Neural Selectivity for Real-World Object Size In Natural Images

Abbreviated title: Neural Selectivity for Real-World Size

Andrew F. Luo (afluo@andrew.cmu.edu)
Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, 15213, USA

Leila Wehbe (lwehbe@cmu.edu)
Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, 15213, USA

Michael J. Tarr (michaeltarr@cmu.edu)
Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, 15213, USA

Margaret M. Henderson (mmhender@cmu.edu)
Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, 15213, USA

1 Abstract

Real-world size is a functionally important high-level visual property of objects that supports interactions with our physical environment. Critically, real-world-size is robust over changes in visual appearance as projected onto our retinae such that large and small objects are correctly perceived to have different real-world sizes. To better understand the neural basis of this phenomenon, we examined whether the neural coding of real-world size holds for objects embedded in complex natural scene images, as well as whether real-world size effects are present for both inanimate and animate objects, whether low- and mid-level visual features can account for size selectivity, and whether neural size tuning is best described by a linear, logarithmic, or exponential neural coding function. To address these questions, we used a large-scale dataset of fMRI responses to natural images combined with per-voxel regression and contrasts. Importantly, the resultant pattern of size selectivity for objects embedded in natural scenes was aligned with prior results using isolated objects. Extending this finding, we also found that size coding exists for both animate and inanimate objects, that low-level visual features cannot account for neural size preferences, and size tuning functions have different shapes for large versus small preferring voxels. Together, these results indicate that real-world size is an ecologically significant dimension in the larger space of behaviorally-relevant cortical representations that support interactions with the world around us.

2 Introduction

Over the past several decades, one of the most striking results in visual neuroscience has been the discovery of cortical regions that preferentially encode object categories (e.g., faces, words, scenes, food; Cohen et al. (2000), Epstein and Kanwisher (1998), Jain et al. (2023), Sergent et al. (1992)) and object properties such as object animacy, real-world size, and reachability (Gallivan et al., 2009, Konkle and Caramazza, 2013, Warrington and Shallice, 1984). While these preferences are typically associated with the process of object recognition, visually invariant, functionally-grounded representations of object properties are also critical across a wide variety of real-world behaviors.

Real-world size is one of the most salient high level properties that supports our ability to interact with objects in the environment (e.g., mistaking tigers for house cats rarely ends well). Importantly, the perception of real-world size remains roughly invariant across variations in lighting, viewpoint, positional displacement, and distance. In that humans intuitively understand and experience real-world size as constant as we move around, farther from or closer to objects (Slater et al., 1990), real-world size estimates are essential in inferring even higher level, behaviorally relevant properties (Grèzes et al., 2003, Tucker and Ellis, 2001). For example, object affordances specify how we can interact with objects, enabling us to perceive small objects as things we can pick up and larger objects as things to be navigated around.

Despite ecologically-grounded arguments for the importance of real-world size, studies demonstrating neural selectivity for real-world size have relied on highly controlled and structured experimental setups, presenting object stimuli in isolation and without naturalistic backgrounds that might provide a variety of visual and contextual cues as to real-world size. In contrast, human perception of real-world size is context dependent and can vary due to background information and the presence of other objects (Bar, 2004). For example, the 3D Ponzo illusion investigated by Murray et al. (2006) leads objects that occupy the same visual angular size, but that appear with different surface contextual cues, to be perceived and neurally represented as having different real-world sizes.

The first question we address is whether selectivity for real-world size can be measured using natural scenes depicting many objects in relation to one another in the context of complex backgrounds. A related issue is whether Konkle and Caramazza (2013)'s finding of real-world size coding only for inanimate objects is a consequence of the decontextualized, non-ecological stimulus conditions used in prior studies. Thus, the second question we address is whether size selectivity can be measured for animate objects (i.e., small animals and large animals) when depicted in natural scenes.

The third question we address is whether neural selectivity for real-world size reflects responses driven by the high level property itself. Alternatively, are these responses driven by visual features associated with different real-world sizes, that is, large and small objects (e.g., curvature and boxiness; Nasr and Tootell, 2012, Nasr et al., 2014)? Supporting this alternative, past work has suggested that mid-level textural features may be sufficient for behavioral judgments of object real-world size (Long and Konkle, 2017, Long et al., 2016), and images containing only mid-level features can elicit patterns of neural selectivity that resemble real-world size selectivity for intact objects (Long et al., 2018). Mid-level features, however, may manifest in different ways for natural scene images as opposed to isolated object images, and it is not clear how strongly mid-level features contribute to real-world size selectivity when considering images of objects embedded in natural contexts.

The fourth question we address pertains to the shapes of real-world size tuning functions in the brain and whether these tuning functions differ in shape between small-preferring and large-preferring neural populations. In particular, preferential responses to different sized objects might be supported by different neural coding schemes that enhances the separability within each domain. In contrast, past studies have only identified which regions of the brain exhibit real-world size selectivity from a qualitative perspective (i.e., large or small).

To preview our findings, voxel-wise analyses using natural images reveal a pattern of real-world size selectivity that is aligned with past studies utilizing single-object images. We also establish a link between size selectivity and the functional roles of different category-selective brain areas. Interestingly, using natural images, we find evidence for size selectivity within the category of animate objects, albeit with a weaker effect than that obtained for inanimate objects. At the same time, we find evidence that selectivity for real-world size is not driven by the presence of correlated low- or mid-level visual features: when controlling for such features we continue to obtain consistent patterns of size selectivity, suggesting that such real-world size effects are largely independent of these features. Finally, a voxel-wise search of monotonic functions (logarithmic, linear, and exponential) to characterize the shape of size tuning functions reveals a difference in the coding used for large and small preferring voxels. Overall, these results provide new evidence for real-world object size as an organizing dimension in natural image representation, and contribute to a more detailed understanding of how real-world size is computed and encoded in the brain.

3 Materials & Methods

fMRI data. We primarily use the Natural Scenes Dataset (NSD; Allen et al. (2022)), with additional visualization and analysis using BOLD5000 (Chang et al., 2019). The detailed data collection procedure can be found in Allen et al. (2022). We will first discuss NSD. Natural images with semantic mask annotations were sourced from the Microsoft Common Objects in Context (MS COCO) dataset. Whole-brain fMRI data was collected using a 7T scanner. Eight subjects each viewed between 9,209 and 10,000 images, with each image repeated three times. Of 70,566 total unique images presented, a total of 1,000 images were viewed by all subjects. Single trial betas were generated using GLMsingle (Prince et al., 2022). This process estimates a per-voxel hemodynamic response function. Cross-validation and ridge regression steps are further applied to estimate the beta coefficients. We compute the mean and standard deviation of each voxel across a given scanning session, and use these values to compute a z-value for each voxel corresponding to each trial. Z-values are averaged across repetitions of an image (approximately 3 trials/image) to compute an averaged response value for each voxel/image combination. In several analyses we summarize our data using functionally-defined regions of interest (ROIs). These include early visual areas (V1, V2, V3, V4), scene-selective areas (RSC, PPA, OPA), body-selective areas (EBA, FBA), face-selective areas (OFA, FFA), and word-selective areas (OWFA, VWFA). These ROI definitions are obtained using independent retinotopic mapping and category localizer scans performed as part of the NSD experiment (see Allen et al. (2022) for details on ROI definitions). For the BOLD5000 dataset, we similarly use GLMsingle derived beta coefficients, and compute the same normalization and z-value calculation. We restrict our BOLD5000 data to visual stimuli that were derived from the MS COCO dataset.

Object real-world size assignment. As the first step in generating real-world size labels for each image in the stimulus set, we assign each semantic object category in COCO a real-world size estimate. We query `GPT-3` for the real-world size of objects using the prompt "What is average size of a `TYPE`", where `TYPE` is the semantic category of an object. We repeat each prompt ten times. When multiple values such as length, width, and height are returned for an object, we select the largest value. All units are converted to centimeters, and averaged. These dimensions are further manually cross-checked against the "dimensions" online database for consistency (dimensions, 2022).

