JNeurosci Accepted Manuscript

JNeurosci
THE JOURNAL OF NEUROSCIENCE

Research Articles: Behavioral/Cognitive

Task-Dependent Warping of Semantic Representations During Search for Visual Action Categories

https://doi.org/10.1523/JNEUROSCI.1372-21.2022

Cite as: J. Neurosci 2022; 10.1523/JNEUROSCI.1372-21.2022

Received: 1 July 2021
Revised: 29 June 2022
Accepted: 6 July 2022

This Early Release article has been peer-reviewed and accepted, but has not been through the composition and copyediting processes. The final version may differ slightly in style or formatting and will contain links to any extended data.

Alerts: Sign up at www.jneurosci.org/alerts to receive customized email alerts when the fully formatted version of this article is published.

Copyright © 2022 the authors

JNeurosci Accepted Manuscript

Task-Dependent Warping of Semantic Representations During Search for Visual Action Categories

Mo Shahdloo[1,2,3], Emin Çelik[2,5], Burcu A. Ürgen[2,4,5], Jack L. Gallant[6], Tolga Çukur[2,3,5,6*]

1. Wellcome Centre for Integrative Neuroimaging, Department of Experimental Psychology, University of Oxford, Oxford/UK
2. National Magnetic Resonance Research Centre (UMRAM), Bilkent University, Ankara/Turkey
3. Department of Electrical and Electronics Engineering, Bilkent University, Ankara/Turkey
4. Department of Psychology, Bilkent University, Ankara/Turkey
5. Neuroscience Program, Aysel Sabuncu Brain Research Centre, Bilkent University, Ankara/Turkey
6. Helen Wills Neuroscience Institute, University of California, Berkeley, CA/USA

[*] Correspondence to:

Tolga Çukur, *cukur@ee.bilkent.edu.tr*

Abbreviated Title: **Search for visual actions warps semantic representations**

Manuscript summary:
Pages: 34
Abstract: 169 words
Significance Statement: 102 words
Introduction: 650 words
Discussion: 1497 words
Figures: 15
Extended Data Figures: 11

JNeurosci Accepted Manuscript

Abstract

Object and action perception in cluttered dynamic natural scenes relies on efficient allocation of limited brain resources to prioritize the attended targets over distractors. It has been suggested that during visual search for objects, distributed semantic representation of hundreds of object categories is warped to expand the representation of targets. Yet, little is known about whether and where in the brain visual search for action categories modulates semantic representations. To address this fundamental question, we studied brain activity recorded from five subjects (1 female) via functional magnetic resonance imaging while they viewed natural movies and searched for either *communication* or *locomotion* actions. We find that attention directed to action categories elicits tuning shifts that warp semantic representations broadly across neocortex, and that these shifts interact with intrinsic selectivity of cortical voxels for target actions. These results suggest that attention serves to facilitate task performance during social interactions by dynamically shifting semantic selectivity towards target actions, and that tuning shifts are a general feature of conceptual representations in the brain.

Keywords: action representation, attention, fMRI, voxelwise modelling, natural stimuli

Significance Statement

The ability to swiftly perceive the actions and intentions of others is a crucial skill for humans, which relies on efficient allocation of limited brain resources to prioritise the attended targets over distractors. However, little is known about the nature of high-level semantic representations during natural visual search for action categories. Here we provide the first evidence showing that attention significantly warps semantic representations by inducing tuning shifts in single cortical voxels, broadly spread across occipitotemporal, parietal, prefrontal, and cingulate cortices. This dynamic attentional mechanism can facilitate action perception by efficiently allocating neural resources to accentuate the representation of task-relevant action categories.

JNeurosci Accepted Manuscript

Introduction

The ability to swiftly perceive the actions and intentions of others is a crucial skill for all social animals. In the human brain this ability has been attributed to a network of occipitotemporal, parietal and premotor areas collectively called the action observation network (AON) (Caspers et al., 2010; Molinari et al., 2013; Oberman et al., 2007; Rozzi and Fogassi, 2017). Recent reports suggest that the AON hierarchically represents diverse information pertaining to actions, ranging from shape and kinematics to action-effector interactions and action categories (Grafton and de C Hamilton, 2007; Handjaras et al., 2015; Oosterhof et al., 2010, 2012, 2013; Urgen et al., 2019; Wurm et al., 2017; Cavina-Pratesi et al., 2018; Lingnau and Downing, 2015). Low-level shape and movement kinematics are represented in occipitotemporal cortex and in the posterior bank of inferior temporal cortex (Jastorff and Orban, 2009). Effector type (e.g., foot, hand) is represented in ventral premotor cortex (Corbo and Orban, 2017; Jastorff et al., 2010), while parietal cortex represents higher-level action categories (Abdollahi et al., 2012; Ferri et al., 2015).

Evidence suggests that selective attention alters population responses to actions across this representational hierarchy. Prior electrophysiology (Muthukumaraswamy and Singh, 2008; Muthukumaraswamy et al., 2004; Puglisi et al., 2017, 2018; Schuch et al., 2010) and neuroimaging studies (Herrington et al., 2012; de Lange et al., 2008; Nicholson et al., 2017; Rowe et al., 2002; Safford et al., 2010) have examined attention to low-level action features. Schuch et al. (2010) reported increased electroencephalography (EEG) responses in AON with attention to action kinematics. Safford et al. (2010) reported enhanced blood oxygen level dependent (BOLD) responses in superior temporal sulcus (STS) with attention to animate actors (i.e., humans) presented via simplified point-light displays (Johansson, 1973). Nicholson et al. (2017) reported enhanced responses in inferior frontal gyrus (IFG), occipitotemporal cortex, and middle frontal gyrus (MFG) with attention to action goals, in parietal cortex and fusiform gyrus with attention to manipulated objects. Few reports have further investigated the effects of attention to higher-level action features (Nastase et al., 2017, 2018). Presenting movie clips from various animal taxonomies performing several actions, Nastase et al. (2017) reported that attending to performed actions versus taxonomy alters multi-variate response patterns across anterior intraparietal sulcus (IPS) and premotor cortex.

Current electrophysiology and neuroimaging findings on visual actions suggest that attention increases AON responses to target features ranging from action kinematics and goals to actors. That said, high-level semantic representations during visual search for specific action categories remain understudied. Furthermore, prior studies did not question whether attending to action features causes baseline and gain changes, or rather elicits dynamic tuning shifts that can alter cortical representation. Recent evidence indicates that visual search for object categories shifts single-voxel category tuning toward target objects (Çukur et al., 2013). Therefore, it is likely that attention to action categories also causes tuning shifts to facilitate visual search. Here we hypothesised that natural visual search for action categories induces semantic tuning shifts in single cortical voxels toward targets. Tuning shift towards target categories elevates the local sampling density near the target actions, and expands target-action representations while compressing behaviourally-irrelevant-action representations by increasing the discriminability in the semantic neighbourhood of the finely-sampled action categories (Fig. 1).

JNeurosci Accepted Manuscript

To test the tuning-shift hypothesis, we recorded whole-brain BOLD responses while human subjects viewed 60min of natural movies and covertly searched for either 14 *communication* actions or 30 *locomotion* actions among 109 action categories in the movies (Fig. 2, Fig. 2-1). Using spatially-informed voxelwise modelling (Çelik et al., 2019), we measured category responses for hundreds of objects and actions in the movies separately for each individual subject and for each search task. We estimated a semantic space underlying action-category responses, and semantic tuning for action categories were measured by projecting voxel-wise model weights onto this space. Finally, semantic tuning profiles during the two search tasks were compared to quantify the magnitude and direction of tuning shifts in single voxels.

Materials and Methods

Subjects

Five healthy adult volunteers with normal or corrected-to-normal vision participated in this study: S1 (male, age 31), S2 (male, age 27), S3 (female, age 32), S4 (male, age 33), S5 (male, age 27). Data were collected at the University of California, Berkeley. The experimental protocol was approved by the Committee for the Protection of Human Subjects at the University of California, Berkeley. All participants gave written informed consent before scanning.

Stimuli and experimental design

Data for the main experiment were collected in six 10min 50s runs in a single session. Continuous natural movies were used as the stimulus in the main experiment. Three distinct 10min movie segments were compiled from short movie clips (10-20secs) without sound. Movie clips were selected from a diverse set of natural movies (see Nishimoto et al. (2011) for details). Movie clips were cropped into a square frame and downsampled to 512×512px. The movie stimulus was displayed at 15Hz on an MRI-compatible projector screen that covered 24ºx24º visual angle. Subjects were instructed to covertly search for target categories in the movies while maintaining fixation. A set of instructions regarding the experimental procedure and exemplars of the search targets were provided to the subjects before the experiment. A colour square of 0.16ºx0.16º at the centre with colour changing at 1Hz was used as the fixation spot. A cue word was displayed before each run to indicate the attention target: *communication* or *locomotion*. The *communication* target contained actions with the intent of communication, including both verbal communication actions and nonverbal gestural communication actions (e.g., talking, shouting, smirking). The *locomotion* target contained locomotion-related actions with the intent of moving animate entities, including humans and anthropomorphized animals (e.g., moving, running, driving). The same movie stimuli were used during each of the two attention tasks. The order of attention conditions was interleaved across runs to minimize subject expectation bias. This resulted in presentation of 1800sec of movies without repetition in each attention condition. Data from the first 20secs and last 30secs of each run were discarded to minimize effects of transient confounds. Following these procedures, 900 data samples for each attention condition were obtained.

JNeurosci Accepted Manuscript

A separate set of functional data were collected while the same set of subjects passively viewed 120min of natural movies (i.e., passive-viewing data; this dataset was also used in Huth et al., 2012 but here it was reanalysed with a focus on action categories). This dataset was used to construct the semantic space and to select voxels subjected to further analyses. Data for the passive-viewing experiment were collected in twelve 10min 50s runs in which 12 separate movie segments were displayed. Presentation procedures were the same between the main experiment and passive-viewing experiment, save for the number of runs. The passive-viewing dataset contained 3600 data samples.

fMRI data collection

Data were collected on a 3T Siemens Tim Trio MRI scanner (Siemens Medical Solutions) via a 32-channel receiver coil. Functional data were collected using a T2*-weighted gradient-echo echo-planar-imaging pulse sequence with the following parameters: TR=2sec, TE=33msec, water-excitation pulse with flip angle=70º, voxel size=2.24mm×2.24mm×4.13mm, field of view=224mm×224mm, 32 axial slices. To construct cortical surfaces, anatomical data were collected using a three-dimensional T1-weighted magnetization-prepared rapid-acquisition gradient-echo (MPRAGE) sequence with the following parameters: TR=2.3sec, TE=3.45 msec, flip angle=10º, voxel size=1mm×1mm×1mm, field of view=256mm×212mm×256mm. Surface flattening and visualisation were done via Freesurfer and PyCortex (Dale et al., 1999; Reuter et al., 2012; Gao et al., 2015).

fMRI data preprocessing

Motion correction was performed using Statistical Parametric Mapping toolbox (SPM12; Friston et al., 1995). Functional volumes were aligned to the first image from the first run in each subject. Brain tissue was identified using the brain extraction tool (BET) from the FSL software package (Smith, 2002). Low-frequency response components were detected using a third order Savitzky-Golay low-pass filter with 240sec temporal window and were removed from voxel responses. Voxel responses were then z-scored to attain zero mean and unit variance. Voxels within the 2mm neighbourhood of the cortical sheet were identified as cortical voxels in each subject (S1, 37791 voxels; S2, 32671 voxels; S3, 36942 voxels; S4, 42090 voxels; S5, 39254 voxels).

