

**Supplementary Materials for:**  
**Biased competition in semantic representation during natural  
visual search**

Mohammad Shahdloo<sup>1,3</sup>, Emin Çelik<sup>2,3</sup>, Tolga Çukur<sup>1,2,3</sup>

1. Department of Electrical and Electronics Engineering, Bilkent University, Ankara/Turkey
2. Neuroscience Program, Sabuncu Brain Research Center, Bilkent University, Ankara/Turkey
3. National Magnetic Resonance Research Center (UMRAM), Bilkent University, Ankara/Turkey

Correspondence to:

Tolga Çukur

Department of Electrical and Electronics Engineering, Room 304

Bilkent University

Ankara, TR-06800, Turkey

TEL: +90 (312) 290-1164

*cukur@ee.bilkent.edu.tr*

Mohammad Shahdloo

National Magnetic Resonance Research Center (UMRAM)

Bilkent University

Ankara, TR-06800, Turkey

TEL: +90 (312) 290-8420

*shahdloo@ee.bilkent.edu.tr*

# 1 Supplementary Methods

**Definition of functional areas.** Functional regions of interest (ROIs) were identified in individual subjects using functional localizers (5). Localizer experiments for category-selective areas (fusiform face area, FFA; extrastriate body area, EBA; parahippocampal place area, PPA; retrosplenial cortex, RSC; lateral occipital complex, LOC) were performed in six 4.5 minute runs of 16 blocks (5). Subjects passively viewed 20 random static images from one of the objects, scenes, body parts, faces, or spatially scrambled objects groups in each block. Each image was shown for 300 ms following a 500 ms blank period. Scene-selective ROIs (PPA, RSC) were identified as voxels with positive scene versus objects contrast ( $t$ -test,  $p < 10^{-4}$ , uncorrected). FFA, EBA, and LOC were defined using face-versus-object, body-part-versus-object, and object-versus-scrambled-object contrasts, respectively ( $t$ -test,  $p < 10^{-4}$ , uncorrected). Localizer experiment for attentional-control areas (intraparietal sulcus, IPS; frontal eye fields, FEF; supplementary eye field, SEF; frontal operculum, FO) contained one 10 minute run. In each 20 second block, either a self-generated saccade task or a resting task was prescribed (2, 11). The run contained 30 blocks. Attentional-control areas were localized using saccade-versus-rest contrast ( $t$ -test,  $p < 10^{-4}$ , uncorrected). Localizer experiment for early visual areas (RET: V1, V2, V3) contained four 9 minute runs. Subjects viewed clockwise and counterclockwise rotating polar wedges in two runs. In the remaining two runs, subjects viewed expanding and contracting rings. Visual angle and eccentricity maps were used to define visual areas V1-3. Localizer experiment for area MT+ consisted of four 90 second runs, containing alternating 16 second blocks of continuous and temporally scrambled natural movies. Area MT+ was localized using continuous-versus-scrambled movie contrast ( $t$ -test,  $p < 10^{-4}$ , uncorrected). ROIs were refined to voxels with contrast level more than half of the maximum near a 2 mm neighborhood of the cortical surface.

**Eye-movement controls.** Although the subjects participating in the experiment were highly trained psychophysical observers who had extensive experience in fixation tasks, residual eye-movements can be a confounding factor. Several control analyses were performed to ensure that eye-movements did not unduly bias the results. During the scans, subjects’ eye positions were monitored at 60 Hz using a custom-built camera system equipped with an infrared source (Avotec) and the ViewPoint EyeTracker software suite (Arrington Research). The eye tracker was calibrated before each run of data acquisition. Kruskal-Wallis tests were used to detect systematic differences in the distribution of eye position and movement. First, distribution of eye position was examined for “attend to humans”, “attend to vehicles”, and “attend to both humans and vehicles” tasks. We find that the distribution of eye position is not affected by attention condition ( $p = 0.22$ ), or by target presence or absence ( $p = 0.09$ ), and no significant interaction is present between these two factors ( $p = 0.62$ ). To test whether eye movement is affected by target or distractor detection, we studied the distribution of eye position during a 1 sec window around target onset and target offset. The eye position distributions are not affected by target onset ( $p = 0.18$ ) or offset ( $p = 0.51$ ), and there are no significant interactions between the aforementioned factors ( $p = 0.45$ ). To determine systematic differences in rapid moment-to-moment variations in eye position across the three tasks, we examined the moving-average standard deviation of eye position in a 200 ms window. There are no significant effects of attention condition ( $p = 0.27$ ), target presence or absence ( $p = 0.69$ ), target onset ( $p = 0.36$ ), or target offset ( $p = 0.08$ ), and there are no significant interactions between these factors ( $p = 0.36$ ).

