mask (7,910 mm^3) from a probabilistic map
325 constructed of 38 healthy subjects (https://neurovault.org/images/60340/).

326 For all ROIs, we set a threshold of ≥50 % of the maximum original mask values to only
327 include mostly unambiguous voxels. For analyses, all ROIs were shifted from MNI152 space
328 into the native space of each participant using the transformation matrix of the T1-weighted
329 image provided by *fMRIPrep* and resampled to the nominal voxel size of the functional
330 volumes.

331 **Reliability-based voxel selection**

332 In order to maximize sensitivity of our analyses across sessions, we used reliability-based
333 voxel selection (Tarhan & Konkle, 2020), which selects a subset of voxels within each ROI
334 that exhibit a stable voxel response profile across conditions. Accordingly, an array
335 containing the beta estimates of all stimuli was computed for each voxel for each session (96
336 beta estimates per session in the current study). The Pearson correlation between these
337 arrays was then computed for each voxel across all pairs of the three sessions, and the mean
338 correlation taken as the overall reliability value. Considering that we computed the reliability
339 over three sessions, as opposed to two, we used the Spearman-Brown prediction formula
340 (Spearman, 1910; Brown, 1910) with a *k*-Factor of 1.5 to correct for the reliability estimates
341 across all three sessions (see below).

342 We ended up including all voxels in further analyses of which the estimated correlation
343 exceeded a threshold of $r \geq 0.15$. Based on visual inspection, this threshold optimized the
344 natural tradeoff between brain coverage and stability of the data. Please note, though, that
345 the results were highly stable irrespective of this voxel selection strategy. Using reliability-
346 based voxel selection allowed us to select the most informative and stable voxels across all
347 three sessions in the multi-voxel pattern analyses (MVPA; Haxby et al., 2001) while
348 discarding noisy voxels, thus maximizing the chance of identifying reliable effects across
349 sessions and conditions.

350 **Multivariate decoding analysis**

351 We used multivariate decoding analysis on the fMRI multivoxel patterns (Haxby et al., 2001;
352 Kriegeskorte et al., 2006; Haynes & Rees, 2005) to determine to what extent image-specific
353 and category-specific information was encoded along the ventral visual pathway. A linear
354 support vector machine (SVM) classifier was trained either with a focus on image-specific
355 representations (training and testing within one image set) or on category-specific
356 representations (training on one image set and testing on the other, and vice versa). All
357 analyses were conducted independently for each subject and each ROI. For a given pair of
358 categories, multivoxel pattern vectors from 2 out of 3 sessions were assigned to the training
359 data, and the left out session was used as the independent test data. Data were cross-
360 validated using 6-fold cross-validation, comprising all possible combinations of training and
361 testing sessions. This analysis was repeated for all pairs of objects, and results were
362 averaged across all comparisons. Group-level analyses of decoding accuracy across all
363 subjects was conducted by means of paired t-tests or one-sample t-tests relative to the
364 classification chance level of 50 %.

365 **Representational similarity analysis: Reliability of the representational geometry across**
366 **sessions and exemplars and model comparisons**

367 Representational similarity analysis (RSA) was used (1) to estimate the stability of the
368 pairwise neural activity patterns across sessions, (2) to compare the neural representations
369 across exemplars of the same concept, and (3) to estimate the representational similarity
370 between neural activity and various layers of different deep neural networks as well as
371 semantic embedding models. All RSA was conducted using custom code in Python largely
372 based on the open-source toolbox *rsatoolbox* (rsatoolbox.readthedocs.io; Schütt et al.,
373 2023).

374 Neural representational dissimilarity matrices (RDMs) for each subject for the different ROIs
375 as well as model RDMs were created using 1 minus Pearson's correlation (*r*) as the similarity
376 metric between the representational patterns of the stimuli. We used 1 minus Pearson
377 correlation as the distance metric since it is commonly used and also applied in previous
378 studies of animacy (e.g. Connolly et al., 2016; Bracci et al., 2019). In addition, Pearson's r
379 showed higher stability in across session representations than Euclidean or cosine distance
380 (Kriegeskorte et al., 2008b). We calculated the neural RDMs by calculating the pairwise
381 correlation distance of stimulus-specific beta patterns, yielding a 96 × 96 fMRI RDM for each
382 ROI for each participant. For the various DNNs (see below), we constructed layer-specific
383 RDMs based on 1 minus the Pearson's correlation between all pairs of vectors of unit
384 responses to each stimulus. For each semantic embedding model (see below), a single RDM
385 was constructed based on the word embedding of each concept. *T*o compare RDMs from
386 different sessions and across modalities, we took the lower triangular part of each matrix
387 excluding the diagonal (Ritchie et al., 2017) and converted them to vectors. Subsequently,

these vectors were compared using Spearman's rank correlation rho for pairwise RDM comparisons.

Since we acquired functional data from three separate sessions across three subsequent days, we sought to determine how reliable the within-exemplar representational relationships were across sessions. To quantify the reliability, we computed Spearman's rho between the context-specific RDMs containing the respective 48 stimuli (48 × 48 RDM) of each ROI across all three sessions. The correlation values were then averaged across all comparisons. This way, we obtained a single correlation value for each subject for each ROI that indicated the reliability of the representational geometry across sessions. We used the context-specific RDMs for within-exemplar correlations in order to have the same modality (48 × 48 RDM) as when looking at across-exemplar similarities.

Having two exemplars of each concept in our data set, we further wanted to test whether there are generalizable representational relationships between different exemplars of the same animal species (e.g. Is the representational relationship between a dog and a mouse generalizable, independent of its visual appearance?). We thus computed the Spearman's rho between the two context RDMs (48 × 48 RDM), correlating context 1 RDM (green) with context 2 RDM (red) across all three sessions and took the average. This way, we obtained a single correlation value for each subject for each ROI that indicated the representational similarity between the two exemplars of all concepts.

Variance Partitioning: Modeling visual, semantic and task effects

To determine the role of visual features, semantics, and task in the estimated representational similarities, we carried out variance partitioning to partial out variance explained by these factors (e.g. Hebart et al., 2018; Cichy et al., 2019). Once an effect has

been partialled out, a remaining significant correlation between the residual RDMs would indicate that the partialled-out factors cannot fully explain the hypothesized relationship between the RDMs.

Deep neural networks (DNNs) reflect an established class of visual models of neural representations along the ventral visual pathway (Khaligh-Razavi & Kriegeskorte, 2014; Cichy et al., 2016) as well as perceived similarity between images (Kubilius et al., 2016, Hebart et al., 2022). We chose AlexNet (Krizhevsky et al., 2012) as it has been used commonly in the field of cognitive computational neuroscience to model the various processing stage of object recognition (e.g. Cichy et al., 2016; Bracci et al., 2019; Cichy et al., 2019; Mehrer et al., 2021). The THINGSvision toolbox (Muttenthaler & Hebart, 2021) was used to extract layer-specific activity patterns from AlexNet, which had been pretrained on the ImageNet dataset (Deng et al., 2009). Based on previous findings (e.g. Cichy et al., 2016; Bankson et al., 2018), we modeled low-level visual features using an early convolutional layer (= layer #2). We calculated the pairwise Pearson's *r* between layer-specific vectors, generating three distinct RDMs: a 96 × 96 RDM for the entire stimulus set, a 48 × 48 RDM for the stimuli in context 1, and a 48 × 48 RDM for the stimuli context 2. In addition, to model mid-level visual features, we used the “ShapeComp” model to account for such visual feature effects, which is a model of perceived shape similarity (Morgenstern et al., 2021). Here again, we calculated the pairwise Pearson's r between shape feature vectors, generating a distinct 96 x 96 RDM indicating the pairwise shape similarity between all stimuli.

To assess fine-grained information about the semantic meaning of the shown stimuli, two similarity models were used. As a first model, we included the THINGS embedding of perceived similarity of Hebart and colleagues (2020), which had been trained on behavioral data during a triplet odd-one-out task (Hebart et al., 2020, 2023) and which captures

435 behaviorally-relevant visual and conceptual features. Here, we generated the similarity
436 matrix from the embedding, assuming the similarity as the probability of not choosing a pair
437 together across all possible pairs of comparisons, and focused the possible triplets in the
438 comparison on the set of 48 objects (48 × 48 RDM). As a second model, we used the CSLB
439 concept property norm model (Devereux et al., 2014) to include sets of semantic features
440 explicitly listed by human participants. In this model, the semantic features are labeled to
441 belong to a certain semantic feature type (e.g. external component, behavior), which allows
442 for the subselection of features. Each object vector consists of binary numbers that reflect
443 the presence or absence of a semantic feature. Since the CSLB model contains only 34 of the
444 48 basic-level concepts in our study, we limited our analyses to these concepts when using
445 the CSLB semantic feature model for comparison. We extracted the semantic feature vectors
446 for each concept and calculated the Pearson's r distance between each pair of vectors (34 ×
447 34 RDM).

448 **Statistical analysis**

449 For testing effects within each ROI, we used a two-tailed one sample t-test on the group
450 results with an expected population mean of zero (e.g. testing whether the activity patterns
451 are reliable in a ROI). For testing effects between ROIs, we used a pairwise t-test (e.g.
452 difference between reliability in two ROIs). Where necessary, we performed a non-
453 parametric permutation test and indicated this in the text. All reported p-values were
454 corrected for multiple comparisons across ROIs using Bonferroni correction. Results are
455 reported as mean values with the respective standard error of the mean (SEM, ±).

456 **Results**

457 Our goal in this study was to determine the degree to which members of a highly
458 homogenous taxonomic subgroup, in our case terrestrial mammals, would elicit distinct
459 voxel activation patterns. More specifically, we aimed to (1) identify the degree to which
460 differences between basic-level objects could be identified that would generalize across
461 different exemplars of the same mammal, and (2) to identify whether a detectable semantic
462 relationship within such a homogenous stimulus set could be found.

463 **Behavioral results**

464 The strict inclusion criteria used in the present study (see Material & Methods) ensured (1)
465 that random guessing was unlikely and (2) that volunteers had a reliable grasp of each
466 stimulus in its respective color context and were sufficiently attentive. This is reflected in the
467 behavioral results. Participants performed very well on all three days, with an accuracy that
468 reached very high levels on the second day (Day 1: 85.02% ± 6.82, Day 2: 95.63% ± 3.80, Day
469 3: 98.13% ± 2.34). We also found a near absence of missed responses in the 2000 ms
470 interstimulus response window across all three scanning days (Misses: 3.15 ± 2.98 out of
471 2,304 trials), corroborating the evidence that participants were paying attention and were
472 alert throughout.