Definition of regions of interest

To define the anatomical regions of interest (ROIs) in each subject, the cortical surface was segmented into 156 regions of the Destrieux atlas (Destrieux et al., 2010) via Freesurfer. Segmentation results were projected from the anatomical space onto the functional space using PyCortex, and each voxel was assigned an anatomical label based on the projections. Functional ROIs were identified in each subject using visual category and retinotopic localizers (Huth et al., 2012). Localizer experiments for visual category-selective areas (fusiform face area, FFA; occipital face area, OFA; parahippocampal place area, PPA; retrosplenial cortex, RSC) were performed in six 4.5 min runs of 16 blocks (Huth et al., 2012). Subjects passively viewed 20 random static images from one of the objects, scenes, body parts, faces, or spatially scrambled

JNeurosci Accepted Manuscript

objects groups in each block. Each image was shown for 300ms following a 500ms blank period. PPA and RSC were identified as voxels with positive scene versus objects contrast (*t*-test, $p<10^{-4}$, uncorrected). FFA and OFA were defined using face-versus-object contrast (*t*-test, $p<10^{-4}$, uncorrected). The boundaries of these areas were hand drawn on the cortical surfaces along the contours at which the contrast level reached half of the maximum. Localizer experiment for early visual areas (RET: V1, V2, V3) contained four 9min 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. Finally, ROIs were refined to voxels inside the drawn boundaries near a 2mm neighbourhood of the cortical sheet.

Abbreviations for regions of interest and important sulci

Several regions of interest and important sulci were labelled on the flattened cortical surfaces to guide the reader.

***Regions of interest*:** pMTG, posterior middle temporal gyrus; pSTS, posterior superior temporal sulcus; AG, angular gyrus; SMG, supramarginal gyrus; IPS, intraparietal sulcus; aIP, anterior intraparietal cortex; PrCu, precuneous; dPMC, dorsal premotor cortex; BA44/45, Brodmann area 44/45; MFG, middle frontal gyrus; SFG, superior frontal gyrus; ACC, anterior cingulate cortex; RET, early visual areas V1-3; FFA, fusiform face area; OFA, occipital face area; PPA, parahippocampal place area; RSC, retrosplenial cortex.

***Sulci*:** TOS, temporo-occipital sulcus; STS, superior temporal sulcus; SF, Sylvian fissure; IFS, inferior frontal sulcus; MFS, middle frontal sulcus; SFS, superior frontal sulcus.

Head motion, eye-movement, and physiological noise

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. The cardiac and respiratory activity during the main experiment were recorded using a pulse oximeter and a pneumatic belt. These data were then used to estimate two regressors to capture respiration and nine regressors to capture cardiac activity (Verstynen and Deshpande, 2011).

To ensure that eye-movements did not unduly bias the results, several control analyses were performed. ViewPoint EyeTracker (Arrington Research) was used to monitor subjects' eye positions at 60Hz, after getting calibrated at the beginning of each experimental run. Kruskal-Wallis tests were used to detect systematic differences in the distribution of eye position and movement. The distribution of eye position during search for *communication* and *locomotion* tasks were examined. We find that the distribution of eye position is not affected by search task (p=0.17), or by target presence or absence (p=0.74), and no significant interactions are present between these two factors (p=0.60). To test whether eye movement is affected by target or distractor detection, the distribution of eye position during a 1 sec window around target onset and target offset was studied. The eye position distribution is not affected by target onset (p=0.73) or offset (p=0.17), and there is no significant interaction between the aforementioned

JNeurosci Accepted Manuscript

factors (p=0.83). Furthermore, the moving-average standard deviation of eye position was studied in a 200ms window to determine systematic differences in rapid moment-to-moment variations in eye position across the two search tasks. There are no significant effects of search task (p=0.11), target presence or absence (p=0.32), target onset (p=0.49), or target offset (p=0.36), and there are no significant interactions between these factors (p=0.16). Finally, 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 maintain subject vigilance, the subjects were instructed to depress a button whenever they detected a member of the target category in the stimulus (i.e., either a communication or a locomotion action depending on the search task). The behavioural responses were initially analysed to ensure that subjects performed the tasks, and that task difficulty was balanced across search targets. The target detection rate was 89±9% for the communication and 91±8% for the locomotion targets (mean±std across subjects), with no significant difference between the two tasks (bootstrap test, p>0.05).

Category features

A category feature space was constructed to encode the information pertaining to object and action categories in the movies. Each second of the movie stimulus was manually labelled using the WordNet lexicon (Miller, 1995) to find the time course for presence of 922 different object and action categories in the movie stimulus. This yielded an indicator matrix where each row represents a one-second clip of the movie stimulus, and each column represents a category. Finally, category features were obtained by downsampling the indicator matrix to 0.5Hz to match the acquisition rate of fMRI.

Motion-energy features

To infer cortical selectivity for low-level scene features, local spatial frequency and orientation information of each frame of the movie stimulus were quantified using a motion-energy filter bank. The filter bank contained 2139 Gabor filters that were computed at eight directions (0 to 350º, in 45º steps), three temporal frequencies (0, 2, and 4Hz), and six spatial frequencies (0, 1.5, 3, 6, 12, and 24 cycles/image). Filters were placed on a square grid spanning the 24ºx24º field of view. The luminance channel was extracted from the movie frames and passed through the filter bank. The outputs were then passed through a compressive nonlinearity to yield the motion-energy features (Lescroart and Gallant, 2019; Nishimoto et al., 2011). Finally, the motion-energy features were temporally downsampled to match the fMRI acquisition rate.

Space-time Interest Points (STIP) features

Intermediate-level kinematic information of the movies were quantified by constructing the Space-Time Interest Point (STIP) features using STIP toolbox (Laptev, 2005; Laptev et al., 2008). STIP features have been successfully leveraged in many computer vision applications to recognize human actions. As detailed in Laptev (2005) and Laptev et al. (2008), Harris operators (Harris and Stephens, 1988) were used to identify spatiotemporal interest points in the movie stimulus at multiple scales $(\sigma_i^2, \tau_j^2)=(2^{1+i}, 2^j)$, $i \in \{1, \ldots, 6\}$, $j \in \{1,2\}$, where σ and τ are the

JNeurosci Accepted Manuscript

standard deviations of the Gaussian kernels in spatial and temporal domains respectively. Histograms of oriented gradients (HoG; Dalal and Triggs, 2005), and histograms of optical flow (HoF; Holte et al., 2010) were calculated in the $(\Delta_{x,i}, \Delta_{y,i}, \Delta_{t,j})$ spatiotemporal neighbourhood of each interest point, where $\Delta_{x,i} = \Delta_{y,i} = 2k\sigma_i$ and $\Delta_{t,j} = 2k\tau_j$, and k is the scale factor. Scale factor was set to 9 according to the default configuration of the toolbox. Finally, normalized histograms were concatenated to construct the collection of 162 STIP features and were downsampled to match the acquisition rate of fMRI.

Model estimation and testing

Separate linearized models were fit in each voxel to estimate model weights that map each set of features (i.e., category, motion-energy, or STIP features) to the measured BOLD responses in each search task in individual subjects. Banded-ridge regression (Nunez-Elizalde et al., 2019) was used to fit the models. To capture the hemodynamic response, delayed feature time-courses were concatenated. Delays of two, three, and four samples, corresponding to 4, 6, and 8secs were used. To account for potential correlations between target detection and BOLD responses, a nuisance target-presence regressor was included in the model. The target-presence regressor contained the category regressor for *communication* during search for *communication* task and the category regressor for *locomotion* during search for *locomotion* task. Model fitting for the two search tasks was performed concurrently by concatenating the features and BOLD responses across search tasks (see Fig. 2). This procedure ensured consistency between the assigned regularization parameters across search tasks and enabled utilisation of the target regressor (Shahdloo et al., 2020).

A nested cross-validation (CV) procedure was used to choose the regularization parameters and estimate model weights. Data from the main experiment were segmented into 60 30-sec blocks. In each of the 10 outer folds, 4 randomly chosen blocks were held-out as validation data. Then, in each of the 10 inner folds, 54 randomly chosen blocks were used as training data and the 2 remaining blocks were used as test data. To fit models for the passive-viewing data, data were segmented into 144 50-sec blocks. In each fold, 8 randomly chosen blocks were held-out as validation data, 132 randomly chosen blocks were used as training data and the 4 remaining blocks were used as test data. For each feature set, regularisation parameters were selected with a random-search; a thousand normalized regularisation parameter candidates were sampled from a Dirichlet distribution and were scaled by 30 log- spaced values ranging from 10^{-5} to 10^{20}. Training data were used to fit models for each set of regularisation parameters independently. Model weights were then used to predict responses in the test data and prediction scores of the fit models were assessed. Prediction scores were taken as product-moment correlation coefficient between measured and predicted voxel responses. The set of regularisation parameters maximizing the average prediction score across inner CV folds was chosen in each voxel. Finally, the optimal set of parameters were used to fit models on the union of training and test data in each outer fold and model weights were averaged across the outer folds.

Finally, prediction performance of the fit models were evaluated. In each outer fold, after discarding the nuisance regressors, responses were predicted for the validation data using the fit models and prediction scores were averaged across the search tasks. Prediction scores were then averaged across the outer folds.

JNeurosci Accepted Manuscript

For each voxel, separate linearized models were estimated to relate each feature representation to the BOLD responses. Specifically, category models were fit to estimate category responses that represented the contribution of each category to single-voxel BOLD responses separately for the data in the main experiment and the passive-viewing data in individual subjects. Furthermore, a motion energy model and a STIP model were fit in each voxel to represent the contribution of the low- and intermediate-level stimulus features to the responses. These alternative models were further used to select analysis voxels (i.e., semantic voxels).

Variance partitioning

Object-action categories can be correlated with low-level visual features of natural movies (Lescroart and Gallant, 2019), and there is evidence for representation of intermediate-level action features (e.g., action kinematics) across cortex (Jastorff et al., 2010). Therefore, there is a possibility that the estimated category responses are confounded by selectivity for low- and intermediate-level scene features. To control for potential confounds, we performed a variance partitioning analysis. This analysis estimates the response variance that is uniquely explained by the category model after accounting for variance that can be attributed to low- and intermediate-level features captured by the motion-energy and STIP models. To do this, we separately measured the variance explained when all three models (category, motion-energy, and STIP) are fit simultaneously (i.e., combined model), and variance explained when only motion-energy and STIP models are fit simultaneously (i.e., control model). Banded ridge regression was used to fit the combined and control models to prevent bias in assigning regularisation parameters across different feature sets. The explained variance (R^2) was calculated as squared prediction scores, separately for the combined and control models. Note that from a model fitting perspective, negative prediction scores correspond to zero explained variance. Finally, unique variance explained by the category model was calculated as

$$\widehat{R^2}_{cat} = R^2_{comb} - R^2_{cont} \quad (1)$$

Here $\widehat{R^2}_{cat}$ is the variance uniquely explained by the category model after accounting for low- and intermediate-level features, R^2_{comb} is the variance explained by the combined model, and R^2_{cont} is the variance explained by the union of motion-energy and STIP models in each voxel.

Action category responses

The fit category responses reflect voxel tuning for each of the 922 object and action categories in the movie stimulus. To infer tuning for action categories, 922-dimensional category responses were masked to select only the 109 action categories. This yielded the voxelwise 109-dimensional action category responses.