Saliency guided object selection. A property of natural images is the potential presence of multiple objects within a single image. To allow each image to be assigned a single real-world size label, we select the single most salient object within each image. Visual saliency is estimated using a neural network trained on the SALICON dataset (Jiang et al., 2015). The architecture of the network consists of stacked 2D convolutional layers (Reddy et al., 2020). The saliency prediction is represented as a single channel image with values between $(0, 1)$, with the same width and height as the input image. Individual object masks are rendered from COCO metadata. For each object, we compute a spatial integral of the saliency over the area defined by the current mask. The object with the maximum saliency integral is retained. To avoid degenerate objects, we do not consider images where the most salient object is less than 0.5% of the image area. Our approach is illustrated in Figure 1.

The final real-world size label for each image is then determined using the estimated real-world size (see previous section) of the object category corresponding to the most salient object. In all analyses, we follow Konkle and Oliva (2012) and exclude any images in which the most salient object is a person, in order to avoid potential biasing effects of this label ("person" is the most common object label in COCO by a substantial margin, and images of people are known to evoke large neural activations in several brain areas; e.g. Kanwisher et al. (1997)). In preliminary analyses, similar results were obtained when "person" images were included. Our combined criteria yields between 5990 and 6631 images for each subject in NSD, and between 814 and 1339 images for each subject in BOLD5000. In our analysis of animate versus inanimate objects, we assigned an object class as animate if it was an animal. For example, objects such as `Cat` and `Dog` are treated as animate, while `Apple` and `Car` are treated as inanimate. Animacy assignment for an image is similarly based on the most salient object. When we

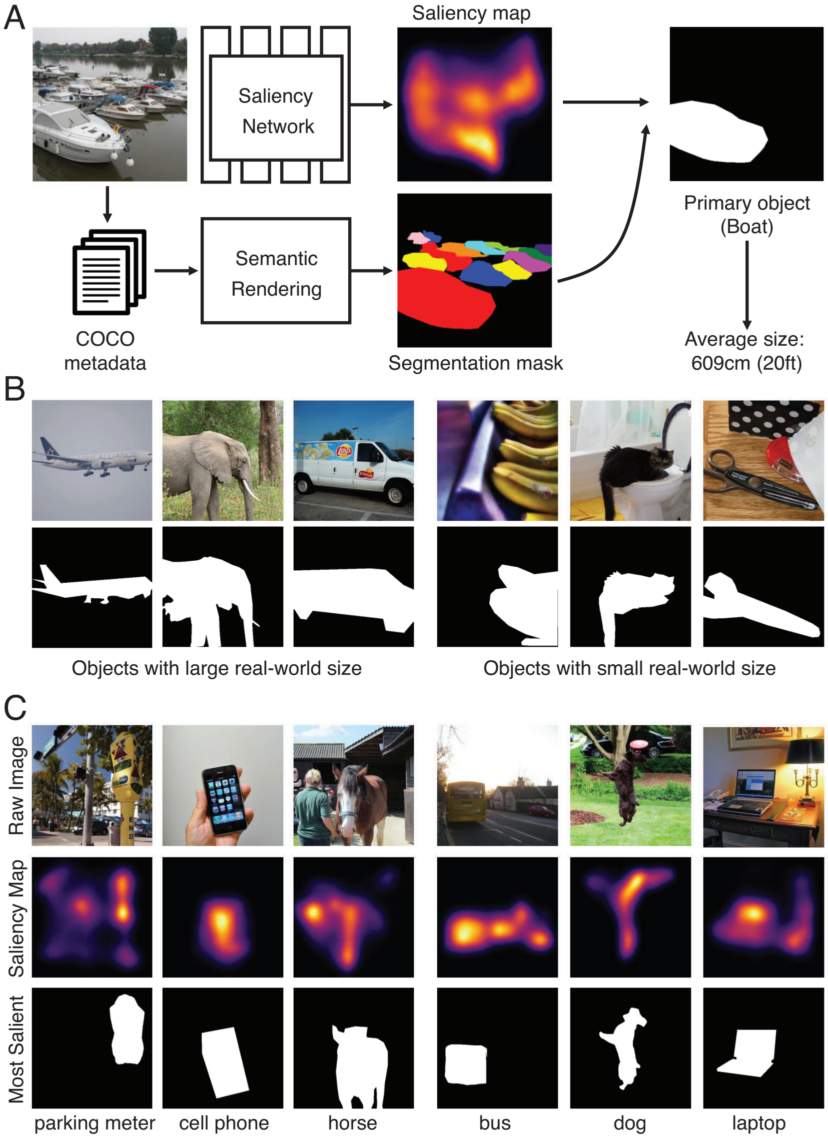

Figure 1: **Labeling real-world size in natural scenes using saliency guided primary object identification. (A)** An image is passed to a convolutional neural network trained to predict human saliency from the SALICON dataset. By combining the semantic mask and saliency information, we identify the size of the most salient object (see *Methods* for details). **(B)** Examples of objects with large and small real-world size. Original color stimulus shown on the top, and the binary mask of the object shown in the bottom. **(C)** Example outputs of our saliency extraction pipeline.

compute the inanimate and animate correlations respectively, for each subject, the number of inanimate images are between 4409 and 4887, while the number of animate images are between 1619 and 2028.

We consider the correlation coefficients when comparing the response of animate to inanimate objects, a metric that is robust to sample size.

Size selectivity analysis. To identify size selective voxels, we perform voxel-wise two-tailed t-tests. To binarize the real-world size labels, we compute a median object size across all valid images, and select images where the most salient object is smaller or larger than this median size. The p-value across all cortical voxels in a subject were further corrected using the Benjamini/Hochberg False Discovery Rate (FDR) method. Results were computed per-subject.

In addition to computing selectivity with these binarized size labels, we use the continuous estimates of real-world size for each image to compute the Spearman's rank correlation coefficient between each voxel's activation and real-world size. This is used to measure the monotonicity between object size and the voxel activation, without making any assumptions about the linearity of real-world size values or neural activation. To evaluate the significance of ROI averaged correlation coefficients, we utilize a one-sample two-tailed t-test to compute a p-value for each ROI. For visualization, we project values onto cortical surface meshes using PyCortex software (Gao et al., 2015).

Controlling for low-level and mid-level visual features. In order to address possible contributions of low and mid-level visual features to our measured real-world size selectivity, we construct voxel-wise encoding models based on several low and mid-level feature spaces: a Gabor model, a GIST model, and a "sketch tokens" model. Both the Gabor and GIST model are chosen as examples of "low-level" feature spaces, as they are each based on orientation and spatial frequency features, while the sketch tokens model is chosen as "mid-level" feature space – it utilizes a computer vision model that learns to categorize image contour patches into a set of data-driven clusters (Lim et al., 2013).

For both the Gabor and sketch tokens models, we use a fitting procedure that takes into account the spatial receptive field of each voxel, or its population receptive field (pRF). An estimate of each voxel's pRF is obtained based on an independent retinotopic mapping task (`prffloc`) performed as part of the NSD experiment (see Allen et al. (2022) and Benson et al. (2018) for details on the task and pRF estimation). The parameters of this estimated pRF (a 2D Gaussian with parameters (μ_x, μ_y)=center, σ=size), are then quantized by rounding the original parameter value to the nearest value in a pre-specified grid over candidate pRF values (this quantization is done to reduce computing time). Within the pRF parameter grid, the size and eccentricity values are each logarithmically distributed with 10 possible values each, with eccentricity ranging from 0° - 7° (° = degrees visual angle), and size ranging from 0.17° to 8.4°. Angular position values are evenly spaced along 16 possible angular values. The quantized parameterization yields a total of $1,456$ possible size/eccentricity/angular position combinations (some position/size combinations are excluded due to being non-overlapping with the image extent).

The pRF estimates are then used to compute a single feature vector for each voxel and each image, by taking a weighted sum between the voxel's pRF (a 2D Gaussian) and a stack of feature maps corresponding to the image of interest (see St-Yves and Naselaris (2017) for details on this approach).

For the Gabor model, these maps are computed at 12 orientations and 8 spatial frequencies, for a total of 96 spatial feature maps (see Henderson et al. (2022) for more details on model construction). For the sketch tokens model, there are 151 spatial feature maps, each indicating the presence of a specific class of contour feature.