473 **Reliability of voxel response profiles decreases along the ventral visual pathway**

474 Having verified that participants had a reliable apprehension of the stimuli, we were next
475 interested in the degree to which each participant's fMRI data exhibited reliable activity
476 patterns across the three different scanning days. To test this, we used reliability-based
477 voxel selection (Tarhan & Konkle, 2020), while at the same time maximizing statistical power
478 across sessions and discarding noisy voxels from comparisons. A qualitative illustration of
479 the results shows that early visual regions contained many highly reliable voxels, with a

480 decline along the ventral visual pathway (Fig. 3a), which is in line with previous findings
481 (Tarhan & Konkle, 2020). A quantitative depiction confirms that EVC contains the largest
482 fraction of reliable voxels out of all ROIs (EVC reliability: 49.73% ± 2.29%, 1367 voxels ± 63;
483 Fig. 3b). In the lateral occipital complex, we also found large fractions of voxels that survived
484 our inclusion threshold (LO reliability: 37.27% ± 2.25%, 316 voxels ± 19; pFus reliability:
485 15.63% ± 1.34%, 174 voxels ± 15; Fig. 3b). Since the fraction of reliable voxels and reliability
486 measures decreased significantly as compared with EVC, it is reasonable to assume that the
487 information encoded in LO and pFus is more difficult to capture than in early visual regions,
488 possibly also due to the reduced MRI signal in these regions. Inspecting the brain regions
489 considered to be at the apex of the ventral visual pathway, we found only a small number of
490 voxels that showed reliable response profiles in HC and PRC (HC reliability: 6.13% ± 0.54, 52
491 voxels ± 5; PRC reliability: 5.85% ± 0.42, 20 voxels ± 2; Fig. 3b). This could at least in part be
492 explained by a combination of signal loss and distortions, which are common in these
493 regions (Weiskopf et al., 2006; Olman et al., 2009). Additionally, the information encoded by
494 the neurons in these regions might be too complex to be captured by voxel patterns
495 recorded with fMRI.
496 In summary, there is large coverage of reliable voxels in the regions along the ventral visual
497 pathway, with many highly reliable voxels in visual areas and a significant decrease of
498 reliability towards the more anterior-ventral regions (Fig. 3a, b).

499 **Decoding of image-specific and category-specific information**

500 Our primary aim was to reveal the degree to which ventral visual cortex contained
501 representations within the highly homogenous stimulus class of mammals that would
502 generalize beyond image-specific effects. To this end, we trained and tested a linear
503 classifier on image-specific and category-specific multivoxel patterns, respectively. We

504 operationalized category-specific effects as information that would generalize across visually
505 distinct exemplars, resulting in above chance cross-decoding accuracy for visually dissimilar
506 exemplars.

507 The image-specific and category-specific decoding results are shown in Figure 4a. We found
508 that image-specific information could be decoded significantly above chance in all ROIs along
509 the ventral visual pathway. Especially in early and high-level visual regions, image-specific
510 decoding accuracies were very high (EVC: 89.54% ± 1.05%, LO: 76.52% ± 0.98%, pFus:
511 76.69% ± 0.70%, $p < 0.001$). This result was expected as these regions are known to
512 represent low- and high-level visual features of the viewed images. Further, voxel patterns in
513 more anterior brain regions still contained sufficient image-specific information to accurately
514 decode single images significantly above chance level (HC: 71.18% ± 0.77%, $p < 0.001$; PRC:
515 67.68% ± 0.60%, $p < 0.001$). The observed phenomenon likely suggests that the strong visual
516 features effects from visual regions extended to higher brain regions along the ventral visual
517 pathway, exerting a predominant influence on the representational patterns.

518 To identify the degree to which higher category-specific information is present in each ROI,
519 we implemented cross-decoding (Fig. 4a). The reasoning for adopting this approach is that, if
520 the classifier trained on one exemplar is able to accurately decode another exemplar from
521 the same basic level category, it can be inferred that the fundamental representational
522 structure of the two visually dissimilar but categorically similar objects is significantly
523 related. For cross-decoding, we found above chance accuracy in EVC (60.95% ± 0.66%, $p <$
524 0.001), LO (61.53% ± 1.00%, $p < 0.001$) and pFus (59.85% ± 0.88%, $p < 0.001$). Decoding
525 accuracy in HC (52.25% ± 0.44%, $p = 0.001$) and PRC (52.20% ± 0.63%, $p = 0.019$) exceeded
526 chance level, but with comparably small effects. Due to the numerically small effects in HC
527 and PRC, we did not carry out follow-up analyses on these regions.

528 Above chance decoding of visually different exemplars of the same basic-level category
529 could hint at higher level categorical information encoded in the voxel patterns.
530 Alternatively, it is possible that exemplars of the same basic-level category still share distinct
531 visual features that make them more similar to each other than to exemplars from other
532 categories allowing for above chance decoding. This idea is supported by the above-chance
533 decoding accuracies in EVC. To test this alternative explanation, we split the categories in
534 different subgroups based on their visual similarity. Visual similarity was computed as a
535 proxy, using Pearson similarity of the first convolutional layer of AlexNet (see Material &
536 Methods). We then focused our analyses on the top- and bottom-quartile of stimuli
537 according to their visual similarity. We then repeated the category-specific decoding analysis
538 but focused all comparisons on object pairs within these quartiles. The results are displayed
539 in Figure 4b. The decoding accuracy was higher for the visually most similar exemplars than
540 for the visually most dissimilar exemplars in all ROIs. This difference is highly significant in
541 EVC and LO (EVC: mean difference = 17.42%; LO: mean difference = 10.00%, $p < 0.001$) and
542 significant in pFus (mean difference = 4.39%, $p = 0.015$), suggesting that visual similarity is a
543 central factor in all ROIs. In EVC, decoding accuracy for the visually most dissimilar categories
544 dropped to chance level (49.12% ± 1.24%), indicating that the visual similarity comparison
545 effectively captured visual effects. In LO and pFus, decoding accuracy for the visually most
546 dissimilar categories remained significantly above chance (LO: 56.54% ± 0.95%; pFus: 58.16%
547 ± 0.78%) indicating category information beyond the basic visual effects captured by the
548 early DNN layer.

549 Taken together these results show that visually identical images can be reliably decoded
550 from all ROIs along the ventral visual pathway and that visual similarity dominates the
551 decoding accuracy between different exemplars of the same basic-level category.

552 **Reliability of representational geometry within and across animal exemplars**

553 Next, we used RSA to compute the across-session reliability of the representational
554 geometry within and across exemplars. While multivariate decoding revealed across
555 exemplar category-specific information contained in the multivoxel patterns, RSA allows to
556 reveal whether the global representational geometry is similar between two sets of animal
557 exemplars, which would more broadly indicate generalizable representations.

558 As participants were exposed to visually identical stimuli during each scan, we predicted the
559 elicited fMRI patterns within-exemplars in early- and mid-level visual areas to be relatively
560 similar across sessions. Indeed, the reliability in EVC was highly significant when compared
561 with zero (mean $r = 0.38 \pm 0.01$, $p < 0.001$, Fig. 5b) and significantly higher than in any other
562 ROI ($p < 0.001$). We were further interested in the extent to which the representations in
563 higher visual areas in occipito-temporal cortex (OTC) showed reliable activity patterns. These
564 representations were expected to encode more complex, possibly semantic, features of
565 object identity (e.g. Konkle & Oliva, 2012; Bracci & Op de Beeck, 2016; Contier et al., 2023).
566 Indeed, the reliability in LO as well as pFus was significantly larger than zero (LO: mean r =
567 0.19 ± 0.01, pFus: mean $r = 0.19 \pm 0.01$, $p < 0.001$, Fig. 5b), with no significant difference
568 between the two ($p = 0.81$). This indicates that the information contained in the
569 representational geometry in LOC can be reliably captured across multiple sessions. Finally,
570 we examined the pattern reliability in the final and most anterior regions of the visual
571 processing hierarchy, which are thought to code for concept-specific information (Clarke &
572 Tyler, 2014; Theves et al. 2020). Even though low, the across-session reliability in these
573 regions remained significant when compared to zero (HC: mean $r = 0.02 \pm 0.004$, PRC: mean
574 $r = 0.05 \pm 0.008$, $p < 0.001$; Fig. 5b). Together, these results suggest that the further one
575 moves along the hierarchy of the ventral visual stream, the lower the reliability of the

representational geometry between objects within a highly homogenous stimulus set. This decline may be explained by a combination of increasingly more complex information encoded in the representational patterns and a decrease of fMRI signal quality. These results are in line with our analyses of the voxel reliability measures (Fig. 3).

Having established the within-exemplar reliability as a noise ceiling, next we investigated whether the representational similarity generalizes across exemplars, which would hint at a basic category-specific representational geometry. To identify such categorical representations, we computed separate RDMs for each of the two contexts for each session within each participant. As they both contained the same 48 basic-level categories, yet different exemplars of each, we were able to correlate the two context RDMs with each other to reduce visually-specific effects in the comparison (Fig. 5a, right panel). In order to compare the results to our within-exemplar reliability measures (Fig. 5b), we correlated the context RDMs across sessions (e.g. green-context RDM from session 1 with red-context RDM from session 2) and not within sessions. We hypothesized that, if there are fine-grained category representations in the patterns, they should show up as significant similarities between the two context RDMs.

Our results showed that there were indeed highly significant correlations between the two RDMs in lower and higher visual regions (EVC: mean $r = 0.09 \pm 0.008$, LO: mean $r = 0.07 \pm 0.01$, pFus: mean $r = 0.09 \pm 0.01$, $p < 0.001$; Fig. 5b). We did not find any significant across-exemplar correlations in the representational geometries in HC or PRC (HC: mean $r = 0.009 \pm 0.006$, $p > 0.05$; PRC: mean $r = -0.002 \pm 0.007$, $p > 0.05$; Fig. 5b). In all ROIs, except for HC, the difference between the within- versus across-exemplar correlations were highly significant ($p < 0.001$).

599 To test whether reliability-based voxel selection introduced any specific bias, we repeated
600 the Pearson correlation analysis on the beta arrays of voxels in a brainstem mask from *FSL*
601 and selected the most reliable voxels ($r >= 0.15$). As expected, subsequent across-session
602 RDM correlation analyses on these voxels indicated no significant effect (within-exemplar:
603 mean $r = 0.01 \pm 0.007$, $p > 0.05$; across-exemplar: mean $r = -0.008 \pm 0.006$, $p > 0.05$), and no
604 significant difference of using reliability-based voxel selection as compared to no selection (p
605 > 0.05), demonstrating that no specific effect was introduced through this procedure.
606 These results suggest that the representational geometry of individual images is highly
607 reliably represented in lower and higher visual regions, namely EVC and LOC, while across-
608 exemplar representational similarities showed smaller yet significant effects. These results
609 suggest that there is a stable representational geometry across different exemplars of the
610 same basic-level category, consistent with our multivariate decoding results (Fig. 4). Post-
611 hoc analyses on whole ROIs without reliability-based voxel selection revealed a similar
612 pattern of results (Fig. 5c), yet with overall smaller effects, demonstrating no specific bias
613 through the voxel selection strategy.