JNeurosci Accepted Manuscript

316 **Semantic representation of actions**

317 Passive-viewing data were used to construct a continuous semantic space for action category
318 representation. In this space, semantically similar action categories would project to nearby
319 points, whereas semantically dissimilar categories would project to distant points (Huth et al.,
320 2012). Category models were fit and action category responses during passive viewing were
321 estimated. A group semantic space was then obtained using principal component analysis (PCA)
322 on the action category responses of cortical voxels pooled across all subjects. To maximize the
323 quality of the semantic space, voxels in which the category model predicted unique response
324 variance after accounting for the variance attributed to low- and intermediate-level stimulus
325 features were selected. These voxels were further refined to include only the top 3,000 best
326 predicted voxels within each subject. The top 12 principal components (PCs) that explained more
327 than 95% of the variance in responses were selected. Subsequent analyses were also repeated
328 using the top 8 PCs that explained more than 90% of the response variance but the results
329 remained consistent. Semantic tuning profile for each voxel under each search task was then
330 obtained by projecting the respective action category responses onto the PCs. To illustrate the
331 semantic content of the PCs, characteristic actions of the movie stimulus were clustered in the
332 semantic space, and cluster centres were projected onto the PCs after getting labelled (see Fig.
333 6).
334

335 **Consistency of the semantic space across subjects**

336 To test whether the estimated semantic space is consistent across subjects, we used a leave-one-
337 out cross-validation procedure. In each cross-validation fold, voxels from four subjects were
338 used to derive 12 PCs to construct a semantic space. In the left-out subject, semantic tuning
339 profile for each voxel was obtained by projecting action category responses during passive
340 viewing onto the derived PCs. Next, product-moment correlation coefficient was calculated
341 between the tuning profiles in the derived space and the tuning profiles in the original semantic
342 space. Results were averaged across semantic voxels in the left-out subject. The cross-validated
343 semantic spaces consistently correlate with the original semantic space (Fig. 7).
344

345 **Characterizing tuning shifts**

346 Attentional tuning shifts toward or away from targets would be reflected in modulation of
347 semantic selectivity for *communication* or *locomotion* action categories. Thus, the magnitude and
348 direction of tuning shifts can be assessed by comparing the semantic selectivity for these
349 categories between the two search tasks. Semantic selectivity for the two target categories was
350 quantified as the similarity between semantic tuning profiles and idealized templates tuned solely
351 for *communication* or *locomotion* action categories. First, idealized category responses were
352 constructed as 109-dimensional vectors that contained ones for target categories (either
353 *communication* or *locomotion* categories) and zeros elsewhere. Idealized templates were then
354 obtained by projecting these idealized category responses onto the semantic space. Semantic
355 selectivity for each target category was quantified as product-moment correlation coefficient
356 between voxel semantic tuning profile and the corresponding template
357

358 $$T_{i,C}=corr(s_i,s'_C) \qquad (1)$$

JNeurosci Accepted Manuscript

$$T_{i,L}=corr(s_i,s'_L) \quad (2)$$

where $T_{i,C}$ and $T_{i,L}$ are the tuning strength for *communication* and *locomotion* during condition $i \in$ {C, L} denoting attend to *communication* or attend to *locomotion*, s_i is the semantic tuning profile during condition i, and s'_C and s'_L denote the idealized semantic tuning templates for *communication* and *locomotion*, respectively. Finally, voxelwise tuning shift index (TSI_{all}) was quantified as

$$TSI_{all} = \frac{(T_{C,C}-T_{C,L})+(T_{L,L}-T_{L,C})}{2-sign(T_{C,C}-T_{C,L})T_{C,L}-sign(T_{L,L}-T_{L,C})T_{L,C}} \quad (3)$$

The numerator of TSI captures the difference in semantic selectivity for the attended versus unattended category, summed over the two attention tasks (i.e., search for communication and search for locomotion). Observing that the maximum possible selectivity for the attended category is 1, obtained when voxel tuning is equivalent to the idealized template, the denominator is cast to normalize the potential range of the TSI metric between 1 and -1 without affecting its sign. Tuning shifts toward the attended category would yield positive values where a TSI_{all} of 1 indicates a complete match between voxel semantic tuning and idealized templates, whereas negative values would indicate shifts away from the attended category where a TSI_{all} of -1 indicates a complete mismatch between voxel tuning and idealized templates. A TSI_{all} of 0 would indicate that the voxel tuning did not shift between the two search tasks.

The TSI metric in Eq. 3 can also be adopted to calculate tuning changes for any given set of action categories. To do this, the 922-dimensional category responses measured during attention tasks were masked to keep only the responses for the given set of actions. The masked tuning vectors and the idealised template for the given set were then projected onto the 12-dimensional semantic space. Semantic selectivity of a voxel to the given set was taken as the correlation coefficient between the projections of voxel tuning and the idealised template in the semantic space. Attentional modulation of semantic tuning for nontarget categories was examined by calculating a separate tuning shift index (TSI_{nt}). Note that this index can be calculated based on Eq. 3, but by zeroing out the category responses for communication and locomotion actions prior to projection onto the semantic space. To study the tuning shifts in an ROI, TSIs were averaged across semantic voxels within the ROI.

The change in voxelwise tuning during attending to the first target (e.g., communication) versus to the second target (e.g., locomotion) was defined as the l_1-norm of the tuning difference between the two conditions. This calculated tuning change can be linearly decomposed into a component explained by the target features (i.e., the union of communication and locomotion features) and a component explained by the nontarget features (i.e., all features excluding the target features). The fraction of tuning change for target/nontarget features was computed by taking the ratio of the respective component to the overall tuning change.

JNeurosci Accepted Manuscript

Characterizing target preference during visual search

To investigate the interaction between tuning shifts and intrinsic selectivity for individual target action categories, we quantified a target preference index (PI∈ [−1,1]) separately during search for communication actions (PI_{com}) and during search for locomotion actions (PI_{loc}). PI during search for each target action was taken as the difference in selectivity for the attended versus the unattended target

$$PI_{com} = \frac{T_{C,C} - T_{C,L}}{1 - sign(T_{C,C} - T_{C,L}) T_{C,L}} \quad (4)$$

$$PI_{loc} = \frac{T_{L,L} - T_{L,C}}{1 - sign(T_{L,L} - T_{L,C}) T_{L,C}} \quad (5)$$

where PI_{com} denotes the relative tuning preference for communication actions during search for communication, and PI_{loc} denotes the relative tuning preference for locomotion actions during search for locomotion. In this scheme, a PI of 1 indicates a complete match between voxel semantic tuning and the idealized template for the target, whereas a PI of -1 indicates a complete mismatch between voxel tuning and the idealized template for the target. Finally, a PI of 0 indicates that the voxel semantic tuning does not shift toward any of the target actions.

Characterizing action category preference during passive viewing

To investigate the interaction between calculated preference index for individual targets and intrinsic selectivity for action categories, we quantified a selectivity index (SI∈ [−1,1]), separately for communication actions (SI_{com}) and for locomotion actions (SI_{loc}). SI for each target in each voxel was calculated as the product-moment correlation coefficient between the voxel category response and idealised template category tuning for the given target.

Action clustering

To facilitate interpretation of stimulus information captured by individual PCs, characteristic action content of the movies was clustered and labelled. Action content ($\bar{C}$) for each short clip was calculated as the number of frames where each of the 109 actions were present ($\overline{N_a}$), and was normalized by the total number of clip frames (N)

$$\bar{C} = \frac{\overline{N_a}}{N} \quad (2)$$

JNeurosci Accepted Manuscript

This yielded a 109-dimensional action content vector for each clip. The action content vectors were then projected onto the semantic space and were grouped into 10 clusters using k-means. The number of clusters was optimized using the elbow method (Thorndike, 1953). Average action content of each clip (A) was calculated as the mean of the clip's action content vector

$$A = \frac{\sum c}{109} \quad (3)$$

where $\bar{C} = [\boldsymbol{c_1}, \boldsymbol{c_2}, \boldsymbol{c_3}, \ldots, \boldsymbol{c_{109}}]$. To label the clusters, five clips with highest average action contents within each cluster were selected. Four candidate labels for each cluster were manually assigned and 15 evaluators were asked to score (from 1 to 5) the correspondence of the selected clips to each of the four candidate labels. Finally, the label with the highest score was selected to represent each cluster.

Statistical Analyses

Bootstrap tests were used to assess statistical significance. To assess significance of the prediction scores, single-voxels predicted responses were resampled 5000 times with replacement. For each bootstrap sample, prediction score was computed. Significance level (p-value) of the prediction scores was taken as the fraction of bootstrap samples in which the prediction scores was greater than 0. Significance level of the unique response variance (Eq. 1) was taken as the fraction of bootstrap samples in which the unique variance explained by the category model was greater than 0. All single-voxel significance levels were corrected to account for multiple comparisons using false-discovery rate correction (FDR; Benjamini and Hochberg 1995).

Significance of TSI_{all}, TSI_{nt}, PI_{com}, and PI_{loc} were assessed for each ROI across subjects. To do this, ROI-wise metrics were resampled across subjects with replacement 10000 times. Significance level was taken as the fraction of bootstrap samples where the test metric averaged across resampled subjects is less than 0 (for right-sided tests) or greater than 0 (for left-sided tests). This procedure was performed in a total of 21 functional ROIs separately. All ROI significance levels were corrected to account for multiple comparisons using FDR.

In ROIs with a significant metric across subjects, the metric was further tested within individual subjects. To do this, semantic voxels within a given ROI were resampled with replacement 10000 times. For each bootstrap sample, mean value of a given metric was computed across resampled voxels. Significance level was taken as the fraction of bootstrap samples in which the tested metric was less than 0 (for right-sided tests) or greater than 0 (for left-sided tests).

Results

Visual search modulates category responses

Little is known on whether and where in the brain natural visual search for action categories warps semantic representations. To answer this question, we investigated voxel-wise tuning for

JNeurosci Accepted Manuscript

hundreds of object and action categories across cortex. Human subjects viewed natural movies and covertly searched for *communication* or *locomotion* actions. Category regressors were constructed to label presence of 922 distinct object and action categories in the movies. Separate category models were then fit in each voxel for each search task. These models enabled us to measure single-voxel category responses during each search task (Fig. 2a, see *Materials and Methods*).

As natural stimuli contain correlations among various levels of features, there is a possibility that estimated category responses are confounded by voxel tuning for low- and intermediate-level scene features. To rule out this potential confound, we measured the response variance explained by low-level motion-energy features, and intermediate-level spatiotemporal interest point (STIP) features. Motion-energy features were constructed using a pyramid of spatiotemporal Gabor filters (Nishimoto and Gallant, 2011). STIP features, providing an intermediate representational basis for human actions, were constructed by measuring optical flow over interest points with significant spatiotemporal variation (Laptev et al., 2008). We identified voxels in which the category model explained unique response variance after accounting for these alternative features via variance partitioning, and subsequent analyses were conducted on this set of uniquely explained voxels. To prevent bias in voxel selection due to attention, variance partitioning was performed on a separate dataset collected for this purpose (i.e., *passive-viewing dataset*, see *Materials and Methods*). We find that the category model explains unique response variance after accounting for low- and intermediate-level features in 25.7±1.6% of cortical voxels (mean±sem across five subjects; bootstrap test, $q(FDR)<0.05$; Figs. 4 and 5), yielding 8,613-13,435 voxels in individual subjects (henceforth called *the semantic voxels*).

Comparison of estimated category responses across search tasks would be justified only if the fit models can accurately predict BOLD responses that were held-out during model fitting. To assess prediction performance of the fit category models, we measured average prediction scores across the two search tasks, taken as product-moment correlation coefficient between the predicted and measured held-out responses (Fig. 2b). Category models have high prediction scores (greater than 1 std above the mean) in 46.9±0.6% of the semantic voxels. These include many voxels spread across the AON comprising occipitotemporal, parietal, and premotor cortices, as well as voxels in prefrontal and cingulate cortices (Fig. 3).

A recent study provided the first evidence that attention can alter single-voxel category tuning profiles during search for object categories (Çukur et al., 2013). We thus hypothesised that visual search for action categories can also cause changes in voxel-wise category tuning. If attentional tuning changes are significant, the category models fit to individual search tasks should yield higher prediction scores than a null model fit by pooling data across the two search tasks. To test this prediction, we compared the prediction scores obtained from the category and null models. We find that the category model significantly outperforms the null model in 46.1±1.8% of semantic voxels (bootstrap test, $q(FDR)<0.05$). Additional control analyses further ensured that these attentional changes cannot be attributed to residual eye-movements, head-motion, physiological noise, or target-detection biases (see *Methods*). Taken together, these results suggest that many cortical voxels in occipitotemporal, parietal, and prefrontal cortices encode high-level category information, and that action-based visual search significantly modulates category responses in single voxels.