To systematically examine the target locations, we used a state-of-the-art object detection toolbox based on deep neural networks, YOLO (9). YOLO was employed to output presence labels and bounding boxes for human and vehicle objects in each stimulus frame. The target location was taken as the center of the bounding box. To account for multiple targets in a given frame, we computed a weighted average of target locations, where the weight for each target was taken as the area of its bounding box. During the divided attention task,

the target location was taken as the average of human and vehicle locations for frames that contained both categories. We find that targets appear at significantly different locations across the three conditions ( $p < 10^{-6}$ , Kruskal-Wallis test). On average, the target is located at  $7.34 \pm 0.13$  (mean $\pm$ sem across frames) degrees away from the fixation spot during search for humans, at  $5.93 \pm 0.15$  degrees during search for vehicles, and at  $6.50 \pm 0.16$  degrees during the divided attention task. In contrast, across frames where the target is present, the average eye position is  $1.52 \pm 0.77$  degrees (mean $\pm$ sem across subjects) during search for humans,  $1.38 \pm 0.64$  degrees during search for vehicles, and  $1.32 \pm 0.60$  degrees during divided attention. Thus, although the target locations vary substantially across attention conditions, distribution of eye positions is similar regardless of condition. Taken together, these controls suggest that the main findings of the current study are not confounded by eye movement biases.

**Voxelwise category model.** For each voxel, response profiles were determined by fitting category models that represented hundreds of objects and actions in natural movies. As a first step, each 1-second clip of the movie stimulus was manually labeled for presence of hundreds of distinct object and action categories (5). Presence of superordinate categories was inferred from the terms in the WordNet lexicon, a lexical database that groups words based on their semantic relationships (6). This procedure yielded time courses for 831 model features (i.e. categories). Each time course was then downsampled to 0.5 Hz to match the acquisition rate of fMRI. Separate finite impulse response (FIR) filters were used for each model feature to capture the hemodynamic response. Filter delays were set to 4, 6, and 8 s. This is equivalent to concatenating feature vectors that are delayed by two, three, and four samples.

To prevent head-motion and physiological noise confounds, estimates of these nuisance factors were regressed out of the BOLD responses. Six affine motion time courses estimated during

the motion-correction stage were taken as the head-motion regressors. Two regressors to capture respiration and nine regressors to capture cardiac activity were estimated using the data collected via a pulse oximeter and a pneumatic belt during the main experimental runs (12). To prevent eye movements confounds, moving-average standard deviation of eye position was included in the model as a nuisance regressor and was regressed out of the BOLD responses.

To reduce spurious correlations between model features and global motion-energy of the movie stimulus, a nuisance regressor was included that reflected the total motion-energy. The motion-energy time course was formed by taking the mean motion-energy in each one second movie clip. Movie frames were transformed into the International Commission on Illumination LAB color space, and the luminance channel was extracted. The luminance was then passed through the motion-energy filter bank. The motion-energy filter bank contained 2139 Gabor filters. Filters were computed at eight directions (0 to 315°, in 45° steps), three temporal frequencies (0, 2, and 4 Hz) and six spatial frequencies (0, 1.5, 3, 6, 12, and 24 cycles/image). Filters were placed on a square grid spanning the 24° × 24° field of view. Finally, the motion-energy time course was assessed by squaring and summing outputs of quadrature filter pairs, and the results were passed through a logarithm compressive nonlinearity and temporally downsampled to match the fMRI acquisition rate (8).