For the GIST model, we do not include an estimate of the voxel's pRF, because this model incorporates features from the entire image. The GIST model is implemented using Matlab code provided by Oliva and Torralba (2001); we use an implementation with a 4×4 spatial grid, 4 spatial frequencies, and 8 orientations, for a combined 512 dimensional feature vector.

For each model, we construct a linear model with bias to map between the feature vector and the voxel activity for each subject. We fit this model using ordinary least squares on visual stimulus for each subject. Using this model, we generate predicted responses of each voxel for each image, and subtract the predictions from the actual brain responses to obtain the residual. Using these residuals and the methods described previously, we estimate the average Spearman correlation between each voxel's residual activity with real-world size. The correlation coefficients are averaged for all voxels in an ROI, and have p-values computed using a one-sample two-tailed t-test. The p-values are FDR corrected across subjects.

Size tuning profiles. To characterize the shape of tuning profiles for object real-world size, we construct functions ($f(x)$) that transform real-world size (x). We search over functions of the parameteric form $f(x) = x^{\gamma}$ where $\gamma \in [\frac{1}{10}, \frac{1}{5}, \frac{1}{2}, 1, 2, 5, 10]$. These functions are monotonically increasing, thus the output preserves the relative ordering of the input. These functions also map input values in the range of $[0,1]$ to outputs in the range of $[0,1]$. Prior to input to $f(\cdot)$ we scale size to lie between $[0,1]$ across the entire dataset. An ordinary least squares is performed to map the transformed size $f(x)$ to the voxel activity for each subject. We perform the fitting on the images unique to each subject. A portion of the images, consisting of the 602 criteria-satisfying images that were shared across all subjects, was left out while fitting the parameters of this model, and the final explained variance (R^2) was computed on this held out data.

4 Results

4.1 Experiment 1: Analyses of object size selectivity in a naturalistic setting

We first examine the size selectivity of voxels in the context of all valid images (including both animate and inanimate objects). We identify size-selective voxels using a two-sample two-tailed t-test comparing responses to large and small objects (see *Methods*; voxels retained if $p < 0.01$ after FDR correction across the cortical surface). For the voxels in which size selectivity is significant, the direction and magnitude of size selectivity is estimated using the Spearman rank correlation coefficient between voxel activation and object size. Averaged results are presented in Figure 2A, while single subject results are presented in Figure 3. Across all subjects, we find size-selective voxels on the ventral, lateral, and dorsal surfaces of occipitotemporal cortex. We observe a gradient of large-to-small size preference along the ventral surface of occipitotemporal cortex: voxels with large selectivity tend to cluster medially, while more lateral regions on the ventral surface tend to be more small selective. In the single-subject results (Figure 3), we also observe a small cluster of large-selective voxels surrounded by small-selective voxels. This cluster partially overlaps with FFA. Due to variability across subjects, this cluster is minimally present when the correlation coefficients are averaged across subjects. On the lateral surface, we observe alternating patches of small and large selective voxels. We observe that V1-V4 (early visual areas) contain slightly more small-selective voxels, with some large-selective voxels present towards the foveal (posterior) portion of these areas.

Looking at the proportion of small and large-selective voxels within functionally-defined ROIs

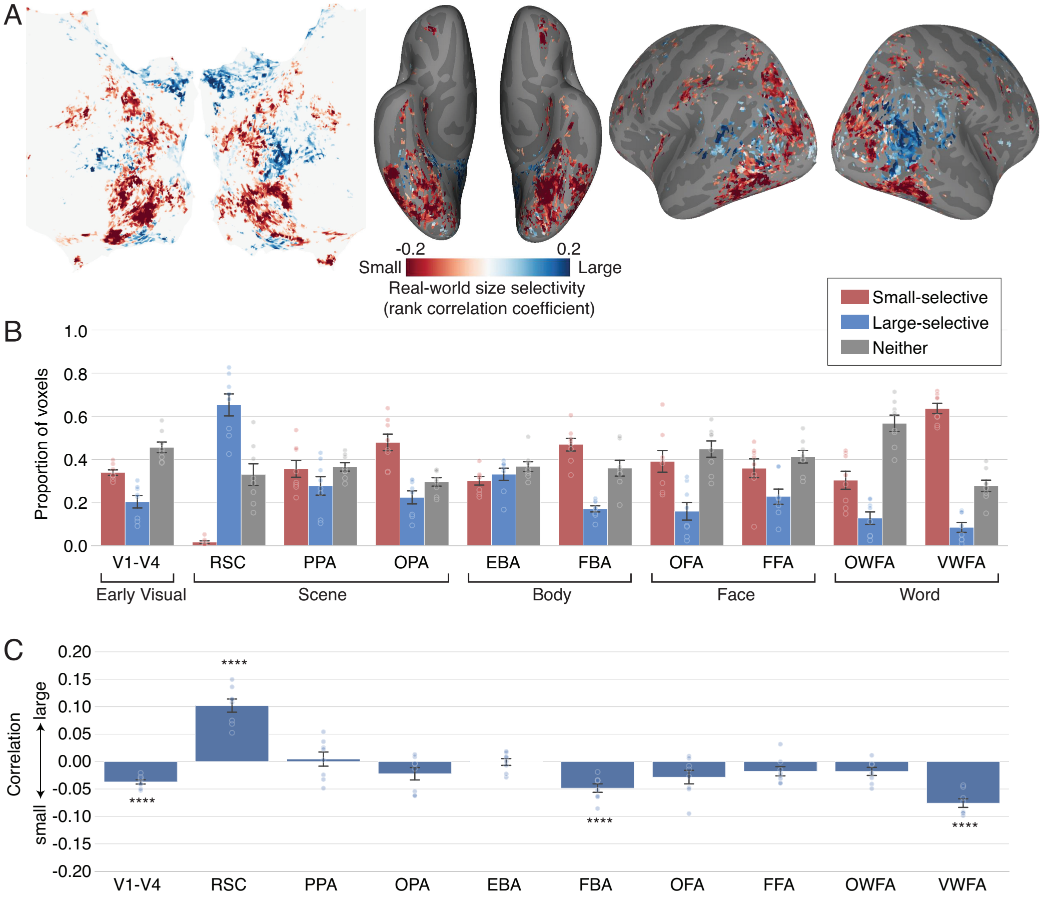

Figure 2: **Neural selectivity for real-world object size. (A)** Flattened, ventral, and expanded visualization of size selectivity using the average of eight subjects in NSD. Voxels having significant size selectivity (evaluated using a two-tailed t-test; $p < 0.01$; FDR corrected across cortex) for two or more subjects are visualized, with color determined using the Spearman rank correlation of the voxel's activity and real-world size of the most salient object. Real-world size is a continuous value estimated using GPT-3; see Figure 1 and *Methods* for more details. Blue indicates selectivity for large objects, and red indicates selectivity for small objects. **(B)** Proportion of voxels in each ROI that exhibit significant selectivity for size. **(C)** Correlation coefficient averaged across all voxels in each ROI. A negative correlation indicates selectivity for small objects, while a positive correlation indicates selectivity for large objects. **** indicates $p < 0.01$ (FDR corrected across subjects) under a one-sample t-test. In both **(B)** and **(C)**, bar heights and errorbars show mean ±1 SEM across subjects, circles represent individual subjects.