JNeurosci Accepted Manuscript

Visual search warps semantic representation of actions

Previous studies suggest that the human brain represents visual categories by embedding them in a continuous semantic space (Huth et al., 2012). Here, we used linear encoding models to map category features of natural movies onto the recorded BOLD responses in single voxels. The model features, namely actions, are fundamental semantic concepts in both language and vision. The models successfully predict brain activity in cortical voxels, after controlling for lower-levels of features (i.e., motion energy and STIP features). Thus, from a quantitative perspective, it could be argued that there is an explicit representation of the semantic categories of actions in the voxel responses (Naselaris et al., 2011). Note that a theoretical characterization of relationships among semantic concepts is difficult. In computational semantics, an empirical approach is adopted instead that is rooted in the distributional hypothesis. This hypothesis states that concepts with similar statistical distributions have similar meaning. Accordingly, co-occurrence statistics of concepts in corpora are used as a proxy metric for similarity of meaning in many methods for learning semantic relationships (Jurafsky and Martin, 2021). Here, to derive a semantic space underlying action category representations, we performed principal component analysis (PCA) on the model weights for action categories. Visual search for actions alters category model weights as reported here, so performing PCA on data from search tasks can bias estimates of the semantic space. Instead, we derived the semantic space using the passive-viewing dataset. Action categories that are semantically close to each other should project to nearby points in this space, whereas semantically dissimilar categories should project to distant points. The top twelve principal components (PCs) that explained more than 95% of the variance in responses were selected, which showed a high degree of inter-subject consistency ($r=0.52\pm0.02$ mean±sem across subjects; Fig. 7). To visually examine the semantic information captured by this space, we projected action categories onto the PCs (Fig. 6a; projections onto the first three dimensions that accounted for 72.8% of the response variance is shown in Fig. 9; loadings for all PCs are shown in Fig. 10). All further quantitative analyses regarding tuning shifts were instead conducted in the full semantic space of 12 dimensions, including all the identified PCs.

Previous evidence suggests that visual search shifts single-voxel tuning profiles to expand the representation of the targets (Çukur et al., 2013). Thus, it is possible that action-based visual search also shifts semantic tuning in single-voxels towards the target category. To investigate this possibility, we projected action category responses onto the semantic space. The first and third PCs maximally differentiated between actions belonging to the target categories (i.e., *communication* versus *locomotion* categories, Fig. 8). Therefore, we visually compared the projections onto these PCs across the two search tasks. We observe that attention causes semantic tuning modulations broadly across cortex (Fig. 6b; see Figs. 6-1 to 6-5 for results in individual brain spaces). Specifically, voxels in inferior posterior parietal cortex (PPC), cingulate cortex, and anterior inferior prefrontal cortex shift their tuning toward communication during search for communication actions. Meanwhile, voxels in superior PPC and medial parietal cortex shift their tuning toward locomotion during search for locomotion actions. Several reports suggest involvement of superior PPC in representing locomotion actions (Corbo and Orban, 2017), and inferior PPC in representing communication actions (Abdollahi et al., 2012, Rizzolatti and Matelli, 2003). Therefore, our findings suggest that during search for a given action

JNeurosci Accepted Manuscript

category, tuning shifts toward the target category are most prominent in voxels that are primarily selective for the target.

Visual search for action categories shifts single-voxel semantic tuning profiles

Our inspection of semantic representations during visual search reveals that attention broadly modulates high-level action representations by shifting semantic tuning profiles in single voxels. To quantify the magnitude and direction of these tuning changes, we separately measured semantic selectivity for *communication* and *locomotion* action categories in each search task. The 922-dimensional category responses for individual voxels measured during attention tasks, and idealised template vectors for the targets were projected onto the semantic space. The template vector for a target is constructed as a 109-dimensional indicator vector containing ones for the target category and all its subordinate categories, and zeros for remaining categories. For instance, the "locomotion" template has ones for *locomotion*, and for *walk, run, crawl, move, ride, etc*. As such, the target template vector indexes the target action as well as actions that are semantically related to the target according to the WordNet hierarchy (see *Methods*). For each attention task, semantic selectivity of a given voxel for a target category was then quantified as the correlation coefficient between projected 12-dimensional vectors characterizing the voxel-wise tuning profile and the idealised template in the semantic space. For each voxel, a tuning shift index ($TSI_{all} \in [-1,1]$) was taken as the difference in semantic selectivity for targets when they were attended versus unattended. A positive TSI_{all} indicates shifts towards the target, a negative TSI_{all} indicates shifts away from the target, and a TSI_{all} of 0 suggests no change in between tasks (see *Materials and Methods*).

We find that voxels across many cortical regions shift their tuning toward the attended category (Fig. 11a; see Figs. 11-1a to 11-5a for results in individual brain spaces). Figure 15a shows respective tuning shifts in relevant regions of interest (ROIs). Tuning shifts are significantly greater than zero in many areas across AON including occipitotemporal cortex (posterior STS, pSTS; posterior MTG, pMTG), posterior parietal cortex (intraparietal sulcus, IPS; AG, SMG), and premotor cortex (Brodmann's areas 44, 45, BA44/45; bootstrap test $q(FDR)<0.05$; Fig. 15a). This result suggests that focused attention to specific action categories shifts semantic tuning toward targets in single-voxels, and that these attentional modulations are present at all levels of the AON hierarchy including occipitotemporal cortex.

Prior evidence suggests that during category-based visual search, semantic tuning shifts grow stronger toward later stages of semantic processing (Çukur et al., 2013). Here, we find that semantic tuning shifts in AG and SMG are significantly stronger than those in occipitotemporal (pSTS, pMTG) and premotor cortices (i.e., averaged over AG and SMG, compared with the average over pSTS and pMTG, and with the average over dorsal premotor cortex (dPMC) and BA44/45; Cohen's $d=1.36$, $p<0.05$). Therefore, the tuning shifts reported here could indicate that AG and SMG are higher nodes in the hierarchy of semantic representation of action categories. In a previous study, we reported that in medial prefrontal cortex visual search for object categories causes tuning shifts toward targets while it causes tuning shifts away from targets in voxels in precuneous (PrCu) and temporo-parietal junction (TPJ; Çukur et al., 2013). Similarly, by qualitative inspection of the flatmaps, here we observe that visual search for action categories causes negative tuning shifts in many voxels across PrCu and TPJ. These results suggest that

JNeurosci Accepted Manuscript

these areas might be involved in distractor detection and in error monitoring during visual search for actions (Corbetta and Shulman, 2002).

Visual search shifts semantic tuning for nontarget action categories

Natural visual search for object categories was previously suggested to cause changes in representations of not only targets but also nontarget categories (Çukur et al., 2013; Seidl et al., 2012). Thus, it is likely that action-based visual search shifts semantic tuning for nontarget categories. To address this important question, we first examined the separate contributions of tuning changes for target versus nontarget categories to the overall tuning shifts. Specifically, we measured the fraction of overall tuning shifts that can be attributed to the target categories versus nontarget categories (i.e., all categories excluding *communication* and *locomotion* actions). We find that both target and nontarget categories significantly contribute to the overall tuning shifts (bootstrap test, $q(FDR)<0.05$).

However, as would be expected, target categories account for a relatively larger fraction of the overall tuning shifts compared to nontarget categories in all studied ROIs, except in early visual cortex ($q(FDR)<0.05$; Fig. 13). Next, to explicitly quantify tuning shifts for nontarget categories, we calculated a separate tuning shift index exclusively on nontarget categories (TSI_{nt}). To calculate TSI_{nt}, the 109-dimensional action category response vectors were masked to select nontarget categories, prior to projection onto the semantic space (see *Materials and Methods*). We observe that tuning shift for nontarget categories is generally smaller than the overall tuning shift (Fig. 11b versus Fig. 11a and Fig. 12; see Figs. 11-1b to 11-5b for results in individual brain spaces). Yet, TSI_{nt} is non-significant in all ROIs except AG, SMG, and BA45 ($q(FDR)<0.05$; Fig. 15b). Note that an insignificant TSI_{nt} does not necessarily suggest that attention has not altered tuning for non-target categories, but rather the direction of tuning changes could be merely not aligned towards or away from the target categories in the semantic space. Thus, these results suggest that, compared to occipitotemporal areas, attention more diversely warps semantic representations in parietal and premotor AON nodes by shifting tuning for both target and nontarget categories.

Tuning shifts interact with intrinsic selectivity of cortical voxels for action categories

A recent study on visual attention has reported that in strongly object-selective regions voxel tuning for a preferred object might be robust against attention directed to a nonpreferred object (e.g., *houses* for fusiform face area, FFA, and *faces* for parahippocampal place area, PPA; Çukur et al., 2013). This previous result suggests that the degree of response modulations in a brain region might depend on the alignment between the search target and the intrinsically preferred object. It is thus likely that tuning shifts during search for an action category also interact with the intrinsic selectivity of cortical voxels for the target category. Tuning shifts as measured by TSI signal an overall increase in relative selectivity for target versus nontarget categories, aggregated across search tasks. Yet, interaction of tuning shifts with intrinsic selectivity for action categories is task-specific by definition. Therefore, to examine potential interactions, we calculated a target preference index ($PI \in [-1,1]$) separately during search for communication actions (PI_{com}) and during search for locomotion actions (PI_{loc}). PI_{com} was taken as the difference in selectivity for *communication* versus *locomotion*, during search for communication actions.

JNeurosci Accepted Manuscript

Analogously, PI_{loc} was taken as the difference in selectivity for locomotion versus *communication*, during search for locomotion actions.

Voxel-wise PI_{com} and PI_{loc} values were projected onto cortical flat maps for visual inspection (Fig. 14; see Figs. 11-1c to 11-5c for results in individual brain spaces) and quantitatively examined in ROIs (Fig. 15c, d). We observe that semantic tuning in areas with indiscriminate selectivity for behaviourally relevant action categories (e.g., selective for low-level visual features or static object categories) show insignificant shifts regardless of the search task. Meanwhile, many voxels across anterior parietal, occipital, and cingulate cortices –with intrinsic action category preferences– show differential preference for one of the two target action categories as indicated by high PI index during either search for communication or search for locomotion actions. Lastly, semantic tuning in voxels across posterior parietal and anterior prefrontal cortices with broad selectivity for actions shift toward the attended category irrespective of the search target. These specific cases are discussed in detail below.

Areas where both PI_{com} and PI_{loc} are non-significant

We find that PI_{com} and PI_{loc} are non-significant in retinotopic early visual areas (RET; bootstrap test, $q(FDR)>0.05$) that represent low-level stimulus features, low-level motion-selective area (hMT; $q(FDR)>0.05$), and object-selective areas (FFA; occipitotemporal face area, OFA; PPA; retrosplenial cortex, RSC; and EBA; $q(FDR)>0.05$). Furthermore, PI_{com} and PI_{loc} are non-significant in anterior intraparietal cortex (aIP; $q(FDR)>0.05$), which is not involved in representing communication or locomotion actions (non-significant SI_{com} and SI_{loc}, $q(FDR)>0.05$) (Noppeney, 2008; Rizzolatti et al., 1997; Urgen and Orban, 2021). These results suggest that during action-based search, semantic tuning does not change substantially in cortical areas that are selective for lower-level visual features or for neutral high-level action categories irrelevant to the task.