To account for potential correlations between target detection and BOLD responses, a target-presence regressor was included in the model. The target-presence regressor contained category regressor for “person” during “attend to humans” task and the category regressor for “conveyance” during “attend to vehicles” task. The target-presence regressor during divided attention task contained the binary union of the “person” and “conveyance” category regressors. The described regressors were aggregated and used as the stimulus matrix.

**Model fitting and testing.** Voxelwise models were fit using regularized linear regression with an  $\ell_2$  penalty to avoid overfitting (7). To prevent target detection bias, model fitting for the three attention tasks was performed concurrently. To do this, the stimulus and BOLD response matrices were aggregated across tasks (Fig. 1). Note that this procedure ensures that the same regularization parameter will be used in each voxel across the three tasks. Furthermore, if the conditions were modeled individually, then the target regressor would be redundant with the human or vehicles category regressors in the single-target tasks. In contrast, using the aggregated stimulus matrix enables employing the target regressor.

A nested cross-validation (CV) procedure was used to estimate response profiles for each voxel. Data were segmented into 58 24-second blocks. In each of the 20 outer folds, 6 blocks were randomly held-out as validation data and the remaining blocks were used for parameter optimization and fitting models on the inner folds. In each of the 20 inner folds, blocks were randomly shuffled and split into 40 blocks as training data and 12 blocks as test data. Models were fit on the training data for regularization parameters in the range  $[2^{-3}, 2^{20}]$ . Using the models found for each regularization parameter, responses were predicted for the test data. Prediction scores were separately calculated for each voxel, taken as the Pearson’s correlation between actual and predicted responses. Prediction scores were then averaged across the inner CV folds. Regularization parameters maximizing the average prediction score were selected in each voxel. Nuisance regressors were discarded from further analyses. Afterwards, optimized parameters were used to fit models on the union of training and test data in each outer fold, yielding category response profiles. To assess model performance, responses were predicted for the validation data using the fit models and prediction scores of each voxel were averaged across the attention tasks. Finally, response profiles and prediction scores for each voxel were averaged across the outer folds. Model fitting was performed using custom-written software in Matlab (MathWorks MA).

**Alternative cross validation procedure.** In the original cross-validation (CV) procedure, each individual movie frame was present in either of the training, validation or test sets without overlap. Yet, splitting the data into fixed-length blocks can cause a single movie clip to be split into separate blocks. For movie clips with gradually changing semantic content, similar visual scenes might then appear in multiple sets. If the proportion of similar scenes that overlap across training and test sets is large, this in turn might cause an upward bias in model performance.

To rule out this potential bias, we performed a control analysis by modifying the CV procedure. Specifically, a variable block size was prescribed to ensure that each movie clip appeared within only a single block. This resulted in 46 blocks with durations in the range [20, 38] seconds, which contained either one or two clips. In each CV fold, we used 31 blocks for training, 10 blocks for test, and 5 blocks for validation. The relative proportions of the three sets were selected to closely match those in the original CV procedure. Category models were then fit in individual subjects. Prediction scores of the models based on the original and modified CV procedures are shown in Fig. 2 below. The prediction scores assessed via original and modified procedures have highly matching distributions. Correlation coefficient between voxel-wise prediction scores is  $0.92 \pm 0.01$  (mean $\pm$ std. across subjects,  $p < 10^{-4}$ ). This control analysis suggests that assessment of model performance is not confounded by similarity of visual scenes among training, test, and validation sets.