(Figure 2B), we find that RSC is the region with the highest ratio of large selective voxels, having far more voxels selective for large objects than small objects. Other scene-selective regions PPA and OPA have more voxels selective for small objects than large. We also find that VWFA is mainly composed of

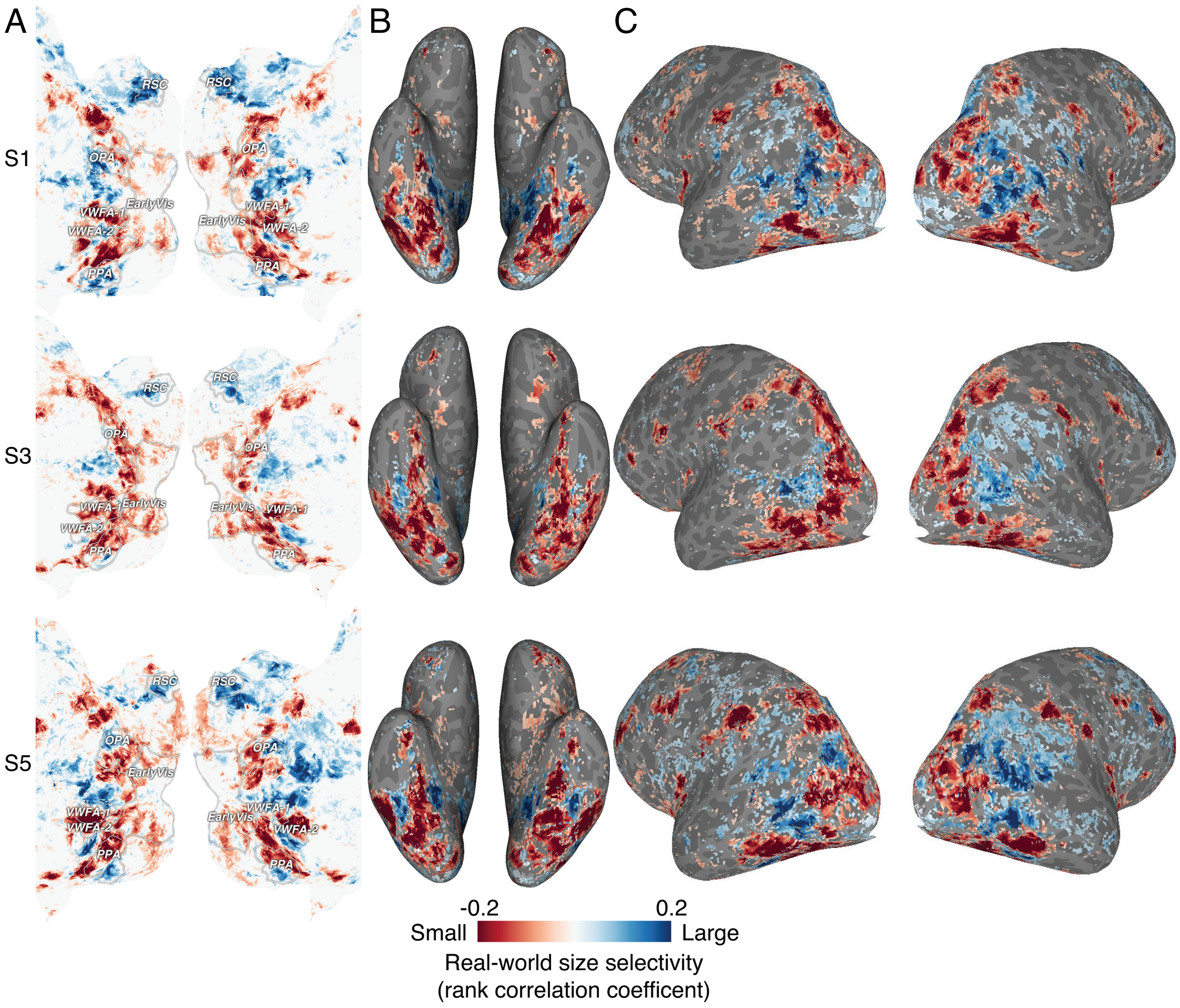

Figure 3: **Neural selectivity for real world object size is consistent within individual subjects.** We visualize the size selectivity estimates (as in Figure 2A) for three individual example subjects, and find the pattern to be largely consistent across subjects, with some additional detailed structure evident. **(A)** Flat map view. **(B)** Ventral view. **(C)** Expanded lateral view. In each panel, blue indicates selectivity for large objects, red indicates selectivity for small objects. Similar results were obtained for the other NSD subjects.

voxels that are small selective, and is the ROI with the highest ratio of small selective voxels. Computing the average correlation coefficients within these ROIs (Figure 2C), we find that V1-V4 is small-selective (V1-V4: $t_7 = 9.510; p = 0.0001$). We find a significantly greater response to large objects in RSC

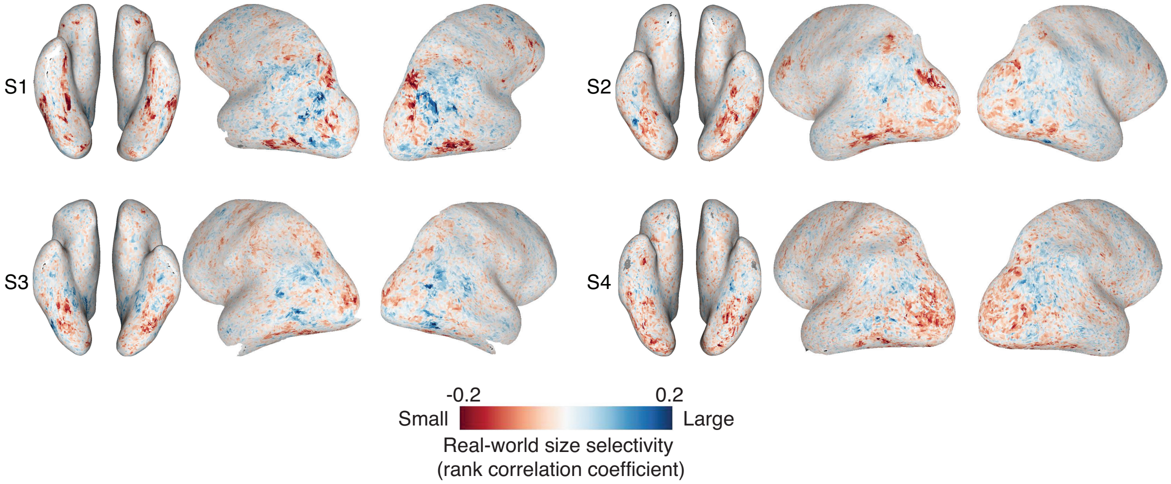

Figure 4: **Neural selectivity for real world size estimated using the BOLD5000 dataset.** Voxel-wise size selectivity is estimated using a Spearman rank correlation coefficient between voxel activation and real-world size of the most salient object (see *Methods* for more details). Values are plotted for each individual subject in BOLD5000. The coefficients for BOLD5000 and NSD demonstrate broad agreement (compare to Figure 2A).

($t_7 = 8.517; p = 0.0002$). However, we find that neither PPA nor OPA have significant size selectivity (PPA: $t_7 = 0.350; p = 0.8185$; OPA: $t_7 = 1.963; p = 0.1130$), a finding that diverges from some past work (Julian et al. (2017), Konkle and Oliva (2012)). For body (EBA, FBA) and face (OFA, FFA) selective regions, we find EBA to be not size selective (EBA: $t_7 = 0.126; p = 0.9030$), while FBA is small selective (FBA: $t_7 = 6.294; p = 0.001$). We identify the OFA and FFA face selective areas to be not size selective (OFA: $t_7 = 2.299; p = 0.0918$; FFA: $t_7 = 2.106; p = 0.1046$). We find OWFA to not be size selective (OWFA: $t_7 = 2.433; p = 0.0901$). In the VWFA, we find significant selectivity for small objects (VWFA: $t_7 = 9.654; p = 0.0001$).

To assess the generality of our results, we also perform the same analyses using an independent dataset consisting of different subjects and images, BOLD5000 (Figure 4). We find the large scale neural organization for real-world size in BOLD5000 to be consistent with what we observe in NSD, with small-selective populations observable on the ventral surface and large-selective populations more medial and lateral to these patches. Together, these results show that a large-scale organization of visual cortex according to size selectivity can be recovered using responses to natural images, and that this finding is robust across datasets. These naturalistic results are largely in agreement with those observed in prior work using isolated object images (comparisons to previous work are elaborated on in *Discussion*).