Areas where either PI_{com} or PI_{loc} are significant

Several previous studies suggest that lateral and medial prefrontal cortices are causally involved in representing communication actions (Van Overwalle, 2009; Wilson-Mendenhall et al., 2013). Here, we find that PI_{loc} is non-significant while PI_{com} is significantly greater than zero in anterior inferior frontal gyrus (BA44/45; d=1.94, $q(FDR)<0.05$; $SI_{com}=0.12$, $q(FDR)<0.05$), in superior frontal gyrus (SFG; d=1.94, $q(FDR)<0.05$; $SI_{com}=0.18$, $q(FDR)<0.05$), and in anterior cingulate cortex (ACC; d=.34, $q(FDR)<0.05$; $SI_{com}=0.18$, $q(FDR)<0.05$). On the other hand, previous reports provide evidence for representation of animate locomotion actions in PPC, including IPS ($SI_{loc}=0.15$, $q(FDR)<0.05$) (Abdollahi et al., 2012; Battelli et al., 2003; Bremmer et al., 2001; Ilg et al., 2004). In accord, we find that PI_{com} is non-significant while PI_{loc} is significantly greater than zero in IPS (d=3.95, $q(FDR)<0.05$). Taken together, our findings suggest that in areas that are strongly selective for specific action categories, visual search for the preferred action shifts tuning more vigorously towards the preferred target category. It is also worth noting that these attentional effects are not limited to the AON, but rather extend to higher-order cortical areas involved in social cognition. Lastly, we find that PI_{loc} is significantly less than zero while PI_{com} is non-significant (d=0.73, $q(FDR)>0.05$; $SI_{loc}=-0.23$, $q(FDR)<0.05$) in dPMC. This result supports

JNeurosci Accepted Manuscript

the view that dPMC enhances the representation of distractors during search for locomotion actions (Anticevic et al., 2010; Toepper et al., 2010; Zhou et al., 2012).

Areas where both PI_{com} and PI_{loc} are significant

Posterior STS (pSTS), posterior middle temporal gyrus (pMTG), and SMG are considered as AON nodes that maintain representation of actions regardless of their semantic category (Caspers et al., 2010; Jastorff et al., 2016; Lui et al., 2008). We find that both PI_{com} and PI_{loc} are significantly greater than zero in pSTS, pMTG, and SMG, consistent with their generic action selectivity. In addition, several previous studies suggest that MFG –as a node in dorsal attention network– facilitates visual search by maintaining the representation of targets (Corbetta and Shulman, 2002; Mars and Grol, 2007; Paneri and Gregoriou, 2017; Ptak et al., 2017). Accordingly, here we find that PI_{com} and PI_{loc} are significantly greater than zero in MFG ($q(FDR)<0.05$). Overall, these results indicate that in areas with generic action selectivity and in high-level cortical areas, attention facilitates action-based search by shifting representations toward targets irrespective of their semantic category.

The results presented here can be explored online via an interactive brain viewer at http://www.icon.bilkent.edu.tr/brainviewer/shahdloo_etal/.

Discussion

Several previous studies have reported response modulations during action-based attention in parietal and prefrontal cortices, but not in occipitotemporal areas (Nastase et al., 2017, 2018; Nicholson et al., 2017). Yet we observe significant attentional tuning shifts in occipitotemporal cortex. Unlike previous studies, our analysis approach enables us to measure single-voxel tuning. Our movie stimulus contains a large set of action categories in natural context in contrast to controlled stimuli with a handful of actions on a homogeneous background. Lastly, we investigate actions that are performed by animate actors, known to elicit robust responses across the occipitotemporal cortex (Isik et al., 2017; Thompson and Parasuraman, 2012; Walbrin and Koldewyn, 2019; Walbrin et al., 2018). These design factors might have enabled us to detect tuning shifts in early stages of AON comprising occipitotemporal areas.

Recent studies emphasize the role of AG and SMG in multi-modal semantic representation while observing actions, hearing action sounds, or reading action words (Bedny and Caramazza, 2011; van Dam et al., 2010; Liljeström et al., 2008; Pizzamiglio et al., 2005). Evidence also suggests that during semantic processing these areas act as central connectivity hubs, passing information from low-level perceptual areas onto higher-level areas in prefrontal cortex (Farahibozorg et al., 2019; Hoeren et al., 2013). We find significant tuning shifts toward targets in AG and SMG, higher than that in occipitotemporal and premotor AON nodes, irrespective of the search target. Our results can be taken to suggest a higher place for AG and SMG in the hierarchy of semantic representations compared to remaining AON nodes. Another potential account could be that areas with stronger action selectivity might undergo stronger tuning shifts and future studies are warranted to investigate this issue more directly.

JNeurosci Accepted Manuscript

Cortical areas selective for an object category are suggested to retain their preferred tuning even when a non-preferred category is the search target (Çukur et al., 2013; Reddy and Kanwisher, 2007; Shahdloo et al., 2020). We find that semantic tuning of voxels in locomotion-action-selective superior parietal cortex are shifted toward locomotion actions only during search for this target. Likewise, semantic tuning of voxels in communication-selective anterior prefrontal cortex are shifted toward communication actions only during search for communication. These results suggest that semantic tuning shifts interact with the intrinsic selectivity for target categories.

We used WordNet to label action categories in the stimulus and create a one-hot-encoded stimulus feature matrix. Thus, it is possible to conduct part of the reported analyses by directly examining modulations of category responses. However, assessments in the 922-dimensional category space would treat each category independently ignoring semantic similarities, and they would be inherently noisier reducing our sensitivity for detecting tuning shifts. To assess TSI for non-target categories, category responses were masked to zero out responses for communication and locomotion actions. If selectivity measurements had been performed based on one-hot category features, this masking would eliminate all information related to target categories. It would then be impossible to quantify whether tuning for non-target categories shifts towards/away from the attended category. Therefore, we performed our analyses in a dense-encoded semantic space obtained via PCA. An alternative is voxel-wise modelling with a dense-encoded stimulus feature matrix derived using embedding models (Mikolov et al., 2013; Devlin et al., 2019). During preliminary experiments in the current study and prior studies from our lab (Huth et al, 2018; Çelik et al., 2021), we compared the category model against dense embedding models and estimates of attentional modulations did not vary significantly by choice of model. As such, we do not expect a profound difference between results from these various models, although there could be practical differences in terms of interpretation and feature similarity assessments.

We employed communication and locomotion as target categories to maximize our chances for detecting semantic tuning shifts, since previous studies suggest that these action categories have broadly distributed and distinctive representations (Urgen and Orban, 2021). Attentional modulations in multi-voxel response patterns were recently reported during search for several other categories related to animal taxonomy or actions (Nastase et al., 2017). We have observed in preliminary experiments that search for many salient categories in natural movies elicits tuning shifts (data not shown here). Thus, it is likely that tuning shifts are a ubiquitous mechanism for response modulation during natural visual search for action categories. However, there may be differences in the strength and cortical distribution of tuning shifts depending on the target action, and future studies are warranted to systematically examine whether and how tuning shifts generalise across action categories. Evidence suggests that attending to an object can modulate responses to features correlated with the target (O'Craven et al., 1999). We have previously reported that attending to a target object (e.g., vehicles) enhances the representation of semantically similar actions (e.g., driving) (Çukur et al., 2013). It is thus possible that attending to a target action could induce tuning shifts for correlated features such as the object categories pertaining to the actor. Since we restricted the target actions to be performed by the same animate actors, we did not examine tuning changes for objects in this study.

JNeurosci Accepted Manuscript

The tuning profile of a voxel refers to its response levels to the examined range of features. Attention can induce different modulations on this profile including baseline changes, gain changes and tuning shifts. Baseline changes imply an additive offset, gain changes imply a multiplicative offset to responses uniformly across features, neither changing the shape of the profile. Instead, tuning shifts alter shape by shifting selectivity towards the target, changing responses to both attended and unattended features. Here, we find that the overall tuning shift is attributed to significant tuning changes for both target and non-target categories. Such broadly distributed changes imply alteration in the shape of the tuning profile. Since our measurements are naturally limited by the spatiotemporal resolution of BOLD responses, we cannot make definitive inferences about the neural mechanisms underlying voxel tuning shifts, which could be attributed to baseline, gain or selectivity changes in single neurons (Connor et al., 1997; David et al., 2008; Reynolds et al., 2000). Further electrophysiological work would be needed to characterize neural tuning shifts during action-based search.

A common practice in fMRI is to collect a relatively limited dataset from a greater number of subjects to increase reliability of across-subject assessments at the expense of individual-subject results. Diverting away from this practice, here we collect a larger amount of data per subject to give greater focus to reliability in single subjects. This procedure substantially increased the amount and diversity of fMRI data collected per subject, which enhanced the quality of resulting models and thereby reliability of individual-subject results. However, we acknowledge that future studies are warranted to assess to what degree the results reported in the current study generalise to a broader population of subjects.

The natural movie stimuli used here have greater ecological validity compared to simplified or controlled movie clips used in many action-perception studies. That said, action categories in natural movies might be correlated with low-level features such as global motion-energy (Nishimoto et al., 2011; Weiss et al., 2006) and intermediate-level features such as scene dynamics (Grossman and Blake, 2002). Substantial correlations can confound the estimated category responses and tuning shifts. We employed several procedures to control for potential biases. First, to minimize correlations between category responses and global motion-energy, we used a nuisance motion-energy regressor (Nishimoto et al., 2011). Second, we restricted analyses to voxels uniquely predicted by the category model after accounting for motion-energy and STIP features. Voxels in areas such as LOC might encode multiple levels of features ranging from motion-energy and kinematics to semantics. Thus, controlling for motion-energy and STIP features might reduce sensitivity for attentional modulation of perceptual selectivity in these areas. Our analyses do not consider attentional tuning shifts that might be evident for motion-energy and STIP features, or other features such as expected action goals (Hudson et al., 2016a, 2016b), and actors' perceived attitude (Bach and Schenke, 2017). Some level of ambiguity will be naturally evident about what specific aspect of the correlated stimulus features are most relevant for measured cortical representations. Addressing this ambiguity requires complete decorrelation of all possible feature sets, yet conclusions derived using decorrelated stimuli deprived from their natural context might no longer be ecologically relevant. It remains important work to assess the effects of category-based search on multiple levels of feature representations.

JNeurosci Accepted Manuscript

In conclusion, we showed that natural visual search for a specific action category modulates semantic representations, causing tuning shifts toward the target in single voxels within and beyond the AON. Attentional modulations further interact with intrinsic selectivity of neural populations for search targets. This dynamic attentional mechanism can facilitate action perception by efficiently allocating neural resources to accentuate the representation of task-relevant action categories. Overall, these findings offer new insights into the effects of category-based visual search on brain responses (Çukur et al., 2013; Erez and Duncan, 2015; Harel et al., 2014; Peelen et al., 2009), as our results help explain humans' astounding ability to perceive others' actions in dynamic, cluttered daily-life experiences.

JNeurosci Accepted Manuscript

Data and software availability

Data supporting the findings of this study are available from the corresponding authors upon request. Results can be explored online via an interactive brain viewer at

http://www.icon.bilkent.edu.tr/brainviewer/shahdloo_etal/.

The codes used to estimate spatially informed voxelwise model weights is freely available on GitHub at https://github.com/icon-lab/SPIN-VM.

Acknowledgements

Funding

This work was supported in part by a Marie Curie Actions Career Integration Grant (PCIG13-GA-2013-618101), by a European Molecular Biology Organisation Installation Grant (IG 3028), by a TUBA GEBIP 2015 fellowship, and by a BAGEP 2017 fellowship.

Author contributions

Conceptualization, J.L.G and T.C.; Methodology, M.S., and T.C.; Software, M.S., and E.C.; Investigation, M.S., B.A.U. and T.C.; Writing-Original Draft, M.S.; Writing-Review and Editing, M.S., T.C., B.A.U., J.L.G, and E.C.; Funding Acquisition, T.C.; Resources, M.S., T.C., and J.L.G; Supervision, T.C.

The authors declare no competing financial interests.

JNeurosci Accepted Manuscript

References

Abdollahi, R.O., Jastorff, J., and Orban, G.A. (2012). Common and Segregated Processing of Observed Actions in Human SPL. Cerebral Cortex *23*, 2734–2753.

Anticevic, A., Repovs, G., and Barch, D.M. (2010). Resisting emotional interference: Brain regions facilitating working memory performance during negative distraction. Cognitive, Affective, & Behavioral Neuroscience *10*, 159–173.

Bach, P., and Schenke, K.C. (2017). Predictive social perception: Towards a unifying framework from action observation to person knowledge. Social and Personality Psychology Compass *11*, *e12312*.