**Significance tests.** A 20-fold CV procedure was implemented to assess significance of the fit models. In each fold, the corresponding fit models and validation data from the model fitting stage were used. Responses were predicted for the validation data using the fit models and then predicted responses were used to calculate the prediction scores. In each fold, predicted BOLD response blocks were resampled 500 times with replacement. This procedure resulted in  $10^4$  bootstrap samples. Prediction scores were averaged across CV folds to assess model performance. A null hypothesis was considered where attention does not

alter response profiles of cortical voxels across tasks. Under this null hypothesis, response profiles are to remain the same across the three attention tasks (3). To assess the prediction score distribution under the null hypothesis, 24-sec blocks of predicted BOLD responses were randomly shuffled across the attention tasks and prediction score was then calculated. The  $p$ -value was taken as the fraction of the bootstrap samples for which the average prediction score across CV folds was lower than that under the null hypothesis. Finally, statistical significance levels were corrected for multiple comparisons using false discovery rate (FDR) control (1).

Bootstrap test was also used to assess significance of BI and LI assessments at different levels. First, we used bootstrap test to assess significance of BI and LI averages across subjects in 11 individual ROIs. Second, we used bootstrap test to compare significant differences between average LI in category-selective areas and average LI in attentional-control areas, in RET, in MT, and in LOC. Finally, we used bootstrap test to assess significance of individual subject BIs. To do this, we compared individual subject BIs against BIs assessed via projection of category responses onto subjects' null semantic spaces. To create null semantic spaces, the order of PCs was resampled  $10^4$  times with replacement independently in individual subjects. Semantic tuning distributions were then assessed by projecting the masked response vectors across voxels within each ROI onto individual subjects' null semantic spaces. Note that null semantic spaces reflect the hypothesis that obtained BIs are not significantly different than zero.

**Visualization on cortical surfaces.** Cortical flatmaps were generated by projecting voxelwise results on individual subject flattened cortical surfaces. Cortical surfaces were constructed for each subject using T1-weighted anatomical brain scans. Freesurfer software was used to construct surfaces (10). Surfaces were then flattened using Pycortex (4).