4.2 Experiment 2: Animacy and size selectivity

Given the observation of real-world object size selectivity across visual cortex, we next assess whether this pattern holds equally well for animate objects and inanimate objects. This analysis is motivated in part by past work (Konkle and Caramazza (2013)) in which size selectivity was found only for inanimate objects, not for animate objects. For the size-selective voxels identified in the previous experiment, we separately estimate their size selectivity (rank correlation coefficient) using only responses to animate or inanimate objects (Figure 5). We find voxel-wise size selectivity to be largely consistent whether estimated using animate or inanimate objects, with the spatial distribution of size-selective voxels having a similar pattern for animate and inanimate objects (Figure 5C). On average, a proportion of 0.8071 (SEM= 0.0175) of the significant voxels across all subjects had the same direction of size selectivity (i.e., same sign of correlation coefficient) for animate and inanimate objects. Computing the cosine similarity of the correlation coefficients computed using animate and inanimate objects, we get an average of 0.5712 (SEM=0.0346). Despite these similarities, we find the average coefficient magnitude to be smaller when estimated using animate objects ($\mathrm{M}_{\mathrm{animate}} = 0.0514$; SEM=0.0020) than inanimate objects ($\mathrm{M}_{\mathrm{inanimate}} = 0.0998$; SEM=0.0015). Reflecting this difference in magnitude, the slope of the relationship between correlation coefficients estimated from animate versus inanimate objects (i.e., red lines in Figure 5B) is 0.3048 in **S1**, and 0.3041 in **S3**. These results indicate that the pattern of size selectivity is highly consistent between animate and inanimate objects, albeit much weaker when viewing animate objects.

4.3 Experiment 3: Low-level features and size selectivity

To further examine the mechanisms of real-world object size selectivity, we explored the extent to which simple visual properties of the images might contribute to neural selectivity for real-world size. Specifically, it may be possible that small and large objects possess different image statistics (an example

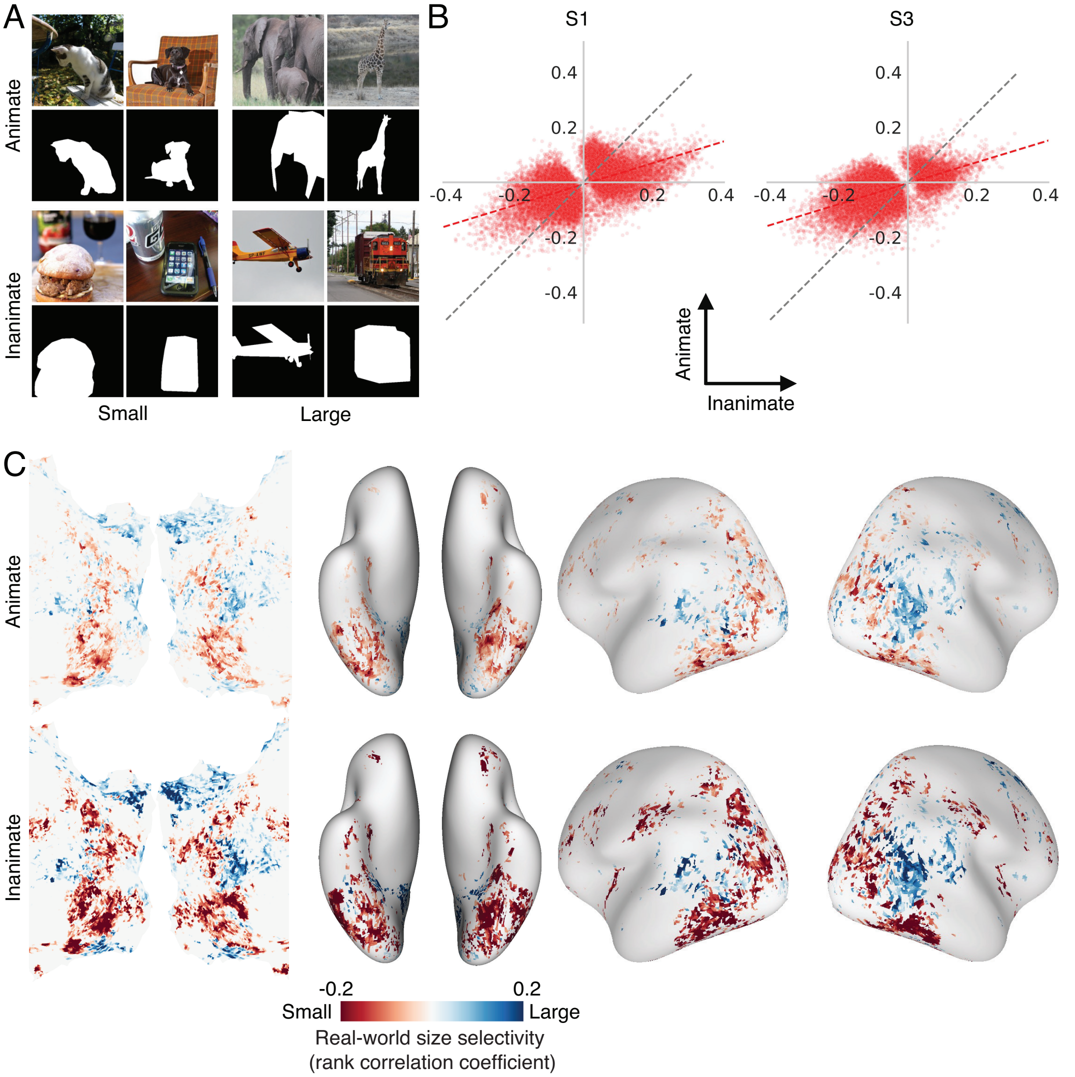

Figure 5: **Comparison of size selectivity measured using animate versus inanimate objects.** **(A)** Examples of images including animate (top) and inanimate (bottom) objects. We visualize both the original RGB stimulus, as well as the mask for the most salient object. **(B)** The size selectivity (Spearman rank correlation) of each voxel when only inanimate objects (x-axis) versus when only animate objects (y-axis) are considered (see *Methods* for details). Grey line indicates where $y = x$, red line indicates best fit. **(C)** Size selectivity estimated over just the animate (top) or just the inanimate (bottom) object images, plotted on the cortical surface. Results reflect the average of eight subjects. Blue indicates selectivity for large objects, red indicates selectivity for small objects; non-significant voxels are set to white for improved visibility. We observe the spatial patterns of size selectivity to be largely consistent between animate and inanimate objects, though the magnitude of the size response is weaker when considering only animate objects.