614 To assess the possibility that these across-exemplar effects are driven by shared low-level
615 visual information, which we expected to be rather high within basic categories, we utilized
616 variance partitioning to control for low-level visual feature effects. To this end, as a proxy of
617 visual features, we took the RDM of the second convolutional layer of the DNN AlexNet layer
618 that best explained the variance in EVC (mean $r = 0.52$; Cichy et al., 2016; Kubilius et al.,
619 2016) and partialled out the explained variance. We refer to the correlations remaining after
620 partialling out the early DNN layer as low-level-controlled (Fig. 5d), as the variance explained
621 by the DNN layer is unlikely to capture mid- and high-level visual features (e.g. shape and
622 texture) contained in the images. Please note that the term “low-level-controlled” does not

623 imply full low-level control, which cannot be achieved without knowing the exact data
624 generating process. These analyses were only applied to the ROIs showing significant across-
625 exemplar correlations, namely EVC, LO and pFus (Fig. 5d).
626 After partialling out the early layer RDM, the correlation values in EVC dropped significantly
627 (mean difference = 0.03, $p < 0.01$; Fig. 5d). This indicates that the shared across-exemplar
628 representations in EVC are partially dictated by low-level visual features. Looking further, in
629 the LOC partialling out the variance explained by the early layer RDM did not lead to a
630 significant decrease in the set of correlation values (mean difference < 0.01; $p = 0.99$; Fig.
631 5d). In line with the multivariate decoding results, this shows that there are some across-
632 exemplar representations in LO and pFus that cannot be reduced to low-level visual features
633 captured by the DNN layer but rather indicate a shared representation at a different level.
634 To test whether these residual effects can be explained by shared mid-level visual features,
635 we additionally partialled out the variance explained by a model of perceived shape
636 similarity (Morgenstern et al., 2021). We found that after controlling for shape similarity
637 between exemplars significant results remained significant and vice versa (low/mid-level-
638 controlled; Fig. 5d). This shows that the across-exemplars representations in LO and pFus
639 cannot be explained by the features of our models chosen to reflect low-level or mid-level
640 visual information, suggesting that other, potentially more high-level information is encoded
641 in these patterns.

642 **Similarities between neural and semantic knowledge model RDMs**

643 The multivariate decoding as well as RSA results showed that the response patterns in LOC
644 contain some form of object-specific information shared between different exemplars, with
645 a stable representational geometry. While this is suggestive of semantic effects, this is only
646 indirectly addressed with the tests of generalization between exemplars. To directly address

647 the contribution of semantic knowledge to these patterns of brain responses, we related
648 two numerical embeddings with fMRI response patterns using RSA (Fig. 6a), the THINGS
649 behavioral similarity embedding (Hebart et al., 2020, 2023) and CSLB semantic feature norm
650 (Devereux et al., 2014). Results are only reported for EVC, LO (Fig. 6b), and pFus, since HC
651 and PRC showed no significant RDM generalization between exemplars (Fig. 5b).
652 We found a highly significant correlation between the THINGS perceived similarity
653 embedding RDM and the RDMs in EVC, LO and pFus (EVC: mean $r = 0.09 \pm 0.008$, LO: mean r
654 $= 0.04 \pm 0.009$, pFus: mean $r = 0.04 \pm 0.009$, $p < 0.001$; Fig. 6b, left). When partialling out the
655 variance explained by the early layer DNN, which we again refer to as "low-level-controlled",
656 the correlation in EVC dropped significantly (mean difference in $r = 0.03$, $p = 0.019$) yet
657 remained significant (mean $r = 0.06 \pm 0.007$, $p < 0.001$; Fig. 6b) indicating that the semantic
658 embedding partially explained variance dominated by low-level visual features in EVC. In LO
659 and pFus, partialling out the DNN-related variance did not significantly change the
660 correlations ($p > 0.05$; Fig. 6b), suggesting that the THINGS model captured information
661 encoded in these regions that transcends the variance explained by the early DNN layer. In
662 order to control for mid-level similarity effects, we next partialled out the variance explained
663 by a model of perceived shape similarity (Morgenstern et al., 2021). We again found that
664 after controlling for the variance explained by the shape similarity model the results of the
665 embedding analysis remained significant (low/mid-level-controlled; Fig. 6b). These results
666 complement our multivariate decoding and across-exemplar RSA results (Fig. 4, 5),
667 suggesting that the information encoded in LO and pFus encompasses object-specific
668 category information that extends beyond low-level and mid-level visual feature effects
669 captured by our models.

670 Lastly, we tested to what extent the CSLB model based on semantic features could capture
671 information encoded in the voxel patterns. The CSLB model significantly correlated with the
672 patterns in EVC (mean $r = 0.05 \pm 0.008$, $p < 0.001$; Fig. 6b, right), which again decreased
673 significantly after partialling out the early DNN layer RDM ($p < 0.01$, low-level-controlled, Fig.
674 6b, right), leading to non-significant remaining effects (mean $r = 0.02 \pm 0.08$, $p = 0.25$). The
675 model was not able to explain any significant variance in LO (mean $r = 0.00 \pm 0.009$, $p = 0.99$;
676 Fig. 6b, right). However, in pFus the correlation with the CSLB model RDM was highly
677 significant (mean $r = 0.06 \pm 0.009$, $p < 0.001$) and was not significantly affected by partialling
678 out the early DNN layer ($p = 0.99$) and remained significant (mean $r = 0.06 \pm 0.009$, $p < 0.001$;
679 Fig. 6b, right). Similar to the analyses above, partialling out the variance explained by our
680 control model for mid-level shape similarity did not change the pattern of results and
681 significant results remained significant (low/mid-level-controlled; Fig. 6c).
682 These findings further suggest that the response patterns in pFus encode higher-level
683 category-specific information, in line with our previous decoding as well as RSA results.

684 **The effect of the color context on the representational geometry**

685 Since our task contained a context component, it was important to rule out the possibility
686 that the color context was adversely affecting the results. We tested for the influence of the
687 context in two ways. First, we repeatedly mixed ($n = 1000$) the color contexts within each
688 session across all stimuli and used this as a basis for forming new, mixed RDMs that were the
689 same in content but mixed in color context. If the effects we reported were driven by color
690 context, then shuffling the contexts should reduce or eliminate these results. However, the
691 results remained unchanged after shuffling ($p > 0.99$), demonstrating that color context was
692 not affecting the pattern similarity across sessions. Additionally, in order to test whether our
693 analyses might miss obvious patterns in the similarity relation between stimuli, we

conducted 2-dimensional metric multidimensional scaling (MDS) on the multivoxel patterns in all ROIs averaged across all participants and sessions. The results showed no obvious pattern related to color context (Fig. 7). To quantify these visually illustrated pairwise distances and test to what extent there were significant differences between the two populations, we computed a t-test comparing the within- with the across-context dissimilarity values within each participant. After controlling for multiple comparisons, we did not find significant differences between the dissimilarity values of within- versus across-context objects in any ROI ($p > 0.05$), again indicating that the context did not significantly influence the similarity values between the stimuli.

Together, these results indicate that color context did not significantly affect the object-specific neural representations.

Discussion

It is commonly assumed that fMRI studies can provide fine-grained information about basic-level objects in higher-level visual cortex (Eger et al., 2008; Cichy et al., 2011; Iordan et al., 2015). However, much previous work has left open to what degree previous findings could in part best be explained by coarse-grained category effects (e.g. animate vs. inanimate; mammal vs. insect) or by low-level visual features of individual objects (Andrews et al., 2010, 2015; Coggan et al., 2016; Bracci et al., 2017; Groen et al., 2017). In this context, the degree to which the global representational geometry remains stable or shifts between different object examples of the same basic-level category, has remained unclear. By putting together a novel stimulus set and leveraging different models, thereby carefully controlling for coarse categorical and visual feature effects, our study reveals three major findings. First, object-related information and the representational geometry along the ventral visual pathway are strongly dominated by image-specific effects. Second, there are much weaker yet reliable effects across images of the same object that cannot be explained by a low- or mid-level visual feature model. Third, a comparison of two embedding models of semantic knowledge showed that these remaining effects may, in part, be explained by high-level semantics.

Consistent with recent approaches of study designs in cognitive neuroscience (Naselaris et al., 2021; Hebart et al., 2020, 2023), our study emphasized extensive fMRI data sampling over multiple days to enhance statistical power and optimize the reliability of our findings. This is especially critical when looking for fine-grained representations encoded in the multivoxel patterns as inter-individual differences induced variability, and hence noise, which becomes particularly apparent at fine spatial scales (Weiner & Grill-Spector, 2012; Naselaris et al., 2021). Additionally, to maximize statistical power, we used strict behavioral

730 inclusion criteria as well as refined model fitting (LS-S) and reliability-based voxel selection
731 (Tarhan & Konkle, 2021).

732 Previous studies have supported the notion that visual features strongly drive responses
733 along the ventral visual pathway, even in higher-level regions, making it challenging to
734 disentangle visual from semantic information (Andrews et al., 2010, 2015; Baldassi et al.,
735 2013; Long et al., 2017, 2018). In line with this, our decoding results revealed highly robust
736 image-specific representations across scanning sessions. As expected, these effects were
737 especially strong in early visual cortex (EVC) and slightly weaker in higher visual areas (LO
738 and pFus, together LOC). Moreover, our RSA results showed a reliable image-specific
739 representational geometry in EVC and, to a smaller extent, in LOC. Taken together, these
740 results demonstrate that both the image-specific information and the image-specific
741 representational geometry remained reliable across scanning sessions in early and later
742 visual regions. Consequently, low-level visual feature effects appeared to be a strong and
743 reliable driver of the representations in the regions.

744 Interestingly, our decoding results showed that reliable image-specific representations
745 extended to regions in the anterior temporal lobe (HC and PRC), also mirrored in the reliable
746 image-specific representational geometry. While in the past it has proven challenging to
747 identify specific effects of visual stimuli in HC and PRC using more traditional vision tasks, it
748 is possible that the color context task used in this study, in combination with the high within-
749 participant statistical power, led to stronger engagement and more reliable effects of these
750 regions, respectively. Beyond the potential for using this approach to studying visual
751 representations in HC and PRC, these results reveal that low-level visual feature effects can
752 in fact be reliably represented in such high-level brain regions and further confirm that
753 supposedly high-level representations are significantly influenced by basic visual features

754 (Bracci et al., 2017; Knapen, 2020; Silson et al., 2021). Thus, it is conceivable that effects
755 previously related to higher semantic information (e.g. Clark & Tyler, 2014) are also affected
756 by low-level visual information.

757 Category-specific representations should not be solely explainable by low-level visual
758 features, but should be determined by distinct shared semantic features (Clarke & Tyler,
759 2015; Clarke, 2019). Prior fMRI research has suggested that there is category-specific
760 information encoded in higher visual regions such as LOC (Eger et al. 2008; Cichy et al., 2011;
761 Iordan et al., 2015). This information supposedly generalizes across exemplars of the same
762 category, which suggests that these effects are best explained by shared visual features,
763 semantics or a combination of both. Our results confirm and extend on these findings, in
764 that we found small but significant shared representations between different exemplars of
765 the same category in LOC. These effects are not significantly influenced by the visual
766 similarity between exemplars, indicating that the representations indeed may contain
767 category-specific information. Supporting these results, we found significant correlations
768 between the fMRI response patterns in LOC and two embedding models that reflect
769 semantic information (Devereux et al., 2014; Hebart et al., 2020, 2023). This indicates that
770 the representations in LOC, beyond containing image-specific visual information, may carry
771 some form of category-specific high-level information.

772 While our results clearly demonstrate that representations along the ventral visual stream
773 are dominated by visual information, it remains possible that our control models did not
774 capture some of the variance related to more low-level or mid-level visual features, and thus
775 remaining across-exemplar information may be explained not by category-related but by
776 visual information. In this context, a potential limitation of our study was that we did not
777 select visually strongly different exemplars within-category, a methodological choice that

could have further reduced the influence of visual features. Future models may show improved predictions of low-level and mid-level visual features and may thus help us more definitely tease apart the effects of low-level, mid-level and category-specific information in similar experimental designs. However, given the effectiveness of DNN models and semantic embeddings in capturing visual and semantic effects, respectively, it appears to be plausible that the remaining effects after low- and mid-level feature control are not primarily visual in nature but may also capture some degree of semantic effects.