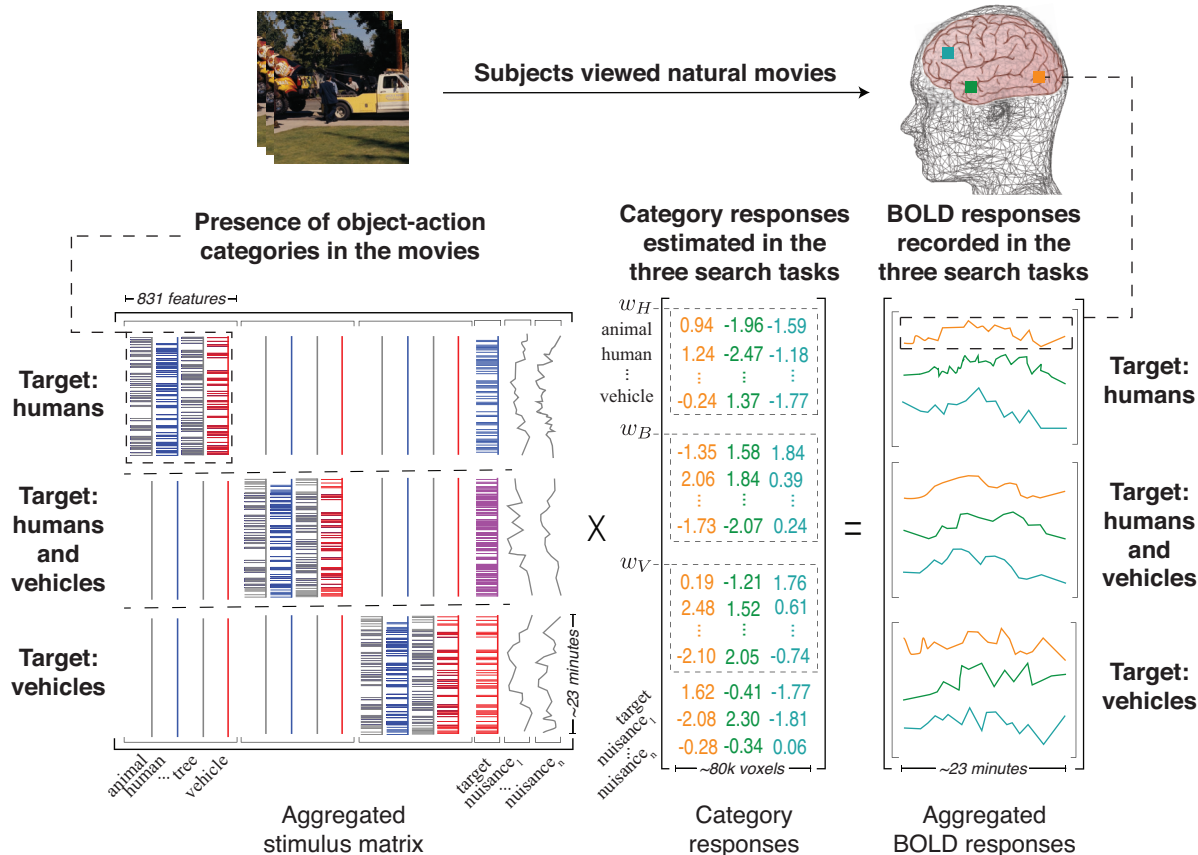
## References

1. Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, August 2001.
2. Maurizio Corbetta, Erbil Akbudak, Thomas E Conturo, Abraham Z Snyder, John M Ollinger, Heather A Drury, Martin R Linenweber, Steven E Petersen, Marcus E Raichle, David C Van Essen, and Gordon L Shulman. A common network of functional areas for attention and eye movements. *Neuron*, 21(4):761–773, October 1998.
3. Tolga Çukur, Shinji Nishimoto, Alexander G Huth, and Jack L Gallant. Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience*, 16(6):763–770, April 2013.
4. James S Gao, Alexander G Huth, Mark D Lescroart, and Jack L Gallant. Pycortex: an interactive surface visualizer for fMRI. *Frontiers in Neuroinformatics*, 9(22):162, September 2015.
5. Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. *Neuron*, 76(6):1210–1224, December 2012.
6. George A Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, November 1995.
7. Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410, May 2011.
8. Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. *Current Biology*, 21(19):1641–1646, October 2011.

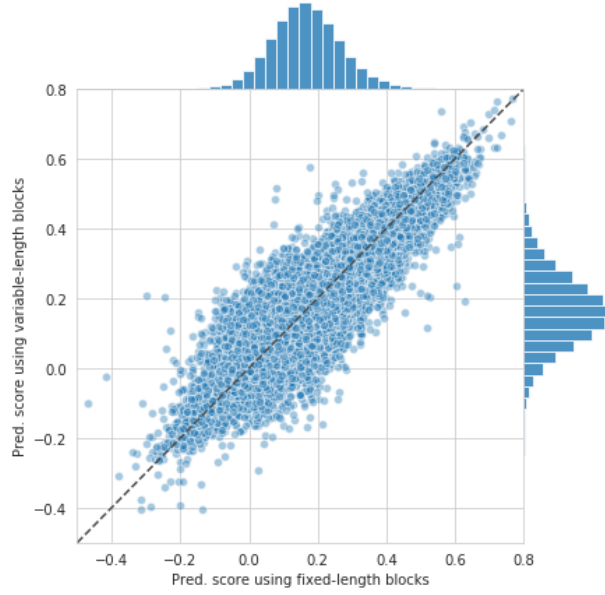
9. Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement. *arXiv.org*, April 2018.
10. Martin Reuter, Nicholas J Schmansky, H Diana Rosas, and Bruce Fischl. Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*, 61(4): 1402–1418, July 2012.
11. John T Serences. Control of Object-based Attention in Human Cortex. *Cerebral Cortex*, 14(12):1346–1357, May 2004.
12. Timothy D Verstynen and Vibhas Deshpande. Using pulse oximetry to account for high and low frequency physiological artifacts in the BOLD signal. *NeuroImage*, 55(4): 1633–1644, April 2011.