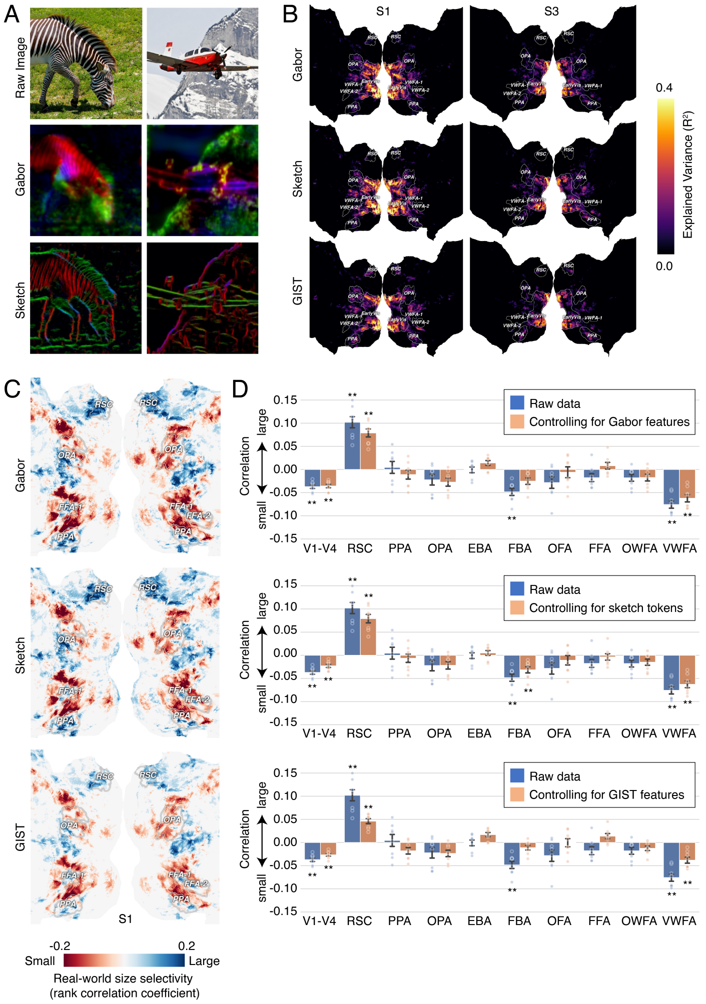

Figure 6: **Controlling for low-level visual features.** To examine the contributions of low-level features to size selectivity, we constructed models using several low- and mid-level feature spaces: Gabor features, sketch tokens features, and GIST features; see *Methods* for more details on these models and feature spaces. **(A)** Top: Two RGB images from the NSD dataset. Middle: PCA projection of the extracted Gabor feature map. Bottom: PCA projection of the extracted sketch token feature map. **(B)** Explained variance on a held-out test set is shown for the models built using each set of features. **(C)** Size selectivity after controlling for each set of low-level features (i.e., computed using the residuals of each low-level model), visualized on a flat map. **(D)** Size selectivity before (blue bars; raw data) and after (orange bars) controlling for each set of features, shown after averaging across all voxels in each ROI and across subjects. The blue bars are the same data as Fig 2C. Bar heights and errorbars indicate mean ±1 SEM across subjects; ** indicate that the average correlation coefficient was significantly different than zero (one-sample t-test; $p < 0.01$; p-values FDR corrected across subjects).

being large objects having more straight lines and rectilinear contours), and that this difference in features is sufficient to explain the difference in response. Supporting this idea, some work has suggested that cortical organization for real-world size can be elicted by mid-level features alone (Long et al. (2018)), and differences between past studies (e.g., Konkle and Oliva (2012) and Julian et al. (2017)) have been speculated to be related to different low-level image statistics. However, an alternative hypothesis is that real-world size selectivity is not driven by low- or mid-level visual features, instead requiring computation of high-level visual features and/or semantic knowledge.

To evaluate these possibilities, we construct voxelwise linear models based on Gabor features, GIST features, and sketch token features in combination with an estimate of each voxel's population receptive field (pRF). These three feature spaces represent local low level, global low level, and mid level features respectively (Figure 6; see *Methods* for details). We first evaluate the amount of variance explained by each feature space on a held out test set (Figure 6B). We find that all three feature descriptors have the highest explained variance within early visual areas. The results are consistent between subjects. The sketch token descriptor has slightly higher explained variance than Gabor descriptors in the higher visual regions. The GIST descriptor provides additional explained variance in RSC that is not explained by Gabor or sketch token features.

To assess how well each of these models can account for each voxel's real-world size selectivity, we re-compute the size selectivity values based on the residuals of each model. As shown in Figure 6C, we observe that the large-scale distribution of object size preferences seen in Figure 2 persists even after controlling for low and mid level features, although the magnitude of size selectivity decreases slightly relative to that computed on the raw data. At the ROI-averaged level (Figure 6D), we find that after controlling for Gabor features, V1-V4, RSC, and VWFA remain size selective (V1-V4: $t_7 = 9.473; p = 0.0002$; RSC: $t_7 = 9.168; p = 0.0002$; VWFA: $t_7 = 7.894; p = 0.0003$); after controlling for sketch token features, V1-V4, RSC, FBA, and VWFA are size selective (V1-V4: $t_7 = 6.530; p = 0.0011$; RSC: $t_7 = 8.690; p = 0.0003$; FBA: $t_7 = 4.758; p = 0.0052$; VWFA: $t_7 = 8.781; p = 0.0003$); after controlling for GIST features, V1-V4, RSC, and VWFA are size selective (V1-V4: $t_7 = 9.618; p = 0.0003$; RSC: $t_7 = 8.634; p = 0.0003$; VWFA: $t_7 = 6.011; p = 0.0018$). For all three control conditions, V1-V4, RSC, and VWFA remain consistently size selective. These results indicate that while low and mid-level features may account for a small portion of the observed size selectivity, a significant portion of the size selectivity cannot be accounted for by these features.

4.4 Experiment 4: Characterizing the form of real-world size tuning profiles

Prior work has explored size selectivity using objects with a discrete size (large or small). However, the size of objects we observe in our daily life can span several magnitudes, and it may be advantageous for neurons that prefer different sizes to utilize different coding schemes – for example, a logarithmic mapping between real-world size and neural activation would lead to highest dynamic range for small objects, while an exponential mapping between real-world size and neural activation would lead to a higher dynamic range for large objects (Figure 7A). To characterize the shape of size tuning functions within visual cortex, we fit functions that predict each voxel's activation as a function of real-world size. For functions taking the form of $f(x) = x^\gamma$, we search through exponential ($\gamma > 1.0$), linear ($\gamma = 1.0$), and logarithmic ($\gamma < 1.0$) like functions. While both logarithmic-like and exponenential-like functions map values to $[0, 1]$, they differ in how they allocate their output range. Logarithmic-like functions allocate more of their output range to small inputs, while exponential-like functions allocate more of their output range to large inputs. We hypothesize that using logarithmic and exponential coding, respectively, may allow voxels to maximize their coding range for small and large sizes.

We evaluate these functions on a held out test-set, and identify the function that maximizes explained variance. Overall, we observe that a logarithmic encoding function leads to the highest predictive accuracy in the majority of visual cortex voxels. Controlling for the size selectivity of the voxel itself, we find that voxels which prefer small objects are overwhelmingly best described by logarithmic coding. Within large-selective voxels, a small majority overall also are better described by logarithmic coding, however the ratio of exponential coding voxels rises notably, with a slight majority of the large-selective voxels in OFA and FFA being better described by an exponential coding function. This is consistent with our prediction of a difference in coding scheme between small-selective and large-selective neural populations.