Battelli, L., Cavanagh, P., and Thornton, I.M. (2003). Perception of biological motion in parietal patients. Neuropsychologia *41*, 1808–1816.

Bedny, M., and Caramazza, A. (2011). Perception, action, and word meanings in the human brain: The case from action verbs. Annals of the New York Academy of Sciences *1224*, 81–95.

Bremmer, F., Schlack, A., Duhamel, J.-R., Graf, W., and Fink, G.R. (2001). Space Coding in Primate Posterior Parietal Cortex. NeuroImage *14*, S46–S51.

Buccino, G., Binkofski, F., Fink, G.R., Fadiga, L., Fogassi, L., Gallese, V., Seitz, R.J., Zilles, K., Rizzolatti, G., and Freund, H.J. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: An fMRI study. European Journal of Neuroscience *13*, 400–404.

Caspers, S., Zilles, K., Laird, A.R., and Eickhoff, S.B. (2010). ALE meta-analysis of action observation and imitation in the human brain. NeuroImage *50*, 1148–1167.

Cavina-Pratesi, C., Connolly, J.D., Monaco, S., Figley, T.D., Milner, D., Schenk, T., Culham, J.C. (2018). Human neuroimaging reveals the subcomponents of grasping, reaching and pointing actions. Cortex *98*, 128-148.

Çelik, E., Dar, S.U.H., Yılmaz, Ö., Keleş, Ü., and Çukur, T. (2019). Spatially informed voxelwise modeling for naturalistic fMRI experiments. NeuroImage *186*, 741–757.

Connor, C.E., Preddie, D.C., Gallant, J.L., and Van Essen, D.C. (1997). Spatial Attention Effects in Macaque Area V4. Journal of Neuroscience *17*, 3201–3214.

Corbetta, M., and Shulman, G.L. (2002). Control of goal-directed and stimulus-driven attention in the brain. Nature Reviews Neuroscience *3*, 215–229.

Corbo, D., and Orban, G.A. (2017). Observing others speak or sing activates SPT and neighboring parietal cortex. Journal of Cognitive Neuroscience *29*, 1002–1021.

Çukur, T., Nishimoto, S., Huth, A.G., and Gallant, J.L. (2013). Attention during natural vision warps semantic representation across the human brain. Nature Neuroscience *16*, 763–770.

JNeurosci Accepted Manuscript

888 Dalal, N., and Triggs, B. (2005). Histograms of oriented gradients for human detection. In IEEE
889 Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), IEEE,
890 886–893.

891 van Dam, W.O., Rueschemeyer, S.-A., and Bekkering, H. (2010). How specifically are action
892 verbs represented in the neural motor system: An fMRI study. NeuroImage *53*, 1318–1325.

893 David, S.V., Hayden, B.Y., Mazer, J.A., and Gallant, J.L. (2008). Attention to stimulus features
894 shifts spectral tuning of V4 neurons during natural vision. Neuron *59*, 509–521.

895 Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep
896 Bidirectional Transformers for Language Understanding. ArXiv:1810.04805.

897 Erez, Y., and Duncan, J. (2015). Discrimination of Visual Categories Based on Behavioral
898 Relevance in Widespread Regions of Frontoparietal Cortex. Journal of Neuroscience *35*, 12383–
899 12393.

900 Farahibozorg, S.-R., Henson, R.N., Woollams, A.M., and Hauk, O. (2019). Distinct roles for the
901 Anterior Temporal Lobe and Angular Gyrus in the spatio-temporal cortical semantic network.
902 BioRxiv, 544114.

903 Ferri, S., Rizzolatti, G., and Orban, G.A. (2015). The organisation of the posterior parietal cortex
904 devoted to upper limb actions: An fMRI study. Human Brain Mapping *36*, 3845–3866.

905 Grafton, S.T., and de C Hamilton, A.F. (2007). Evidence for a distributed hierarchy of action
906 representation in the brain. Human Movement Science *26*, 590–616.

907 Griffiths, T. L., Steyvers, M., and Tenenbaum, J. B. (2007). Topics in semantic representation.
908 Psychological Review, 114(2), 211–244.

909 Grosbras, M.H., Beaton, S., and Eickhoff, S.B. (2012). Brain regions involved in human
910 movement perception: A quantitative voxel-based meta-analysis. Human Brain Mapping *33*,
911 431–454.

912 Grossman, E.D., and Blake, R. (2002). Brain areas active during visual perception of biological
913 motion. Neuron *35*, 1167–1175.

914 Hamilton, A.F. de C., and Grafton, S.T. (2006). Goal representation in human anterior
915 intraparietal sulcus. Journal of Neuroscience *26*, 1133–1137.

916 Handjaras, G., Bernardi, G., Benuzzi, F., Nichelli, P.F., Pietrini, P., and Ricciardi, E. (2015). A
917 topographical organisation for action representation in the human brain. Human Brain Mapping
918 *36*, 3832–3844.

919 Harel, A., Kravitz, D.J., and Baker, C.I. (2014). Task context impacts visual object processing
920 differentially across the cortex. PNAS *111*, E962-71.

JNeurosci Accepted Manuscript

Harris, C., and Stephens, M. (1988). A combined corner and edge detector. In In Proc. Fourth Alvey Vision Conference, 147–152.

Herrington, J., Nymberg, C., Faja, S., Price, E., and Schultz, R. (2012). The responsiveness of biological motion processing areas to selective attention towards goals. NeuroImage *63*, 581–590.

Hoeren, M., Kaller, C.P., Glauche, V., Vry, M.-S., Rijntjes, M., Hamzei, F., and Weiller, C. (2013). Action semantics and movement characteristics engage distinct processing streams during the observation of tool use. Experimental Brain Research *229*, 243–260.

Holte, M.B., Moeslund, T.B., and Fihl, P. (2010). View-invariant gesture recognition using 3D optical flow and harmonic motion context. Computer Vision and Image Understanding *114*, 1353–1361.

Hudson, M., Nicholson, T., Ellis, R., and Bach, P. (2016a). I see what you say: Prior knowledge of other's goals automatically biases the perception of their actions. Cognition *146*, 245–250.

Hudson, M., Nicholson, T., Simpson, W.A., Ellis, R., and Bach, P. (2016b). One step ahead: The perceived kinematics of others' actions are biased toward expected goals. Journal of Experimental Psychology: General *145*, 1–7.

Huth, A.G., Nishimoto, S., Vu, A.T., and Gallant, J.L. (2012). A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. Neuron *76*, 1210–1224.

Ilg, U.J., Schumann, S., and Thier, P. (2004). Posterior parietal cortex neurons encode target motion in world-centered coordinates. Neuron *43*, 145–151.

Isik, L., Koldewyn, K., Beeler, D., and Kanwisher, N.G. (2017). Perceiving social interactions in the posterior superior temporal sulcus. PNAS *114*, E9145–E9152.

Jastorff, J., and Orban, G.A. (2009). Human Functional Magnetic Resonance Imaging Reveals Separation and Integration of Shape and Motion Cues in Biological Motion Processing. Journal of Neuroscience *29*, 7315–7329.

Jastorff, J., Begliomini, C., Fabbri-Destro, M., Rizzolatti, G., and Orban, G.A. (2010). Coding observed motor acts: Different organisational principles in the parietal and premotor cortex of humans. Journal of Neurophysiology *104*, 128–140.

Jastorff, J., Abdollahi, R.O., Fasano, F., and Orban, G.A. (2016). Seeing biological actions in 3D: An fMRI study. Human Brain Mapping *37*, 203–219.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. Perception & Psychophysics *14*, 201–211.

Jurafsky, D. and Martin, J. (2021). Speech and language processing. 3rd ed., Prentice Hall.

JNeurosci Accepted Manuscript

955 Kilintari, M., Raos, V., and Savaki, H.E. (2014). Involvement of the Superior Temporal Cortex
956 in Action Execution and Action Observation. Journal of Neuroscience *34*, 8999–9011.

957 Kiremitçi, I., Yilmaz, Ö., Çelik, E., Shahdloo, M., Huth, A.G., and Çukur, T. (2021). Attentional
958 Modulation of Hierarchical Speech Representations in a Multi-Talker Environment. Cerebral
959 Cortex, bhab136

960 de Lange, F.P., Spronk, M., Willems, R.M., Toni, I., and Bekkering, H. (2008). Complementary
961 Systems for Understanding Action Intentions. Current Biology *18*, 454–457.

962 Laptev, I. (2005). On space-time interest points. International Journal of Computer Vision, 107–
963 123.

964 Laptev, I., Marszałek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human
965 actions from movies. 26th IEEE Conference on Computer Vision and Pattern Recognition
966 (CVPR), 1-8.

967 Lescroart, M.D., and Gallant, J.L. (2019). Human Scene-Selective Areas Represent 3D
968 Configurations of Surfaces. Neuron *101*, 178-192.e7.

969 Liljeström, M., Tarkiainen, A., Parviainen, T., Kujala, J., Numminen, J., Hiltunen, J., Laine, M.,
970 and Salmelin, R. (2008). Perceiving and naming actions and objects. NeuroImage *41*, 1132–
971 1141.

972 Lingnau, A., and Downing, P. E. (2015). The lateral occipitotemporal cortex in action. *Trends in*
973 *cognitive sciences*, *19*(5), 268-277.

974 Lui, F., Buccino, G., Duzzi, D., Benuzzi, F., Crisi, G., Baraldi, P., Nichelli, P., Porro, C.A., and
975 Rizzolatti, G. (2008). Neural substrates for observing and imagining non-object-directed actions.
976 Social Neuroscience *3*, 261–275.

977 Mars, R.B., and Grol, M.J. (2007). Dorsolateral prefrontal cortex, working memory, and
978 prospective coding for action. Journal of Neuroscience *27*, 1801–1802.

979 Miller, G.A. (1995). WordNet: a lexical database for English. Communications of ACM *38*, 39–
980 41.

981 Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word
982 Representations in Vector Space. ArXiv:1301.3781

983 Molinari, E., Baraldi, P., Campanella, M., Duzzi, D., Nocetti, L., Pagnoni, G., and Porro, C.A.
984 (2013). Human parietofrontal networks related to action observation detected at rest. Cerebral
985 Cortex *23*, 178–186.

986 Muthukumaraswamy, S.D., and Singh, K.D. (2008). Modulation of the human mirror neuron
987 system during cognitive activity. Psychophysiology *45*, 896–905.

JNeurosci Accepted Manuscript

988 Muthukumaraswamy, S.D., Johnson, B.W., and McNair, N.A. (2004). Mu rhythm modulation
989 during observation of an object-directed grasp. Cognitive Brain Research *19*, 195–201.

990 Nastase, S.A., Connolly, A.C., Oosterhof, N.N., Halchenko, Y.O., Guntupalli, J.S., Di Oleggio
991 Castello, M.V., Gors, J., Gobbini, M.I., and Haxby, J.V. (2017). Attention selectively reshapes
992 the geometry of distributed semantic representation. Cerebral Cortex *27*, 4277–4291.

993 Nastase, S.A., Halchenko, Y.O., Connolly, A.C., Gobbini, M.I., and Haxby, J.V. (2018). Neural
994 responses to naturalistic clips of behaving animals in two different task contexts. Frontiers in
995 Neuroscience *12*, 316.

996 Nelissen, K., Vanduffel, W., and Orban, G.A. (2006). Charting the Lower Superior Temporal
997 Region, a New Motion-Sensitive Region in Monkey Superior Temporal Sulcus. Journal of
998 Neuroscience *26*, 5929–5947.

999 Nelissen, K., Borra, E., Gerbella, M., Rozzi, S., Luppino, G., Vanduffel, W., Rizzolatti, G., and
1000 Orban, G.A. (2011). Action observation circuits in the macaque monkey cortex. Journal of
1001 Neuroscience *31*, 3743–3756.