The absence of high-level semantic information in the representations in PRC and HC could be explained by limitations of current fMRI techniques. First, the spatial and temporal resolution of fMRI might be too coarse to capture such fine-grained representations within such a homogenous object space. At the same time, image-specific information was captured to a certain degree, indicating that semantic information may be represented in a different format across voxels than visual information. Second, fMRI data from the anterior parts of the medial temporal lobe are known to contain weak and noisy signals (Weiskopf et al., 2006; Olman et al., 2009). Thus, such nuanced representations might not be picked up with fMRI in these regions even if the information is encoded there. An additional plausible explanation is that there is dynamic information flow between posterior and anterior regions of the ventral visual pathway, which in accordance allows for the dissociation between semantically similar objects (Clarke et al., 2011; Devereux & Clarke, 2018; Clarke, 2019). Capturing and discriminating between these dynamic representations requires data acquisition techniques with high temporal resolution. Future studies looking for basic-level object representations could use higher spatial resolution with fMRI (e.g. 7T) together with more dynamic neural recordings techniques (e.g. MEG), and could combine both techniques (Cichy et al., 201; Hebart et al., 2018).

802 In summary, our study went beyond previous work by using (1) a substantially larger group
803 of categories with different exemplars, (2) a more fine-grained level of abstraction, (3)
804 revealing not only generalizable decoding but also a stable representational pattern, and (4)
805 demonstrating potential sources of this information using DNNs and semantic models (Eger
806 et al. 2008; Cichy et al., 2011; Iordan et al., 2015). We deliberately chose categories of the
807 class of animals, more specifically land mammals, as these stimuli have been extensively
808 studied and we know them to be firmly represented along the ventral visual pathway
809 (Kriegeskorte et al., 2008; Connolly et al., 2012, 2016; Sha et al., 2015; Bracci et al., 2019;
810 Ritchie et al., 2021). Only including land mammals ensured that the effects between
811 different exemplars of the category are indeed best explained by fine-grained semantics
812 rather than coarse-grained categorical effects (e.g. having a face, degree of animacy; Sha et
813 al., 2015; Proklova & Goodale, 2022). Our work highlights the dominance of image-specific
814 visual feature representations yet demonstrates the presence of small, generalizable effects.
815 This extends our understanding of the function of the ventral visual pathway and opens the
816 door to a more detailed understanding of the nature of representations across these
817 regions.

818

819 **References**

820 Andrews, T.J., Clarke, A., Pell, P., Hartley, T., 2010. Selectivity for low-level features of
821 objects in the human ventral stream. NeuroImage 49, 703–711.
822 https://doi.org/10.1016/j.neuroimage.2009.08.046

823 Andrews, T.J., Watson, D.M., Rice, G.E., Hartley, T., 2015. Low-level properties of natural
824 images predict topographic patterns of neural response in the ventral visual pathway.
825 Journal of Vision 15, 3. https://doi.org/10.1167/15.7.3

Baldassi, C., Alemi-Neissi, A., Pagan, M., DiCarlo, J.J., Zecchina, R., Zoccolan, D., 2013. Shape Similarity, Better than Semantic Membership, Accounts for the Structure of Visual Object Representations in a Population of Monkey Inferotemporal Neurons. PLoS Comput Biol 9, e1003167. https://doi.org/10.1371/journal.pcbi.1003167

Bankson, B.B., Hebart, M.N., Groen, I.I.A., Baker, C.I., 2018. The temporal evolution of conceptual object representations revealed through models of behavior, semantics and deep neural networks. NeuroImage 178, 172–182. https://doi.org/10.1016/j.neuroimage.2018.05.037

Bracci, S., Op de Beeck, H., 2016. Dissociations and Associations between Shape and Category Representations in the Two Visual Pathways. J. Neurosci. 36, 432–444. https://doi.org/10.1523/JNEUROSCI.2314-15.2016

Bracci, S., Ritchie, J.B., Kalfas, I., Op de Beeck, H.P., 2019. The Ventral Visual Pathway Represents Animal Appearance over Animacy, Unlike Human Behavior and Deep Neural Networks. J. Neurosci. 39, 6513–6525. https://doi.org/10.1523/JNEUROSCI.1714-18.2019

Brants, M., Baeck, A., Wagemans, J., Op De Beeck, H.P., 2011. Multiple scales of organization for object selectivity in ventral visual cortex. NeuroImage 56, 1372–1381. https://doi.org/10.1016/j.neuroimage.2011.02.079

Chao, L.L., Haxby, J.V., Martin, A., 1999. Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. Nat Neurosci 2, 913–919. https://doi.org/10.1038/13217

Cichy, R.M., Chen, Y., Haynes, J.-D., 2011. Encoding the identity and location of objects in human LOC. NeuroImage 54, 2297–2307. https://doi.org/10.1016/j.neuroimage.2010.09.044

Cichy, R., Pantazis, D. & Oliva, A., 2014. Resolving human object recognition in space and time. Nat Neurosci 17, 455–462. https://doi.org/10.1038/nn.3635

Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., Oliva, A., 2016. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci Rep 6, 27755. https://doi.org/10.1038/srep27755

Cichy, R.M., Kriegeskorte, N., Jozwik, K.M., van den Bosch, J.J.F., Charest, I., 2019. The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. NeuroImage 194, 12–24. https://doi.org/10.1016/j.neuroimage.2019.03.031

Clarke, A., Taylor, K. I., & Tyler, L. K., 2011. The evolution of meaning: Spatiotemporal dynamics of visual object recognition. Journal of Cognitive Neuroscience, 23(8), 1887–1899. https://doi.org/10.1162/jocn.2010.21544

Clarke, A., Tyler, L.K., 2014. Object-Specific Semantic Coding in Human Perirhinal Cortex. J. Neurosci. 34, 4766–4775. https://doi.org/10.1523/JNEUROSCI.2828-13.2014

Clarke, A., Tyler, L.K., 2015. Understanding What We See: How We Derive Meaning From Vision. Trends in Cognitive Sciences 19, 677–687. https://doi.org/10.1016/j.tics.2015.08.008

Clarke, A., 2019. Neural dynamics of visual and semantic object processing, in: Psychology of Learning and Motivation. Elsevier, pp. 71–95. https://doi.org/10.1016/bs.plm.2019.03.002

Coggan, D.D., Baker, D.H., Andrews, T.J., 2016. The Role of Visual and Semantic Properties in the Emergence of Category-Specific Patterns of Neural Response in the Human Brain. eneuro 3, ENEURO.0158-16.2016. https://doi.org/10.1523/ENEURO.0158-16.2016

Connolly, A.C., Guntupalli, J.S., Gors, J., Hanke, M., Halchenko, Y.O., Wu, Y.-C., Abdi, H., Haxby, J.V., 2012. The Representation of Biological Classes in the Human Brain. Journal of Neuroscience 32, 2608–2618. https://doi.org/10.1523/JNEUROSCI.5547-11.2012

Connolly, A.C., Sha, L., Guntupalli, J.S., Oosterhof, N., Halchenko, Y.O., Nastase, S.A., di Oleggio Castello, M.V., Abdi, H., Jobst, B.C., Gobbini, M.I., Haxby, J.V., 2016. How the Human Brain Represents Perceived Dangerousness or "Predacity" of Animals. J. Neurosci. 36, 5373–5384. https://doi.org/10.1523/JNEUROSCI.3395-15.2016

Contier, O., Baker, C.I., Hebart, M.N., 2023. Distributed representations of behaviorally relevant object dimensions in the human visual system. bioRxiv. https://doi.org/10.1101/2023.08.23.553812

Devereux, B.J., Tyler, L.K., Geertzen, J., Randall, B., 2014. The Centre for Speech, Language and the Brain (CSLB) concept property norms. Behav Res 46, 1119–1127. https://doi.org/10.3758/s13428-013-0420-4

Devereux, B. J., Clarke, A., & Tyler, L. K. (2018). Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. Scientific Reports, 8(1). 10636. https://doi.org/10.1038/s41598018-28865-1.

Downing, P.E., Jiang, Y., Shuman, M., Kanwisher, N., 2001. A Cortical Area Selective for Visual Processing of the Human Body. Science 293, 2470–2473. https://doi.org/10.1126/science.1063414

DiCarlo, J.J., Zoccolan, D., Rust, N.C., 2012. How Does the Brain Solve Visual Object Recognition? Neuron 73, 415–434. https://doi.org/10.1016/j.neuron.2012.01.010

Eger, E., Ashburner, J., Haynes, J.-D., Dolan, R.J., Rees, G., 2008. fMRI Activity Patterns in Human LOC Carry Information about Object Exemplars within Category. Journal of Cognitive Neuroscience 20.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., ... Gorgolewski, K. J. (2020). FMRIPREP 20.2.1: A robust preprocessing pipeline for functional MRI. Nature Methods, 17(2), 111-116.

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., ... & Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Scientific data, 3(1), 1-9.

Grill-Spector, K., Weiner, K.S., 2014. The functional architecture of the ventral temporal cortex and its role in categorization. Nat Rev Neurosci 15, 536–548. https://doi.org/10.1038/nrn3747

Groen, I.I.A., Silson, E.H., Baker, C.I., 2017. Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. Phil. Trans. R. Soc. B 372, 20160102. https://doi.org/10.1098/rstb.2016.0102

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science, 293(5539), 2425-2430. doi: 10.1126/science.1063736.

Haynes, J.-D., Rees, G., 2005. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat Neurosci 8, 686–691. https://doi.org/10.1038/nn1445

Hebart, M.N., Bankson, B.B., Harel, A., Baker, C.I., Cichy, R.M., 2018. The representational dynamics of task and object processing in humans. eLife 7, e32816. https://doi.org/10.7554/eLife.32816

Hebart, M.N., Zheng, C.Y., Pereira, F., Baker, C.I., 2020. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. Nat Hum Behav 4, 1173–1185. https://doi.org/10.1038/s41562-020-00951-3

Hebart, M.N., Contier, O., Teichmann, L., Rockter, A.H., Zheng, C.Y., Kidder, A., Corriveau, A., Vaziri-Pashkam, M., 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. https://doi.org/10.7554/eLife.82580

Hebart, M. N., Kaniuth, P., & Perkuhn, J. (2022). Efficiently-generated object similarity scores predicted from human feature ratings and deep neural network activations. Journal of Vision, 22(14), 4057. https://doi.org/10.1167/jov.22.14.4057

Iordan, M.C., Greene, M.R., Beck, D.M., Fei-Fei, L., 2015. Basic Level Category Structure Emerges Gradually across Human Ventral Visual Cortex. Journal of Cognitive Neuroscience 27, 1427–1446. https://doi.org/10.1162/jocn_a_00790

Jagadeesh, A.V., Gardner, J.L., 2022. Texture-like representation of objects in human visual cortex. Proc. Natl. Acad. Sci. U.S.A. 119, e2115302119. https://doi.org/10.1073/pnas.2115302119

Jozwik, K.M., Kriegeskorte, N., Mur, M., 2016. Visual features as stepping stones toward semantics: Explaining object similarity in IT and perception with non-negative least squares. Neuropsychologia 83, 201–226. https://doi.org/10.1016/j.neuropsychologia.2015.10.023