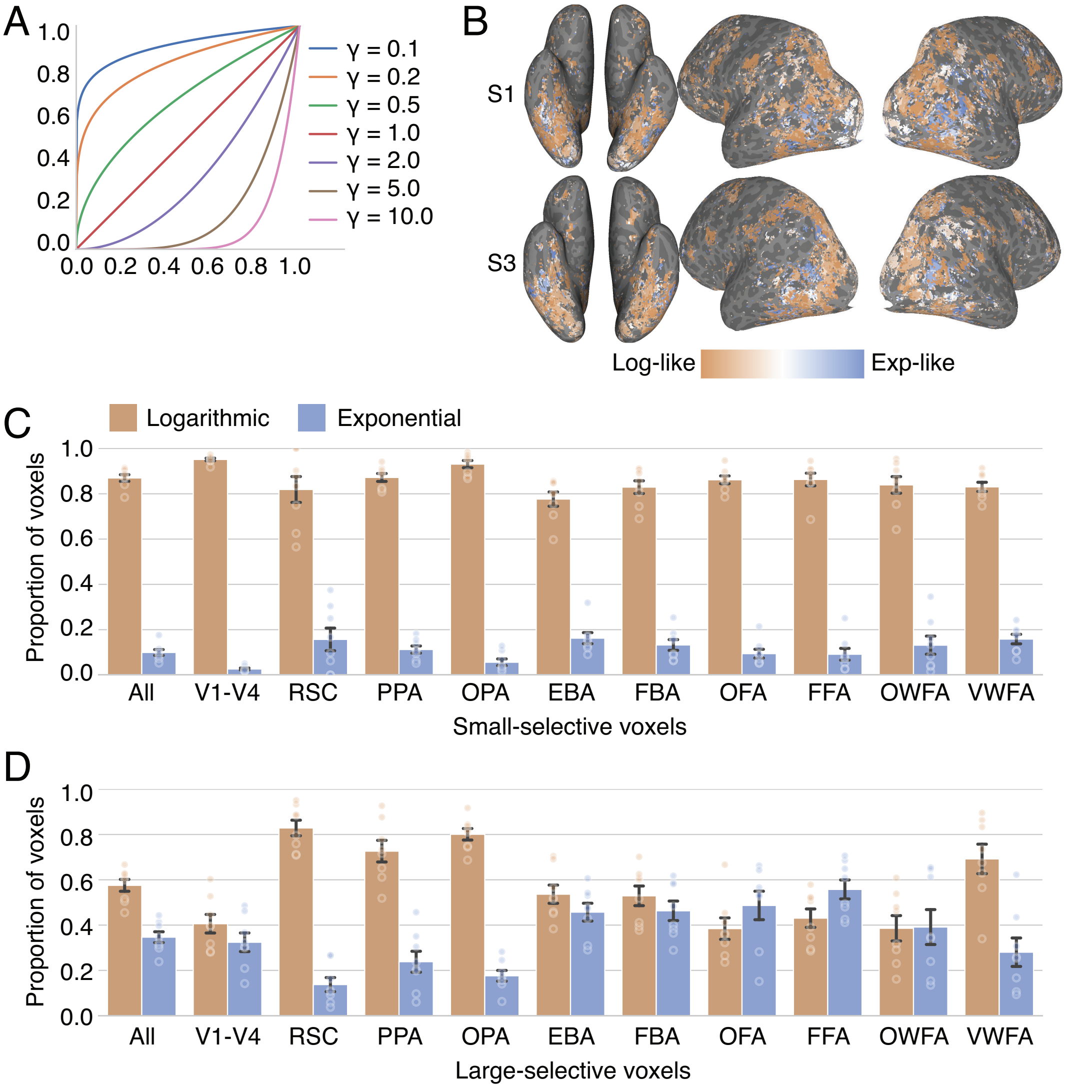

Figure 7: **Quantifying the encoding of real-world size in the brain. (A)** Visualization of the set of functions that we search, which differ in their γ parameter. When $\gamma < 1$ we have functions that are logarithmic-like, when $\gamma = 1$ we have a function that is linear, and when $\gamma > 1$ we have functions that are exponential-like; see *Methods* for more details. **(B)** The best gamma value for each voxel (i.e., the value that yields the highest R^2 on the test set) is plotted on an inflated cortical surface for two example NSD subjects. We find that a logarithmic coding function tends to dominate. **(C)** Proportion of voxels that are best described by a logarithmic function (orange) or an exponential function (blue), out of all the small-selective voxels within each ROI. **(D)** As in **(C)**, but computed for all the large-selective voxels in each ROI. Bar heights and errorbars indicate mean ±1 SEM, circles represent individual subjects.

5 Discussion

As a high-level property of objects in our environment, real-world object size provides us with critical information for interacting with the world. Experientially, humans have an intuitive understanding that real-world size is stable across viewing conditions. That is, the perceived real-world size of an object remains unchanged as it moves farther or closer from us, or undergoes changes in lighting or in orientation. Given the importance of real-world size to a diverse set of tasks, our work investigated:

- How is information about the real-world size of objects embedded within natural scenes represented in the brain?
- Is neural selectivity for real-world size exhibited for both inanimate objects and animate objects?
- Can real-world object size selectivity be explained, in part, by low- and mid-level visual features?
- How is real-world size quantitatively encoded?

With these questions in mind, our work makes several contributions. First, in contrast to prior studies of real-world size, our study employs a diverse dataset of natural images, where each object is presented in contextually-relevant and diverse backgrounds. Our present results are largely (sic) consistent with past results on the coarse-scale organization of size-selective neural populations that relied on single, isolated objects as stimuli (Huang et al., 2022, Julian et al., 2017, Konkle and Caramazza, 2013, Konkle and Oliva, 2012). Second, our more ecologically-grounded approach reveals that size selectivity is present for both animate and inanimate objects, which contrasts with prior studies that suggested that size selectivity was exclusive to inanimate objects (Konkle and Caramazza, 2013). A possible explanation is the higher sensitivity of our study – in terms of both fMRI methods and the range, number and content of stimuli – for detecting somewhat weaker effects. Third, we include a variety of controls for the possibility that size-related neural responses might originate from a distributional difference for low- or mid-level visual features across large and small objects (Julian et al., 2017, Long et al., 2016). By regressing out image visual features on a per-voxel basis using each voxel's population receptive field (pRF), we establish that the size-selective responses we measure cannot be entirely explained by differences in low- and mid-level image statistics. Fourth, we quantitatively characterize the shape of real-world size-related response profiles in visual cortex. Previous behavioral studies in humans suggest that size representations follow Weber-Fechner-like scaling (Konkle and Oliva, 2011) which follows a logarithmic function. We find that while a logarithmic encoding captures the encoding properties of most size-selective voxels, there exists a difference in representational characteristics of small-selective and large-selective voxels.

Our findings regarding size organization in the brain using complex natural images align well with the results from prior fMRI experiments. In particular, we observed a medial-lateral gradient and alternating selectivity for large-small selectivity along the ventral and lateral occipitotemporal cortex respectively – a pattern consistent with previous work (Julian et al., 2017, Konkle and Caramazza, 2013). This correspondence is somewhat surprising given the importance of context in real-world size judgments. For example, in the Ebbinghaus illusion, an object appears relatively larger when surrounded by small objects, but relatively smaller when surrounded by large objects (Doherty et al., 2010). Similar observations have been made using the Jastrow illusion, Delboeuf illusion, and vertical–horizontal illusion; where real-world size perception is affected by contextual cues. Interestingly, our results suggest a similar coding of real-world object size across different contexts; as such, it is possible that context comes into play when the size of an object is ambiguous or unknown. However, for most natural images, real-world size estimates are straightforward, explaining why we did not find many differences between isolated and natural images.

We found evidence for significant size selectivity in several functionally-defined ROIs. Some of these ROI-level results are aligned with prior work that used decontextualized objects, while other aspects of our results differ from past work. First, we found that scene-selective areas PPA and OPA each have slightly more small-selective than large-selective voxels. This result is different from previous reports by Julian et al. (2017), Konkle and Caramazza (2013), who found all three scene regions show a preference for large objects. Second, we found the word-selective area VWFA to be strongly small-selective, while the word-selective area OWFA showed no significant selectivity for size. These preferences are consistent with the functional roles of these regions in the processing of visual words (McCandliss et al., 2003). Third, across all ROIs, VWFA was observed to have the largest proportion of small-selective voxels, while RSC was observed to have the largest proportion of large-selective voxels.

Such selectivity likely reflects the natural statistics of these categories, where words typically manifest as a smaller physical entities (McCandliss et al., 2003), whereas scenes are typically composed of larger objects and navigable environments (Epstein and Baker, 2019).

In contrast to prior work, we observed neural selectivity for real-world size for both animate and inanimate object categories (Huang et al., 2022, Julian et al., 2017, Konkle and Caramazza, 2013, Konkle and Oliva, 2012). Moreover, when comparing the qualitative neural patterns associated with animate and inanimate objects, we see a high degree of agreement in real-world size selectivity, but with the magnitude of the effect for animate objects being weaker (which may be one reason why prior studies did not detect the effect). Methodological advances may be contributing factors: we leveraged data that was collected using a higher resolution scanner (7T), we used recently introduced BOLD processing algorithms that boost SNR (Prince et al. (2022)), and we used continuous size labels for each category. Such improvements may help to detect signals that were previously too weak or noisy to observe. Even given these advances, size selectivity for animate objects was weaker than that for inanimate objects. One reason for this may be that animate objects have higher intra-category size variability than inanimate objects. Many animate objects (e.g., animals) undergo size changes over their lifespan, and thus our weaker animate size selectivity effect may reflect a weaker ecological basis for size in animate object recognition. Future work should test this hypothesis using finer-grained intra-class size labels to create a diverse collection of animate objects over their development.

Our analyses also investigated the contributions of low-level visual features to real-world object size representations. In particular, there is some evidence that real-world size information may correlate with underlying mid-level image statistics, with large and small objects having different amounts of boxiness or curvature (Long et al. (2016, 2018)). Other low-level features also co-vary with real-world object size in natural images (Henderson et al. (2022)). Controlling for low-level and mid-level image statistics using local (Gabor, sketch tokens) and global (GIST) features did reduce the magnitude of size selectivity across visual cortex, and controlling for either Gabor or GIST features led to a loss of significant size selectivity in FBA. Importantly, however, real-world size selectivity remained significant after these manipulations in early visual areas (V1-V4), RSC, and VWFA. This suggests that real-world size processing in the brain includes integration of high-level information that cannot be accounted for by simple visual properties.