1002 Newman-Norlund, R., van Schie, H.T., van Hoek, M.E.C., Cuijpers, R.H., and Bekkering, H.
1003 (2010). The role of inferior frontal and parietal areas in differentiating meaningful and
1004 meaningless object-directed actions. Brain Research *1315*, 63–74.

1005 Nicholson, T., Roser, M., and Bach, P. (2017). Understanding the Goals of Everyday
1006 Instrumental Actions Is Primarily Linked to Object, Not Motor-Kinematic, Information:
1007 Evidence from fMRI. PLoS ONE *12*, e0169700.

1008 Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J.L. (2011).
1009 Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. Current
1010 Biology *21*, 1641–1646.

1011 Noppeney, U. (2008). The neural systems of tool and action semantics: A perspective from
1012 functional imaging. Journal of Physiology *102*, 40–49.

1013 Nunez-Elizalde, A. O., Huth, A. G., and Gallant, J. L. (2019). Voxelwise encoding models with
1014 non-spherical multivariate normal priors. NeuroImage, *197*:482–492.

1015

1016 Oberman, L.M., Pineda, J.A., and Ramachandran, V.S. (2007). The human mirror neuron
1017 system: A link between action observation and social skills. Social Cognitive and Affective
1018 Neuroscience *2*, 62–66.

1019 Oosterhof, N.N., Wiggett, A.J., Diedrichsen, J., Tipper, S.P., and Downing, P.E. (2010). Surface-
1020 based information mapping reveals crossmodal vision-action representations in human parietal
1021 and occipitotemporal cortex. Journal of Neurophysiology *104*, 1077–1089.

JNeurosci Accepted Manuscript

1022 Oosterhof, N.N., Tipper, S.P., and Downing, P.E. (2012). Viewpoint (in)dependence of action
1023 representations: an MVPA study. Journal of Cognitive Neuroscience *24*, 975–989.

1024 Oosterhof, N.N., Tipper, S.P., and Downing, P.E. (2013). Crossmodal and action-specific:
1025 Neuroimaging the human mirror neuron system. Trends in Cognitive Sciences *17*, 311–318.

1026 Paneri, S., and Gregoriou, G.G. (2017). Top-down control of visual attention by the prefrontal
1027 cortex. Functional specialization and long-range interactions. Frontiers in Neuroscience *11*, 545.

1028 Peelen, M.V., Fei-Fei, L., and Kastner, S. (2009). Neural mechanisms of rapid natural scene
1029 categorization in human visual cortex. Nature *460*, 94–97.

1030 Pizzamiglio, L., Aprile, T., Spitoni, G., Pitzalis, S., Bates, E., D'Amico, S., and Di Russo, F.
1031 (2005). Separate neural systems for processing action- or non-action-related sounds. NeuroImage
1032 *24*, 852–861.

1033 Popham, S. F., Huth, A. G., Bilenko, N. Y., Deniz, F., Gao, J. S., Nunez-Elizalde, A. O., and
1034 Gallant, J. L. (2021). Visual and linguistic semantic representations are aligned at the border of
1035 human visual cortex. Nature Neuroscience, 24(11), 1628–1636.

1036 Ptak, R., Schnider, A., and Fellrath, J. (2017). The Dorsal Frontoparietal Network: A Core
1037 System for Emulated Action. Trends in Cognitive Sciences *21*, 589–599.

1038 Puglisi, G., Leonetti, A., Landau, A., Fornia, L., Cerri, G., and Borroni, P. (2017). The role of
1039 attention in human motor resonance. PLoS ONE *12*, e0177457.

1040 Puglisi, G., Leonetti, A., Cerri, G., and Borroni, P. (2018). Attention and cognitive load modulate
1041 motor resonance during action observation. Brain & Cognition *128*, 7–16.

1042 Ramsey, R., and Hamilton, A.F.D.C. (2010). Understanding actors and object-goals in the
1043 human brain. NeuroImage *50*, 1142–1147.

1044 Reddy, L., and Kanwisher, N.G. (2007). Category Selectivity in the Ventral Visual Pathway
1045 Confers Robustness to Clutter and Diverted Attention. Current Biology *17*, 2067–2072.

1046 Reynolds, J.H., Pasternak, T., and Desimone, R. (2000). Attention Increases Sensitivity of V4
1047 Neurons. Neuron *26*, 703–714.

1048 Rizzolatti, G., and Matelli, M. (2003). Two different streams form the dorsal visual system:
1049 Anatomy and functions. Experimental Brain Research *153*, 146–157.

1050 Rizzolatti, G., Fogassi, L., and Gallese, V. (1997). Parietal cortex: from sight to action. Current
1051 Opinion in Neurobiology *7*, 562–567.

1052 Rowe, J., Friston, K., Frackowiak, R., and Passingham, R. (2002). Attention to action: Specific
1053 modulation of corticocortical interactions in humans. NeuroImage *17*, 988–998.

JNeurosci Accepted Manuscript

1054 Rozzi, S., and Fogassi, L. (2017). Neural coding for action execution and action observation in
1055 the prefrontal cortex and its role in the organisation of socially driven behavior. Frontiers in
1056 Neuroscience *11*, 276.

1057 Safford, A.S., Hussey, E.A., Parasuraman, R., and Thompson, J.C. (2010). Object-based
1058 attentional modulation of biological motion processing: spatiotemporal dynamics using
1059 functional magnetic resonance imaging and electroencephalography. Journal of Neuroscience *30*,
1060 9064–9073.

1061 Schuch, S., Bayliss, A.P., Klein, C., and Tipper, S.P. (2010). Attention modulates motor system
1062 activation during action observation: Evidence for inhibitory rebound. Experimental Brain
1063 Research *205*, 235–249.

1064 Seidl, K.N., Peelen, M.V., and Kastner, S. (2012). Neural evidence for distracter suppression
1065 during visual search in real-world scenes. Journal of Neuroscience *32*, 11812–11819.

1066 Shahdloo, M., Çelik, E., and Çukur, T. (2020). Biased competition in semantic representation
1067 during natural visual search. NeuroImage *216*, 116383.

1068 Tarhan, L., and Konkle, T. (2020). Sociality and interaction envelope organise visual action
1069 representations. Nature Communications *11*, 3002.

1070 Thompson, J., and Parasuraman, R. (2012). Attention, biological motion, and action recognition.
1071 NeuroImage *59*, 4–13.

1072 Toepper, M., Gebhardt, H., Beblo, T., Thomas, C., Driessen, M., Bischoff, M., Blecker, C.R.,
1073 Vaitl, D., and Sammer, G. (2010). Functional correlates of distractor suppression during spatial
1074 working memory encoding. Neuroscience *165*, 1244–1253.

1075 Urgen, B.A., and Orban, G.A. (2021). The unique role of parietal cortex in action observation:
1076 Functional organisation for communicative and manipulative actions. NeuroImage *273*, 118220.

1077 Urgen, B.A., Pehlivan, S., and Saygin, A.P. (2019). Distinct representations in occipito-temporal,
1078 parietal, and premotor cortex during action perception revealed by fMRI and computational
1079 modeling. Neuropsychologia *127*, 35–47.

1080 Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. Human Brain
1081 Mapping *30*, 829–858.

1082 Walbrin, J., and Koldewyn, K. (2019). Dyadic interaction processing in the posterior temporal
1083 cortex. NeuroImage *198*, 296–302.

1084 Walbrin, J., Downing, P., and Koldewyn, K. (2018). Neural responses to visually observed social
1085 interactions. Neuropsychologia *112*, 31–39.

1086 Weiss, P.H., Rahbari, N.N., Lux, S., Pietrzyk, U., Noth, J., and Fink, G.R. (2006). Processing the
1087 spatial configuration of complex actions involves right posterior parietal cortex: An fMRI study
1088 with clinical implications. Human Brain Mapping *27*, 1004–1014.

JNeurosci Accepted Manuscript

Wilson-Mendenhall, C.D., Simmons, W.K., Martin, A., and Barsalou, L.W. (2013). Contextual processing of abstract concepts reveals neural representations of nonlinguistic semantic content. Journal of Cognitive Neuroscience *25*, 920–935.

Wurm, M.F., and Caramazza, A. (2019). Lateral occipitotemporal cortex encodes perceptual components of social actions rather than abstract representations of sociality. NeuroImage *202*, 116153.

Wurm, M.F., Caramazza, A., and Lingnau, A. (2017). Action Categories in Lateral Occipitotemporal Cortex Are Organised Along Sociality and Transitivity. Journal of Neuroscience *37*, 562–575.

Zhou, X., Katsuki, F., Qi, X.-L., and Constantinidis, C. (2012). Neurons with inverted tuning during the delay periods of working memory tasks in the dorsal prefrontal and posterior parietal cortex. J Neurophysiology *108*, 31–38.

Figure Captions

Figure 1. Hypothesised changes in semantic representation of action categories. Recent evidence suggests that the human brain organises hundreds of object and action categories in a semantic space that is distributed systematically across the cerebral cortex (Huth et al., 2012). **a.** Semantic representation for a single subject from Çukur et al. (2013) is shown on flattened cortical surface and on inflated hemispheres. Colours indicate tuning for different object or action categories (see colour legend). Regions of interest identified using conventional functional localizers are denoted by white borders. Abbreviations for regions of interest are listed in *Materials and Methods*. **b.** In the semantic space, action categories that are semantically similar to each other are mapped to nearby points and semantically dissimilar actions are mapped to distant points. There is evidence that visual search for object categories warps semantic representation in favour of the targets by shifting single-voxel tuning for object categories toward target objects (Çukur et al., 2013). Thus, we hypothesised that visual search for a given action category should similarly expand the semantic representation of the target and semantically similar categories.

Figure 2. Model fitting and validation procedure. Undergoing fMRI, human subjects viewed 60mins of natural movies and covertly searched for *communication* or *locomotion* action categories while fixating on a central dot. **a.** An indicator matrix was constructed that identified the presence of each of the 922 object and action categories in each 1-sec clip of the movies (see Fig. 2-1). Nuisance regressors were included to account for head-motion, physiological noise, and eye-movement confounds. An additional nuisance regressor was included to account for target detection confounds. In a nested cross-validation (CV) procedure, regularized linear regression was used to estimate separate category model weights (i.e., category responses) for each search task that mapped each category feature to the recorded BOLD responses in single voxels. **b.** Accuracy of the fit models was cross-validated by measuring prediction performance on the held-out data in each CV fold, after discarding the nuisance regressors and the target regressor. Prediction score of the fit models was taken as product-moment correlation coefficient between estimated and measured BOLD responses, averaged across the two search tasks.

Figure 3. Prediction performance of the category model. To test the performance of fit category models, prediction score was calculated on held-out data as the product-moment correlation coefficient between the predicted category responses and measured BOLD responses, and it was averaged across the two search tasks. **a.**

JNeurosci Accepted Manuscript

Prediction scores of the category model are plotted on flattened cortical surfaces of individual subjects. A variance partitioning analysis was used to quantify the response variance that was uniquely predicted by the category model after accounting for low- and intermediate-level stimulus features (see *Materials and Methods*, Fig. 4). Voxels where the category model did not explain unique response variance after accounting for these features were masked (bootstrap test, q(FDR)<0.05; see Fig. 11). **b.** To visualise single-subject results in a common space, prediction score values are shown following projection onto the standard brain template from Freesurfer and averaging across subjects, after getting thresholded in single subjects. Only voxels that were identified as semantic in all individual subjects were averaged and displayed in the template. Regions of interest are illustrated by white borders. Several important sulci are illustrated by dashed grey lines. Abbreviations for regions of interest and sulci are listed in *Materials and Methods*. The category model predicts responses across ventral-temporal, parietal, and frontal cortices well, suggesting that visual categories are broadly represented across visual and nonvisual cortex. Results can be explored via an interactive brain viewer at http://www.icon.bilkent.edu.tr/brainviewer/shahdloo_etal/.