Jozwik, K.M., Najarro, E., van den Bosch, J.J.F., Charest, I., Cichy, R.M., Kriegeskorte, N., 2022. Disentangling five dimensions of animacy in human brain and behaviour. Commun Biol 5, 1247. https://doi.org/10.1038/s42003-022-04194-y

Julian, J.B., Doeller, C.F., 2021. Remapping and realignment in the human hippocampal formation predict context-dependent spatial behavior. Nat Neurosci 24, 863–872. https://doi.org/10.1038/s41593-021-00835-3

Kanwisher, N., 2010a. Functional specificity in the human brain: A window into the functional architecture of the mind. Proc. Natl. Acad. Sci. U.S.A. 107, 11163–11170. https://doi.org/10.1073/pnas.1005062107

Khaligh-Razavi, S.-M., Kriegeskorte, N., 2014. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. PLoS Comput Biol 10, e1003915. https://doi.org/10.1371/journal.pcbi.1003915

Konkle, T., Oliva, A., 2012. A Real-World Size Organization of Object Responses in Occipitotemporal Cortex. Neuron 74, 1114–1124. https://doi.org/10.1016/j.neuron.2012.04.036

Konkle, T., Caramazza, A., 2013. Tripartite Organization of the Ventral Stream by Animacy and Object Size. Journal of Neuroscience 33, 10235–10242. https://doi.org/10.1523/JNEUROSCI.0983-13.2013

Kriegeskorte, N., Mur, M., Bandettini, P., 2008. Representational similarity analysis – connecting the branches of systems neuroscience. Front. Sys. Neurosci. https://doi.org/10.3389/neuro.06.004.2008

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. Commun. ACM 60, 84–90. https://doi.org/10.1145/3065386

Libby, L. A., Reagh, Z. M., Bouffard, N. R., Ragland, J. D., & Ranganath, C. 2019. The Hippocampus Generalizes across Memories that Share Item and Context Information. Journal of cognitive neuroscience, 31(1), 24–35. https://doi.org/10.1162/jocn_a_01345

Long, B., Yu, C.-P., 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. https://doi.org/10.1073/pnas.1719616115

Mehrer, J., Spoerer, C.J., Jones, E.C., Kriegeskorte, N., Kietzmann, T.C., 2021. An ecologically motivated image dataset for deep learning yields better models of human vision. Proc. Natl. Acad. Sci. U.S.A. 118, e2011417118. https://doi.org/10.1073/pnas.2011417118

Morgenstern, Y., Hartmann, F., Schmidt, F., Tiedemann, H., Prokott, E., Maiello, G., Fleming, R.W., 2021. An image-computable model of human visual shape similarity. PLoS Comput Biol 17, e1008981. https://doi.org/10.1371/journal.pcbi.1008981

Mumford, J.A., Turner, B.O., Ashby, F.G., Poldrack, R.A., 2012. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. NeuroImage 59, 2636–2643. https://doi.org/10.1016/j.neuroimage.2011.08.076

Muttenthaler, L., Hebart, M.N., 2021. THINGSvision: A Python Toolbox for Streamlining the Extraction of Activations From Deep Neural Networks. Front. Neuroinform. 15, 679838. https://doi.org/10.3389/fninf.2021.679838

Naselaris, T., Allen, E., Kay, K., 2021. Extensive sampling for complete models of individual brains. Current Opinion in Behavioral Sciences 40, 45–51. https://doi.org/10.1016/j.cobeha.2020.12.008

Nasr, S., Tootell, R.B.H., 2012. A Cardinal Orientation Bias in Scene-Selective Visual Cortex. J. Neurosci. 32, 14921–14926. https://doi.org/10.1523/JNEUROSCI.2036-12.2012

Olman, C.A., Davachi, L., and Inati, S. (2009). Distortion and signal loss in medial temporal lobe. PLOS ONE, 4(12), e8160. doi: 10.1371/journal.pone.0008160.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. Behavior research methods , 51 (1), 195-203.

Proklova, D., Kaiser, D., Peelen, M.V., 2016. Disentangling Representations of Object Shape and Object Category in Human Visual Cortex: The Animate–Inanimate Distinction. Journal of Cognitive Neuroscience 28, 680–692. https://doi.org/10.1162/jocn_a_00924

Proklova, D., Goodale, M.A., 2022. The role of animal faces in the animate-inanimate distinction in the ventral temporal cortex. Neuropsychologia 169, 108192. https://doi.org/10.1016/j.neuropsychologia.2022.108192

Ritchie, J.B., Bracci, S., Op De Beeck, H., 2017. Avoiding illusory effects in representational similarity analysis: What (not) to do with the diagonal. NeuroImage 148, 197–200. https://doi.org/10.1016/j.neuroimage.2016.12.079

Ritchie, J.B., Zeman, A.A., Bosmans, J., Sun, S., Verhaegen, K., Op de Beeck, H.P., 2021. Untangling the Animacy Organization of Occipitotemporal Cortex. J. Neurosci. 41, 7103–7119. https://doi.org/10.1523/JNEUROSCI.2628-20.2021

Schütt, H.H., Kipnis, A.D., Diedrichsen, J., Kriegeskorte, N., 2023. Statistical inference on representational geometries. eLife 12, e82566. https://doi.org/10.7554/eLife.82566

Silson, E.H., Zeidman, P., Knapen, T., Baker, C.I., 2021. Representation of Contralateral Visual Space in the Human Hippocampus. J. Neurosci. 41, 2382–2392. https://doi.org/10.1523/JNEUROSCI.1990-20.2020

Sha, L., Haxby, J.V., Abdi, H., Guntupalli, J.S., Oosterhof, N.N., Halchenko, Y.O., Connolly, A.C., 2015. The Animacy Continuum in the Human Ventral Vision Pathway. Journal of Cognitive Neuroscience 27, 665–678. https://doi.org/10.1162/jocn_a_00733

Singer, J.J.D., Cichy, R.M., Hebart, M.N., 2023. The Spatiotemporal Neural Dynamics of Object Recognition for Natural Images and Line Drawings. J. Neurosci. 43, 484–500. https://doi.org/10.1523/JNEUROSCI.1546-22.2022

Tarhan, L., Konkle, T., 2020. Reliability-based voxel selection. NeuroImage 207, 116350. https://doi.org/10.1016/j.neuroimage.2019.116350

Thaler, L., Schütz, A.C., Goodale, M.A., Gegenfurtner, K.R., 2013. What is the best fixation target? The effect of target shape on stability of fixational eye movements. Vision Research 76, 31–42. https://doi.org/10.1016/j.visres.2012.10.012

Theves, S., Fernández, G., Doeller, C.F., 2020. The Hippocampus Maps Concept Space, Not Feature Space. J. Neurosci. 40, 7318–7325. https://doi.org/10.1523/JNEUROSCI.0494-20.2020

Thorat, S., Proklova, D., Peelen, M.V., 2019. The nature of the animacy organization in human ventral temporal cortex. eLife 8, e47142. https://doi.org/10.7554/eLife.47142

Wang, L., Mruczek, R.E.B., Arcaro, M.J., Kastner, S., 2015. Probabilistic Maps of Visual Topography in Human Cortex. Cereb. Cortex 25, 3911–3931. https://doi.org/10.1093/cercor/bhu277

Weiner, K.S., Grill-Spector, K., 2012. The improbable simplicity of the fusiform face area. Trends in Cognitive Sciences 16, 251–254. https://doi.org/10.1016/j.tics.2012.03.003

Weiskopf, N., Hutton, C., Josephs, O., Deichmann, R., 2006. Optimal EPI parameters for reduction of susceptibility-induced BOLD sensitivity losses: A whole-brain analysis at 3 T and 1.5 T. NeuroImage 33, 493–504. https://doi.org/10.1016/j.neuroimage.2006.07.029

Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., DiCarlo, J.J., 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. U.S.A. 111, 8619–8624. https://doi.org/10.1073/pnas.1403112111

Figures legend

Figure 1: Stimulus set and experimental paradigm.
Figure 2: Definition of the ROIs.
Figure 3: Voxel reliability.
Figure 4: Decoding accuracy for an image-specific and category-specific SVM.
Figure 5: Reliability of representational similarity within and across exemplars.
Figure 6: Comparison between neural voxel patterns and semantic knowledge models.
Figure 7: 2D MDS representations on the multivoxel patterns.

1038 **Figure 1**: **Stimulus set and experimental paradigm.**

1039 ***A)*** Stimulus set: The 96 terrestrial mammal images used in the current study. Each of the 48
1040 animals is represented by two different exemplars. Each exemplar was assigned to one of
1041 two color contexts, either green (context 1, left) or red (context 2, right), indicated by the
1042 colored rectangle. Images were collected via web search with public domain licenses.
1043 Animals are shown in alphabetical order. **B)** Experimental paradigm: The experiment
1044 consisted of a memorization task (left) and a retrieval task (right). In the memorization part,
1045 participants passively viewed the stimuli in their respective color context and were asked to
1046 memorize each exemplar. In the retrieval part, participants viewed the stimuli on a gray
1047 background. Following each presentation, the participants were asked to assign the
1048 previously seen stimulus to the correct color context with a button press.

1049 **Figure 2: Definition of the ROIs.**

1050 We defined the following ROIs to describe the ventral visual pathway including extended
1051 ROIs: EVC (V1 – V3), lateral occipital cortex (LO), posterior fusiform (pFus), hippocampal
1052 cortex (HC) and perirhinal cortex (PRC). LO and pFus are together referred to as lateral
1053 occipital complex (LOC).

1054 **Figure 3**: **Voxel reliability.**

1055 **A)** Glass brain display of the whole-brain Spearman-Brown reliability estimates
1056 (unthresholded) averaged across all subjects using the reliability-based voxel selection
1057 method (Tarhan and Konkle, 2020). **B)** Barplot display of the fraction of reliable voxels
1058 exceeding our threshold of $r \geq 0.15$ in percentage (y-axis) in each ROI (x-axis).
1059 Error bars reflect 95% confidence intervals.

1060 **Figure 4: Decoding accuracy for an image-specific and category-specific SVM.**
1061 **A**) ROI-wise decoding accuracies of pairwise comparisons at the image-specific level (training

1062 and testing within exemplar) and the category-specific level (training and testing across
1063 exemplars of the same object). The red line indicates chance-level for pair-wise decoding
1064 (50%). **B**) Significant ROI-wise category-specific effects were additionally split for visual
1065 similarity, as defined by the representational similarity within an early convolutional DNN
1066 layer. Results reflect the top quartile and bottom quartile of visual similarity.
1067 Asterisks indicate significance when compared with zero at significance levels: *** $p < 0.001$,
1068 ** $p < 0.01$, * $p < 0.05$, Bonferroni-corrected. Error bars reflect 95% confidence intervals.