Finally, our results provide new insight into the representational mapping between real-world object size and neural activation in visual cortex. In both human and non-human animals, past behavioral work has found that the perception of real-world object size follows a logarithmic scaling function in accordance with Weber–Fechner's Law (Kelley and Kelley (2014), Konkle and Oliva (2011)). Our results reveal a potential neural basis for this psychophysical effect, as we find that a logarithmic coding function provides a good fit for the responses of almost all size-selective ROIs. However, on a per-voxel basis, voxels that are large-selective tend to have a different coding function than voxels that are small-selective, with many large-selective voxels being well-described by an exponential size coding function. This pattern may be understood by considering how the dynamic range of neural responses is allocated to large and small objects within each of these coding schemes. A logarithmic encoding of object size, as suggested by past work, would lead to decreased sensitivity to changes of size as size increases. By contrast, an exponential encoding of object size allows for more sensitivity for larger sizes. Future work should address how these differences in encoding across large- and small-selective neural populations are related to previously-reported psychophysical effects.

In conclusion, we find that neural selectivity for real-world object size is robust for objects embedded in natural scenes. Moreover, the ecological relevance of real-world size is reflected in our finding of size selectivity for both inanimate and animate objects, as well as in the fact that selectivity for size across different brain regions is consistent with each region's ascribed functional role. We also establish that the observed selectivity is driven, as least in part, by high-level properties, although low-level features may also contribute. Reinforcing the ecological import of real-world size selectivity in the brain, we find that large and small selective voxels represent size according to different coding functions, thereby enhancing their sensitivity to particular size ranges in an ecologically adaptive manner. *In toto*, the robustness of size selectivity that we demonstrate using complex, natural scenes advances our understanding of the neural basis of real-world object size perception, a key aspect of our visual experience and interactions with the physical world.

References

E. J. Allen, G. St-Yves, Y. Wu, J. L. Breedlove, J. S. Prince, L. T. Dowdle, M. Nau, B. Caron, F. Pestilli, I. Charest, et al. A massive 7t fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, 2022.

M. Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629, 2004.

N. C. Benson, K. W. Jamison, M. J. Arcaro, A. T. Vu, M. F. Glasser, T. S. Coalson, D. C. V. Essen, E. Yacoub, K. Ugurbil, J. Winawer, and K. Kay. The human connectome project 7 tesla retinotopy dataset: Description and population receptive field analysis. *Journal of Vision*, 18:23–23, 12 2018.

N. Chang, J. A. Pyles, A. Marcus, A. Gupta, M. J. Tarr, and E. M. Aminoff. Bold5000, a public fMRI dataset while viewing 5000 visual images. *Scientific Data*, 6(1):1–18, 2019.

L. Cohen, S. Dehaene, L. Naccache, S. Lehéricy, G. Dehaene-Lambertz, M.-A. Hénaff, and F. Michel. The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain*, 123(2):291–307, 2000.

d. dimensions. Database of dimensioned drawings, Jan 2022. URL https://www.dimensions.com/.

M. J. Doherty, N. M. Campbell, H. Tsuji, and W. A. Phillips. The ebbinghaus illusion deceives adults but not young children. *Developmental Science*, 13(5):714–721, 2010.

R. Epstein and N. Kanwisher. A cortical representation of the local visual environment. *Nature*, 392 (6676):598–601, 1998.

R. A. Epstein and C. I. Baker. Scene perception in the human brain. *Annual review of vision science*, 5:373–397, 9 2019.

J. P. Gallivan, C. Cavina-Pratesi, and J. C. Culham. Is that within reach? fmri reveals that the human superior parieto-occipital cortex encodes objects reachable by the hand. *Journal of Neuroscience*, 29 (14):4381–4391, 2009.

J. S. Gao, A. G. Huth, M. D. Lescroart, and J. L. Gallant. Pycortex: an interactive surface visualizer for fMRI. *Frontiers in Neuroinformatics*, 9, 2015.

J. Grèzes, M. Tucker, J. Armony, R. Ellis, and R. E. Passingham. Objects automatically potentiate action: an fMRI study of implicit processing. *European Journal of Neuroscience*, 17(12):2735–2740, 2003.

M. Henderson, M. J. Tarr, and L. Wehbe. Low-level tuning biases in higher visual cortex reflect the semantic informativeness of visual features. *bioRxiv*, 2022.

T. Huang, Y. Song, and J. Liu. Real-world size of objects serves as an axis of object space. *Commun. Biol.*, 5(1):1–12, 2022.

N. Jain, A. Wang, M. M. Henderson, R. Lin, J. S. Prince, M. J. Tarr, and L. Wehbe. Selectivity for food in human ventral visual cortex. *Commun. Biol.*, 6(1):175, 2023.

M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1072–1080, 2015.

J. B. Julian, J. Ryan, and R. A. Epstein. Coding of object size and object category in human visual cortex. *Cerebral Cortex*, 27(6):3095–3109, 2017.

N. Kanwisher, J. McDermott, and M. M. Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11):4302–4311, 1997.

L. A. Kelley and J. L. Kelley. Animal visual illusion and confusion: the importance of a perceptual perspective. *Behavioral Ecology*, 25(3):450–463, 2014.

T. Konkle and A. Caramazza. Tripartite organization of the ventral stream by animacy and object size. *Journal of Neuroscience*, 33(25):10235–10242, 2013.

T. Konkle and A. Oliva. Canonical visual size for real-world objects. *Journal of Experimental Psychology: Human Perception and Performance*, 37(1):23, 2011.

T. Konkle and A. Oliva. A real-world size organization of object responses in occipitotemporal cortex. *Neuron*, 74(6):1114–1124, 2012.

J. J. Lim, C. L. Zitnick, and P. Dollár. Sketch tokens: A learned mid-level representation for contour and object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3158–3165, 2013.

B. Long and T. Konkle. A familiar-size stroop effect in the absence of basic-level recognition. *Cognition*, 168:234–242, 2017.

B. Long, T. Konkle, M. A. Cohen, and G. A. Alvarez. Mid-level perceptual features distinguish objects of different real-world sizes. *Journal of Experimental Psychology: General*, 145(1):95, 2016.

B. Long, C.-P. Yu, and T. Konkle. Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, 115(38): E9015–E9024, 2018.

B. D. McCandliss, L. Cohen, and S. Dehaene. The visual word form area: expertise for reading in the fusiform gyrus. *Trends Cogn Sci*, 7(7):293–299, 2003.

S. O. Murray, H. Boyaci, and D. Kersten. The representation of perceived angular size in human primary visual cortex. *Nature Neuroscience*, 9(3):429–434, 2006.

S. Nasr and R. B. Tootell. A cardinal orientation bias in scene-selective visual cortex. *Journal of Neuroscience*, 32(43):14921–14926, 2012.

S. Nasr, C. E. Echavarria, and R. B. Tootell. Thinking outside the box: rectilinear shapes selectively activate scene-selective cortex. *Journal of Neuroscience*, 34(20):6721–6735, 2014.

A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

J. S. Prince, I. Charest, J. W. Kurzawski, J. A. Pyles, M. J. Tarr, and K. N. Kay. Improving the accuracy of single-trial fMRI response estimates using glmsingle. *eLife*, 11:e77599, nov 2022.

N. Reddy, S. Jain, P. Yarlagadda, and V. Gandhi. Tidying deep saliency prediction architectures. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10241–10247. IEEE, 2020.

J. Sergent, S. Ohta, and B. MacDonald. Functional neuroanatomy of face and object processing: A positron emission tomography study. *Brain*, 115:15–36, 1992.

A. Slater, A. Mattock, and E. Brown. Size constancy at birth: Newborn infants' responses to retinal and real size. *Journal of Experimental Child Psychology*, 49(2):314–322, 1990.

G. St-Yves and T. Naselaris. The feature-weighted receptive field: An interpretable encoding model for complex feature spaces. *NeuroImage*, 6 2017.

M. Tucker and R. Ellis. The potentiation of grasp types during visual object categorization. *Visual Cognition*, 8(6):769–800, 2001.

E. K. Warrington and T. Shallice. Category specific semantic impairments. *Brain*, 107(3):829–853, 1984.