# List of Figures

- 1    **Experimental and modeling procedures.** Subjects viewed 69 minutes of natural movies. BOLD responses were recorded using functional MRI (fMRI) while subjects performed covert search for “humans”, “vehicles” (i.e. single-target attention tasks), or “both humans and vehicles” (i.e. divided attention task) in the movies. Movies were labeled for presence of object and action categories. Presence of superordinate categories was inferred using the WordNet lexicon that resulted in a total of 831 object and action categories. Models for the three attention tasks were fit simultaneously using the aggregated stimulus and BOLD response matrices ( $w_H$ ,  $w_V$ , and  $w_B$  for the attend to “humans”, attend to “vehicles”, and attend to “both humans and vehicles” tasks, respectively). A target-presence regressor was used to account for BOLD response modulations resulting from detection of targets in the scenes. Target-presence regressor was comprised of the human regressor (red series), vehicle regressor (blue series), and the binary union of the two (cyan series) to indicate the presence of “humans”, “vehicles”, and “both humans and vehicles”, depending on the task. Nuisance regressors were used to account for head motion, physiological noise, and eye movements. Category models were fit independently for each voxel using regularized linear regression. The fit models define a response profile for the contribution of 831 object-action categories to BOLD responses. . . . . 11
  
- 2    **Prediction scores of models fit based on the original and modified CV procedures.** Prediction scores (Pearson’s correlation coefficient) of category models fit via the original CV procedure (fixed data block size), and the control CV procedure (variable data block size) were measured for single cortical voxels. Cortical voxels are pooled across subjects, and each voxel is represented with a dot. The horizontal and vertical histograms display prediction score distributions for the original and control CV procedures, respectively. The prediction scores assessed via the original and control procedures have highly matching distributions. Correlation coefficient between voxel-wise prediction scores is  $0.92 \pm 0.01$  (mean $\pm$ std. across subjects,  $p < 10^{-4}$ ). . . . . 12



Supporting Fig. 1: **Experimental and modeling procedures.** Subjects viewed 69 minutes of natural movies. BOLD responses were recorded using functional MRI (fMRI) while subjects performed covert search for “humans”, “vehicles” (i.e. single-target attention tasks), or “both humans and vehicles” (i.e. divided attention task) in the movies. Movies were labeled for presence of object and action categories. Presence of superordinate categories was inferred using the WordNet lexicon that resulted in a total of 831 object and action categories. Models for the three attention tasks were fit simultaneously using the aggregated stimulus and BOLD response matrices ( $w_H$ ,  $w_V$ , and  $w_B$  for the attend to “humans”, attend to “vehicles”, and attend to “both humans and vehicles” tasks, respectively). A target-presence regressor was used to account for BOLD response modulations resulting from detection of targets in the scenes. Target-presence regressor was comprised of the human regressor (red series), vehicle regressor (blue series), and the binary union of the two (cyan series) to indicate the presence of “humans”, “vehicles”, and “both humans and vehicles”, depending on the task. Nuisance regressors were used to account for head motion, physiological noise, and eye movements. Category models were fit independently for each voxel using regularized linear regression. The fit models define a response profile for the contribution of 831 object-action categories to BOLD responses.



Supporting Fig. 2: **Prediction scores of models fit based on the original and modified CV procedures.** Prediction scores (Pearson’s correlation coefficient) of category models fit via the original CV procedure (fixed data block size), and the control CV procedure (variable data block size) were measured for single cortical voxels. Cortical voxels are pooled across subjects, and each voxel is represented with a dot. The horizontal and vertical histograms display prediction score distributions for the original and control CV procedures, respectively. The prediction scores assessed via the original and control procedures have highly matching distributions. Correlation coefficient between voxel-wise prediction scores is  $0.92 \pm 0.01$  (mean $\pm$ std. across subjects,  $p < 10^{-4}$ ).