1069 **Figure 5**: **Reliability of representational similarity within and across exemplars.**
1070 **A)** Analysis approach for comparing representational similarity within and across exemplars,
1071 with reliability estimated across scanning sessions. **B)** Pattern reliability within and across
1072 exemplars, estimated across scanning sessions, where across exemplar comparisons reflect a
1073 representational geometry that is abstracted away from individual images. **C)** Pattern
1074 reliability within and across exemplars but without reliability-based voxel selection, showing
1075 overall smaller effects yet a similar pattern of results. This suggests that reliability-based
1076 voxel selection did not induce any specific bias. **D)** Comparison of the across-exemplar
1077 correlation effects before and after solely partialling out the early convolutional layer RDM
1078 from AlexNet ("low-level-controlled") as well as after additionally partialling out the shape
1079 model ("low/mid-level-controlled"). Results are shown only for ROIs with significant across
1080 exemplar reliability.
1081 Asterisks indicate significance when compared with zero at significance levels: *** $p < 0.001$,
1082 ** $p < 0.01$, * $p < 0.05$, Bonferroni-corrected. Error bars reflect 95% confidence intervals.

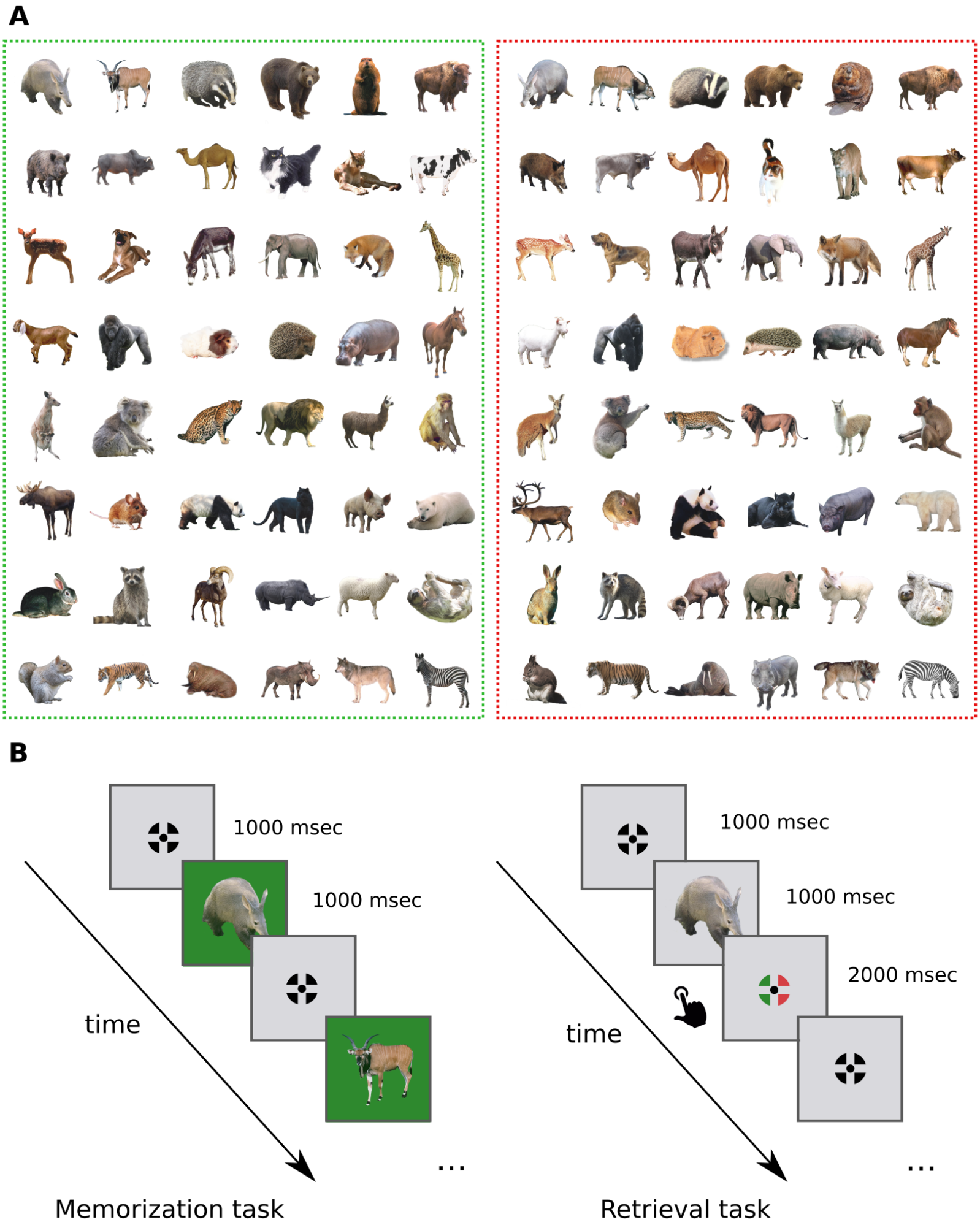
1083 **Figure 6**: **Comparison between neural voxel patterns and semantic knowledge models.**
1084 **A)** Analysis approach for uncovering semantic information in the representational geometry
1085 of the fMRI signal. Two embedding model RDMs (THINGS embedding, CSLB) were correlated

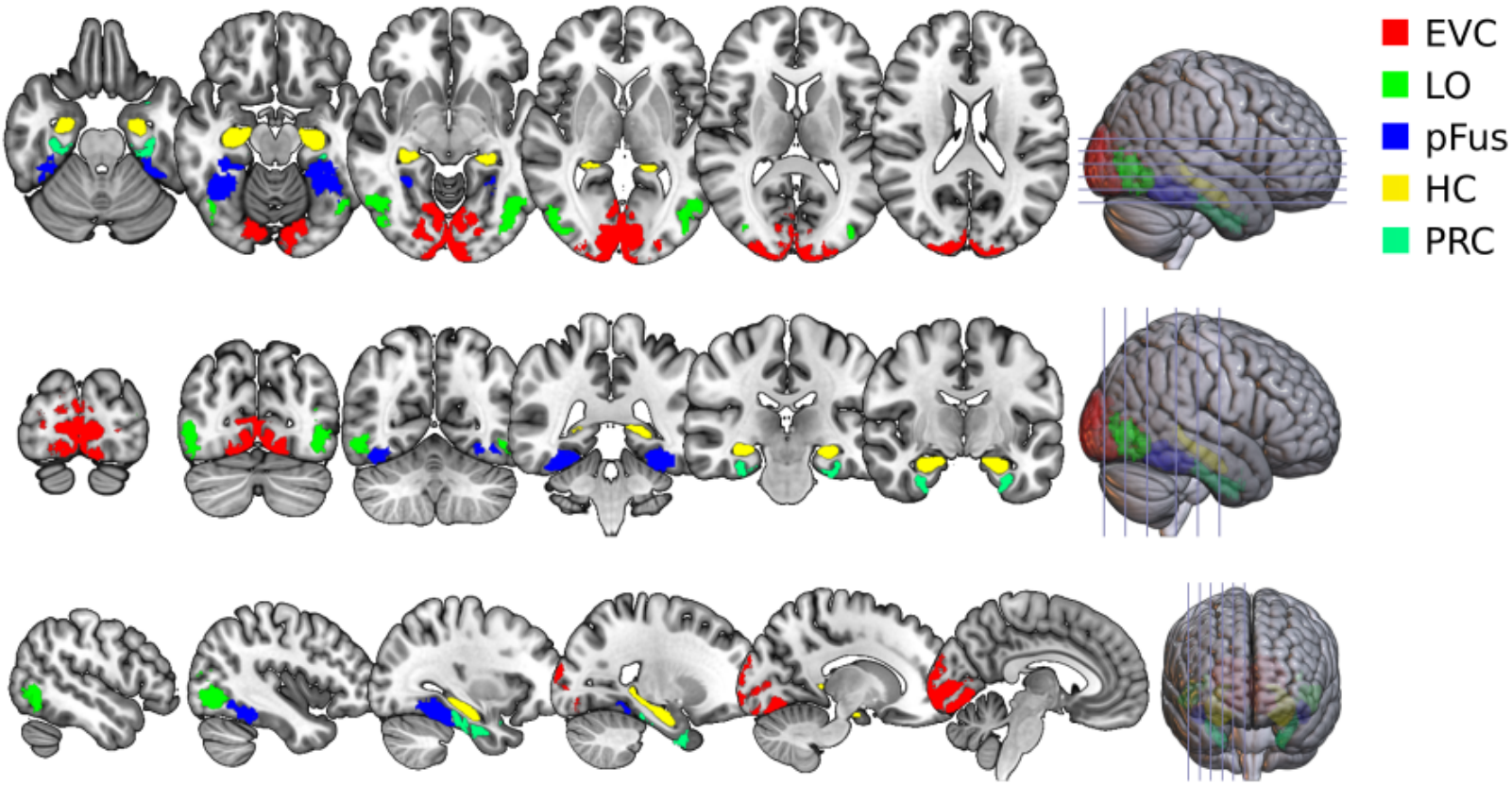
with the ROI-specific fMRI RDMs to detect significant correspondence. **B)** Spearman correlation values between the THINGS embedding (left) as well as the CSLB semantic (right) model RDMs and the ROI-specific pattern RDMs from each session and each participant. Effects are shown without variance partitioning ("THINGS", "CSLB"), after solely partialling out the early convolutional layer RDM from AlexNet ("low-level-controlled") as well as after additionally partialling out the shape model ("low/mid-level-controlled"). Results are shown only for ROIs with significant across exemplar reliability.

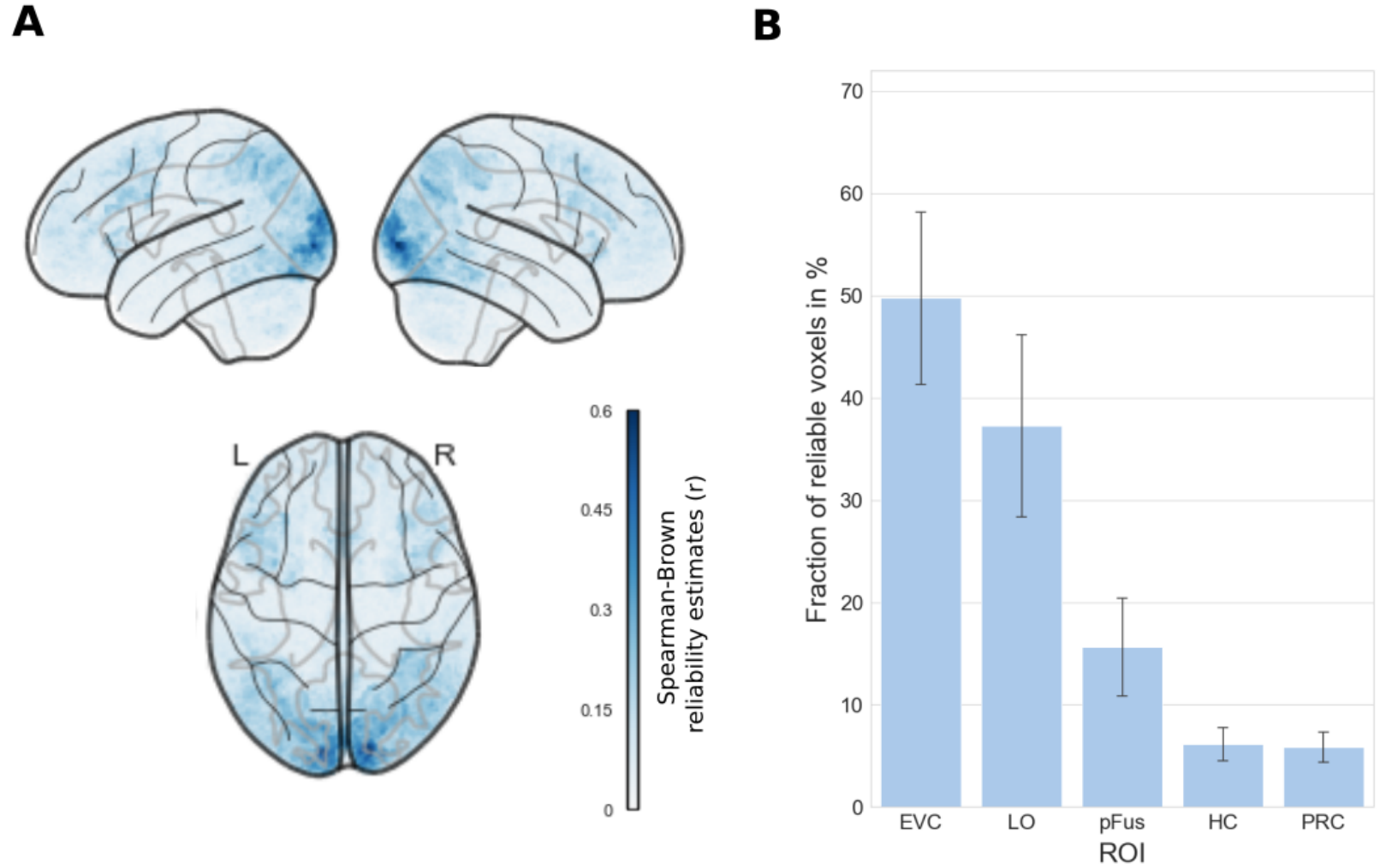
Asterisks indicate significance when compared with zero at significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, Bonferroni-corrected. Error bars reflect 95% confidence intervals.

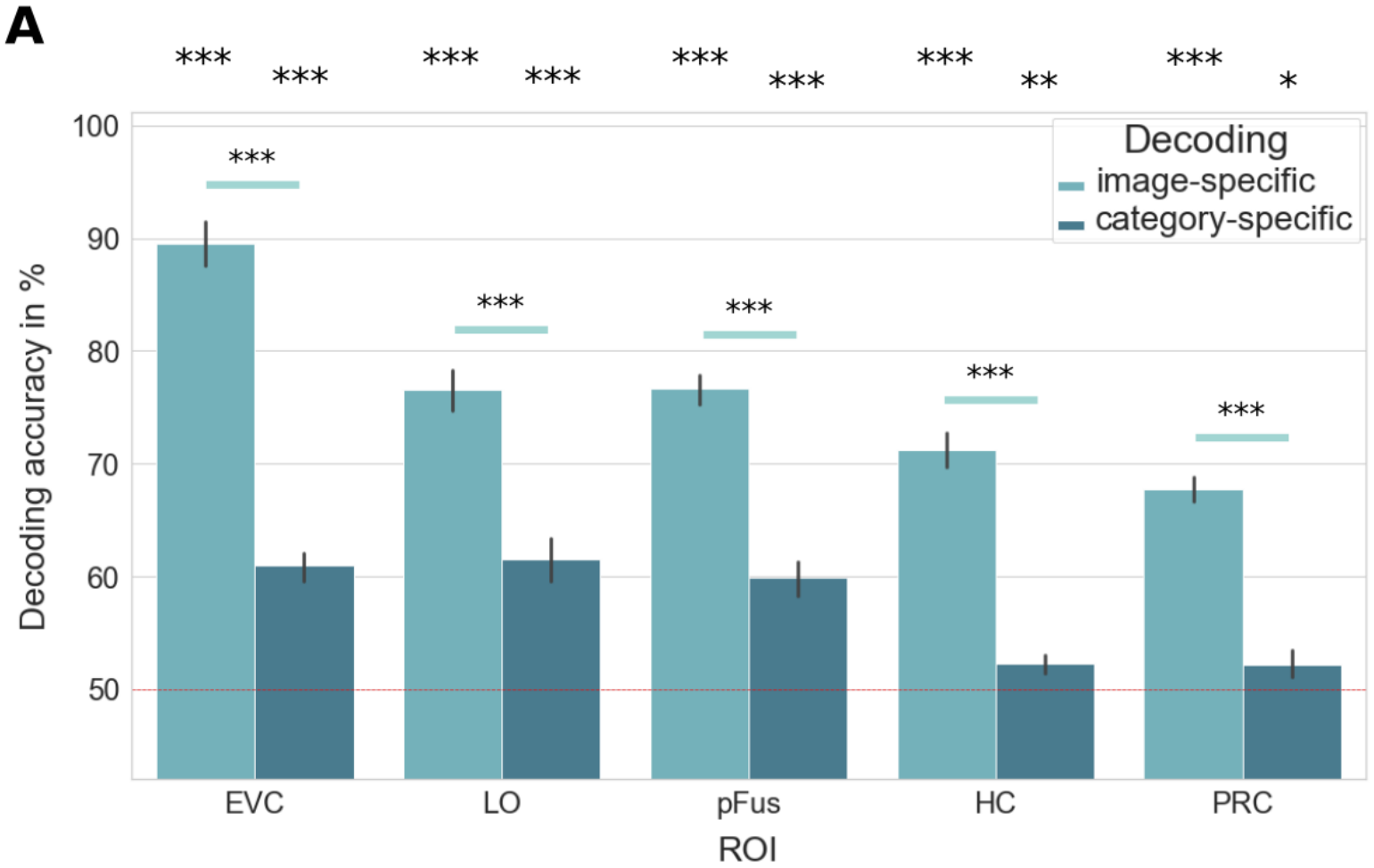
Figure 7: 2D MDS representations on the multivoxel patterns.

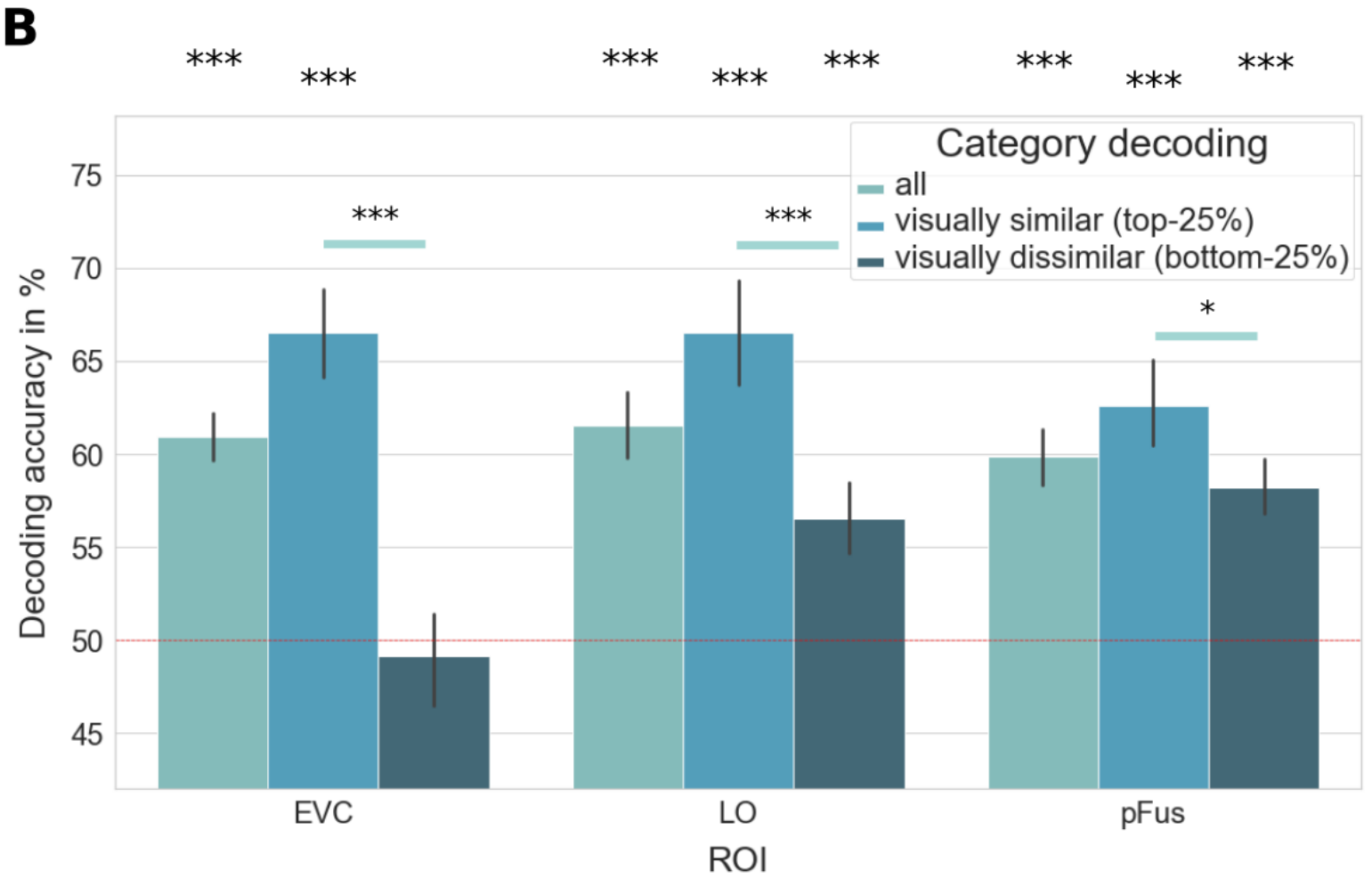
A 2-dimensional multidimensional scaling (MDS) display derived from the average RDMs across all participants and sessions for each ROI showing pairwise distances between all stimuli. The numbering ranges from 1 to 48 and displays the basic-level mammal category arranged in alphabetical order (as in Fig. 1a). The numbers are colored to indicate the color-context each stimulus belongs to.

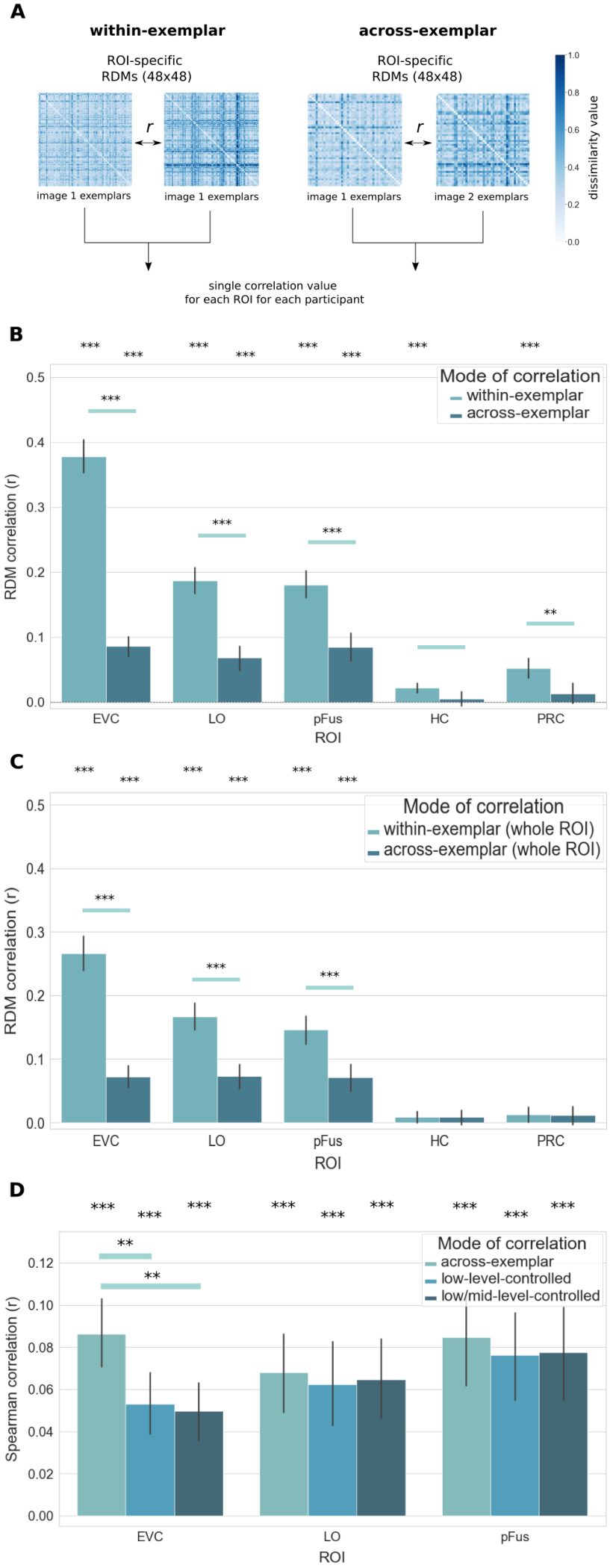

A
B
1000 msec
1000 msec
time
...
Memorization task
1000 msec
1000 msec
2000 msec
time
...
Retrieval task

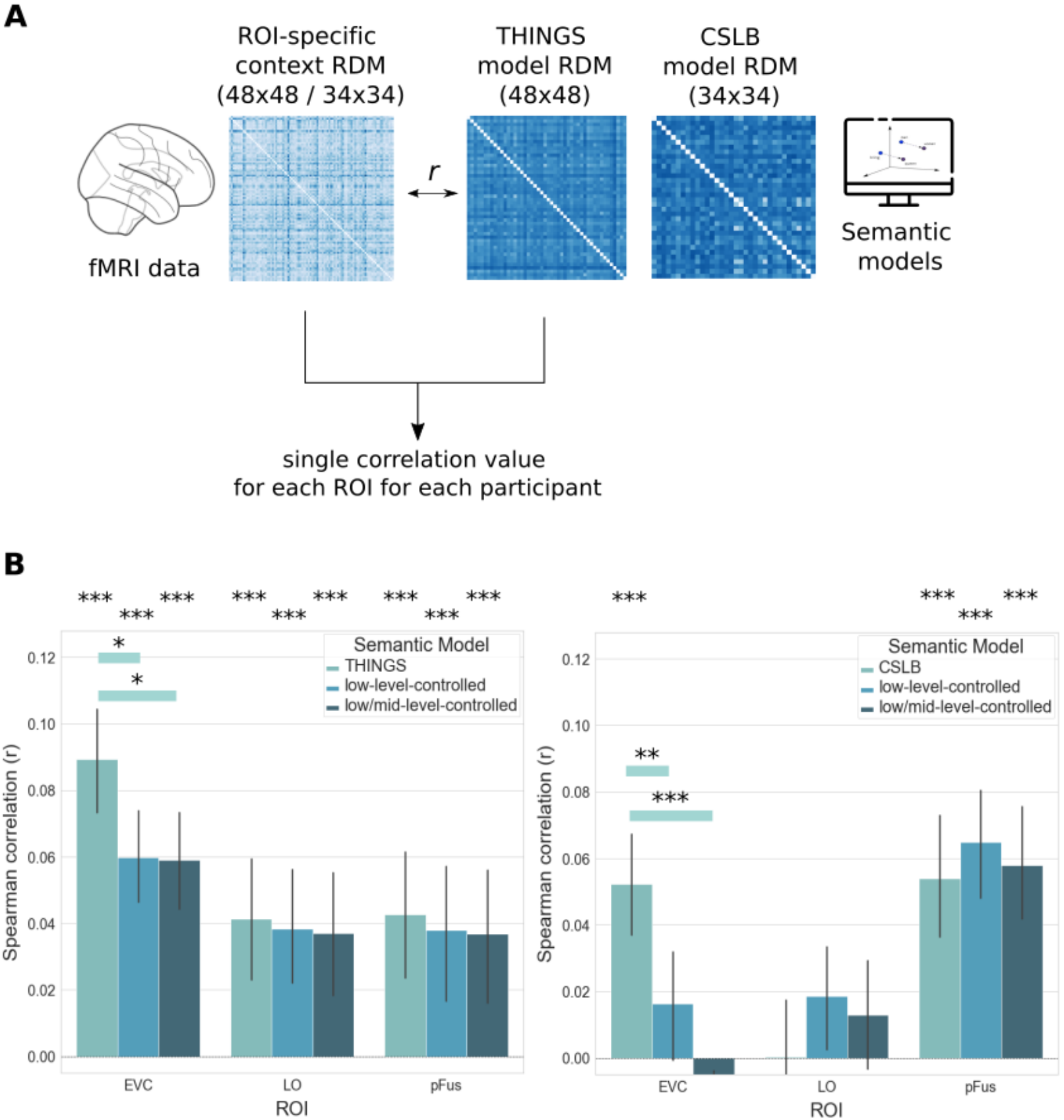

EVC
LO
pFus
HC
PRC

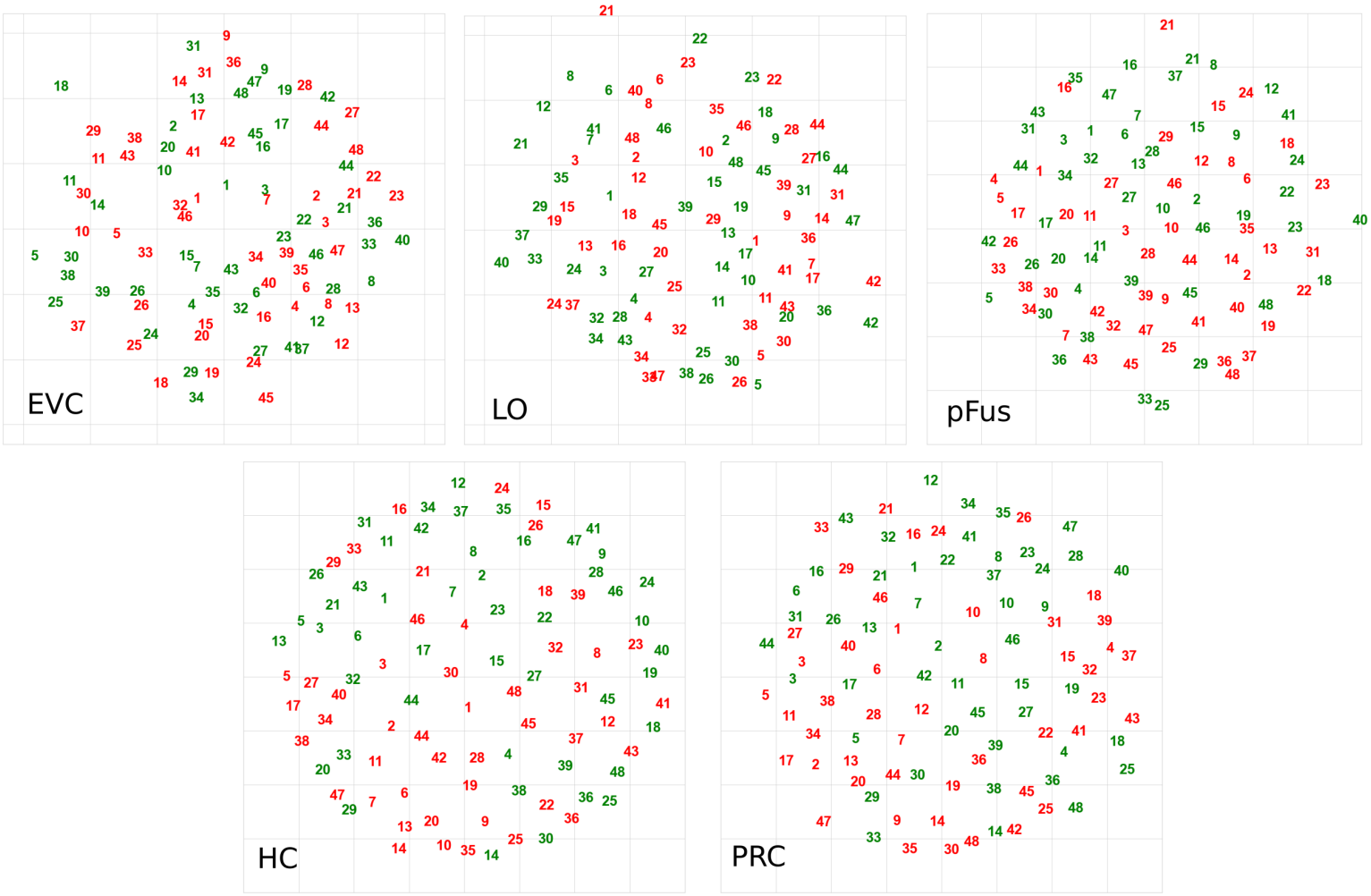

A
L
R
0.6
0.45
0.3
0.15
0
Spearman-Brown
reliability estimates (r)
B
Fraction of reliable voxels in %
70
60
50
40
30
20
10
0
EVC
LO
pFus
HC
PRC
ROI


A
Decoding
image-specific
category-specific
Decoding accuracy in %
100
90
80
70
60
50
EVC
LO
pFus
HC
PRC
ROI


B
Category decoding
all
visually similar (top-25%)
visually dissimilar (bottom-25%)
Decoding accuracy in %
75
70
65
60
55
50
45
EVC
LO
pFus
ROI


A
within-exemplar
across-exemplar
ROI-specific RDMs (48x48)
ROI-specific RDMs (48x48)
r
r
image 1 exemplars
image 1 exemplars
image 1 exemplars
image 2 exemplars
dissimilarity value
1.0
0.8
0.6
0.4
0.2
0.0
single correlation value for each ROI for each participant
B
Mode of correlation
within-exemplar
across-exemplar
RDM correlation (r)
0.5
0.4
0.3
0.2
0.1
0.0
EVC
LO
pFus
HC
PRC
ROI
C
Mode of correlation
within-exemplar (whole ROI)
across-exemplar (whole ROI)
RDM correlation (r)
0.5
0.4
0.3
0.2
0.1
0.0
EVC
LO
pFus
HC
PRC
ROI
D
Mode of correlation
across-exemplar
low-level-controlled
low/mid-level-controlled
Spearman correlation (r)
0.12
0.10
0.08
0.06
0.04
0.02
0.00
EVC
LO
pFus
ROI


A
ROI-specific context RDM (48x48 / 34x34)
THINGS model RDM (48x48)
CSLB model RDM (34x34)
r
fMRI data
Semantic models
single correlation value for each ROI for each participant
B
Semantic Model
THINGS
low-level-controlled
low/mid-level-controlled
CSLB
Spearman correlation (r)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
EVC
LO
pFus
ROI


EVC
LO
pFus
HC
PRC