Figure 4. Comparison of category and control models. The prediction scores (raw product-moment correlation coefficient) of the category and control (the collection of motion-energy and STIP regressors) models were measured for all cortical voxels. Voxels across all subjects are displayed. Each voxel is represented with a dot. Red versus blue dots indicate whether the category model or the control model yields higher prediction scores. Black dots indicate voxels where none of the models has high prediction scores. The category model outperforms the control model in 53.75±3.29% of cortical voxels (mean±sem; average over five subjects).

Figure 5. Fraction of uniquely predicted voxels in regions of interest (ROIs). We identified voxels in which the category model explained unique response variance after accounting for low-level motion-energy, and intermediate-level STIP stimulus features by performing a variance partitioning analysis (see *Materials and Methods*). Fraction of these *semantic voxels* is shown across ROIs, in individual subjects. Asterisk shows across subject significance (bootstrap test, $q(FDR)$<0.05).

Figure 6. Attention warps semantic representation of action categories. To assess attentional changes, we projected voxel-wise tuning profiles onto a continuous semantic space. **a.** The semantic space was derived from principal components analysis (PCA) of tuning vectors measured during a separate passive-viewing task, and was tested to be consistent across subjects (Fig. 7). To illustrate the semantic information embedded within this space, action categories were projected onto PC1 and PC3 that best delineate the target actions (Fig. 8; words in regular font show projections of individual categories, see Fig. 9). To illustrate the semantic content of the PCs, characteristic actions of the movie stimulus were clustered in the semantic space, and cluster centres were projected onto the PCs after getting labelled (bold-italic words; see *Materials and Methods*, Fig. 10). Average location of the *communication* and *locomotion* actions are specified with red and green dots. **b.** Action category responses during passive viewing and during the two search tasks were projected onto the semantic space, and a two-dimensional colourmap was used to colour each voxel based on the projection values along PC1 and PC3 (see legend). Projections in individual subjects were mapped onto the standard brain template from Freesurfer, and average projections across subjects are displayed (see Figs. 6-1 to 6-5 for data in individual subjects). Figure formatting is identical to Fig. 3. Many voxels across occipitotemporal, parietal, and prefrontal cortices shift their tuning toward targets, suggesting that attention warps semantic representations of actions. Specifically, voxels in inferior posterior parietal cortex, cingulate cortex, and anterior inferior prefrontal cortex shift their tuning toward *communication* during search for *communication* actions. Meanwhile, voxels in superior posterior and medial parietal cortex shift their tuning toward *locomotion* during search for *locomotion* actions. Results can be explored via an interactive brain viewer at http://www.icon.bilkent.edu.tr/brainviewer/shahdloo_etal/.

Figure 7. Consistency of the semantic space across subjects. To test whether the estimated semantic space is consistent across subjects, leave-one-out cross-validation was performed. In each cross-validation fold, best-predicted voxels from four subjects were used to derive 12 PCs to construct a semantic space. In the left-out subject, semantic tuning profile for each voxel was obtained by projecting action category responses during passive viewing onto the derived PCs. Next, product-moment correlation coefficient was calculated between the tuning profiles in the derived space and the tuning profiles in the original semantic space. Results were averaged across semantic voxels in the left-out subject. Correlation coefficients are shown for each PC and each subject. The cross-validated semantic spaces consistently correlate with the original semantic space.

JNeurosci Accepted Manuscript

Figure 8. The distance between target actions in subspaces spanned by different pairs of PCs. To visualise attentional modulation of semantic representation in Fig. 6, we compared projections of action category responses onto a pair of PCs across the search tasks. To maximize our sensitivity in visualising the attentional modulations, we chose the pair of dimensions that maximally separates the actions belonging to the two target categories (i.e., *communication* and *locomotion* categories). The Mahalanobis distance between communication actions and locomotion actions (mean±sem across communication and locomotion actions) in the subspace spanned by each pair of PCs is shown. Target actions are maximally separated across the subspace spanned by the first and third PCs.

Figure 9. Distribution of action categories across PCs. To illustrate the distribution of action categories embedded within the semantic space, action categories were projected onto the PCs. Projections onto the first three PCs are shown (words in regular font show projections of individual categories). To facilitate illustration, categories were collapsed into 10 clusters and cluster centres were also projected onto the PCs (bold-italic words; see *Materials and Methods*). Average location of the *communication* and *locomotion* actions are specified with red and green dots. The estimated semantic space captures reasonable semantic variance across action categories in natural movies.

Figure 10. Projections of action category clusters onto PCs. Each of the 109 action categories were projected onto the twelve semantic principal components (PCs). The projections were then clustered into 10 groups using k-means and labelled for interpretation (see *Materials and Methods*). The projections of the cluster centres onto 12 PCs are shown. The first three dimensions were used to visualise the semantic space. The first dimension distinguishes between self-movements (e.g., swirl, consume) and actions that are targeted toward other humans or objects (e.g., reach, talk). The second dimension distinguishes between dynamic (e.g., drive, chase) versus static actions (e.g., consume, struggle). The third dimension distinguishes between actions that involve humans (e.g., talk, reach) and dynamic actions (e.g., fly, swirl).

Figure 11. Cortical distribution of tuning shifts. a. To quantify the tuning shifts for the attended versus unattended categories, a tuning shift index ($TSI_{all} \in [-1,1]$) was calculated for each voxel. Tuning shifts toward the attended category would yield positive TSI (red colour), whereas negative TSI would indicate shifts away from the attended category (blue colour). TSI_{all} values from individual subjects were projected onto the standard brain template and averaged across subjects (see Figs. 11-1a to 11-5a for data in individual subjects). Figure formatting is identical to Fig. 3. AON is outlined by green dashed lines. Voxels across many cortical regions shifted their tuning toward the attended category. These include regions across AON (occipitotemporal cortex, posterior parietal cortex, and premotor cortex), lateral prefrontal cortex, and anterior cingulate cortex. **b.** To examine how representation of nontarget action categories changes during visual search, we measured a separate tuning shift index specifically for these categories (TSI_{nt}). TSI_{nt} values from individual subjects were projected onto the standard brain template and averaged across subjects (see Figs. 11-1b to 11-5b for data in individual subjects). TSI_{nt} shows a similar distribution to TSI_{all} shown in **a**, albeit with lower magnitude (Fig. 12). Tuning shift for nontarget categories is positive across many voxels within posterior parietal cortex and anterior prefrontal cortex, suggesting a more flexible semantic representation of actions in these cortices, compared to occipitotemporal AON nodes. Results can be explored via an interactive brain viewer at http://www.icon.bilkent.edu.tr/brainviewer/shahdloo_etal/.

Figure 12. Difference in tuning shift for target, versus non-target categories. The difference between absolute values of TSI_{all} and TSI_{nt} were calculated in individual ROIs. TSI_{all} is significantly larger than TSI_{nt} in all areas with significant tuning shift.

Figure 13. Fraction of the overall tuning shifts. Fraction of the overall tuning shifts explained by shifts in tuning for target categories (mean±sem across subjects) and nontarget categories (i.e., excluding the union of communication and locomotion categories) is shown. Target categories explain a greater portion of the overall

JNeurosci Accepted Manuscript

tuning shifts broadly across ROIs, except for early retinotopic areas. At the same time, nontarget categories significantly contribute to the overall tuning shifts.

Figure 14. Interaction of tuning shifts with intrinsic selectivity for individual targets. To examine the interaction between tuning shifts and the intrinsic selectivity for individual targets, separate target preference indices (PI) were calculated during search for *communication* (PI_{com}), and *locomotion* (PI_{loc}) categories. PI during search for a specific target action was taken as the difference in selectivity for the target versus distractor during search for that target. PI_{com} and PI_{loc} values are shown following projection onto the standard brain template (see Figs. 11-1c to 11-5c for data in individual subjects). A two-dimensional colourmap was used to annotate each voxel based on PI_{com} and PI_{loc} values (see legend). Figure format is identical to Fig. 3. AON is outlined by green dashed lines. Semantic tuning in voxels across posterior parietal and anterior prefrontal cortices shift toward the attended category irrespective of the search target. However, tuning in many voxels in anterior parietal, occipital, and cingulate cortices shift toward the attended category only during search for communication or only during search for locomotion actions.

Figure 15. Attentional tuning changes in regions of interest. Average **(a)** TSI_{all}, **(b)** TSI_{nt}, **(c)** PI_{com}, and **(d)** PI_{loc} values were examined in cortical areas (mean±sem across five subjects). Significant values are denoted by green bars and grey bars denote non-significant values (bootstrap test, $q(FDR)>0.05$). Values for individual subjects are indicated by dots. Grey dots show values in areas with non-significant mean, green dots show non-significant values in areas with significant mean, and green crosses show significant values in areas with significant mean. Tuning shift is significantly greater than zero in many regions across all levels of the AON including occipitotemporal cortex (pSTS, pMTG), posterior parietal cortex (IPS, AG, SMG), and premotor cortex (BA44, BA45), and in regions across prefrontal and cingulate cortices (SFG, ACC). Compared to occipitotemporal areas, attention more diversely modulates semantic representations in parietal and premotor AON nodes, manifested as significantly positive tuning shift for nontarget categories in posterior parietal cortex (AG, SMG) and anterior inferior frontal cortex (BA45). PI_{com} is significantly greater than zero in BA44/45, SFG, and ACC. In contrast, PI_{loc} is significantly greater than zero in IPS and AG and is significantly less than zero in dPMC. Both PI_{com} and PI_{loc} are significantly greater than zero in pSTS, pMTG, SMG, and MFG. Tuning shifts interact with the attention task, and with intrinsic selectivity of cortical areas for target action categories.

Figure 2-1. Stimulus action categories. Actions belonging to the two target categories, and non-target categories are specified. Occurrence frequency of each action, calculated as the number of movie frames where a given action was present, is indicated in parentheses.

Figure 6-1. Cortical flat maps of semantic representation for subject S1. Action category responses during **a.** passive viewing, **b.** search for *communication*, and **c.** search for *locomotion* categories were projected onto the semantic space in subject S1. A two-dimensional colourmap was used to colour each voxel based on the projection values along the first and third semantic dimensions (see colour legend). Voxels where the category model does not explain unique response variance after accounting for low- and intermediate-level stimulus features are masked (bootstrap test, $q(FDR)<0.05$). Regions of interest are illustrated by white borders. Several important sulci are illustrated by dashed grey lines. Abbreviations for regions of interest and sulci are listed in *Materials and Methods*. Many voxels across occipitotemporal, parietal, and prefrontal cortices shift their tuning toward targets.

Figure 6-2. Cortical flat maps of semantic representation for subject S2. Action category responses during **a.** passive viewing, **b.** search for *communication*, and **c.** search for *locomotion* categories were projected onto the semantic space in subject S2. Formatting is identical to Fig. 6-1.

Figure 6-3. Cortical flat maps of semantic representation for subject S3. Action category responses during **a.** passive viewing, **b.** search for *communication*, and **c.** search for *locomotion* categories were projected onto the semantic space in subject S3. Formatting is identical to Fig. 6-1.

JNeurosci Accepted Manuscript

Figure 6-4. Cortical flat maps of semantic representation for subject S4. Action category responses during **a.** passive viewing, **b.** search for *communication*, and **c.** search for *locomotion* categories were projected onto the semantic space in subject S4. Formatting is identical to Fig. 6-1.

Figure 6-5. Cortical flat maps of semantic representation for subject S5. Action category responses during **a.** passive viewing, **b.** search for *communication*, and **c.** search for *locomotion* categories were projected onto the semantic space in subject S5. Formatting is identical to Fig. 6-1.

Figure 11-1. Cortical flat maps of TSI and PI for subject S1. a. Tuning shift index for all action categories (TSI_{all}), **b.** tuning shift for nontarget categories (TSI_{nt}), and **c.** preference index values (PI_{com}, PI_{loc}) were projected onto the semantic space in subject S1 (see legends in Fig. 11 and 14 for the colour map). Only significant voxels are shown (bootstrap test, $q(FDR)<0.05$). Formatting is identical to Fig. 6-1.

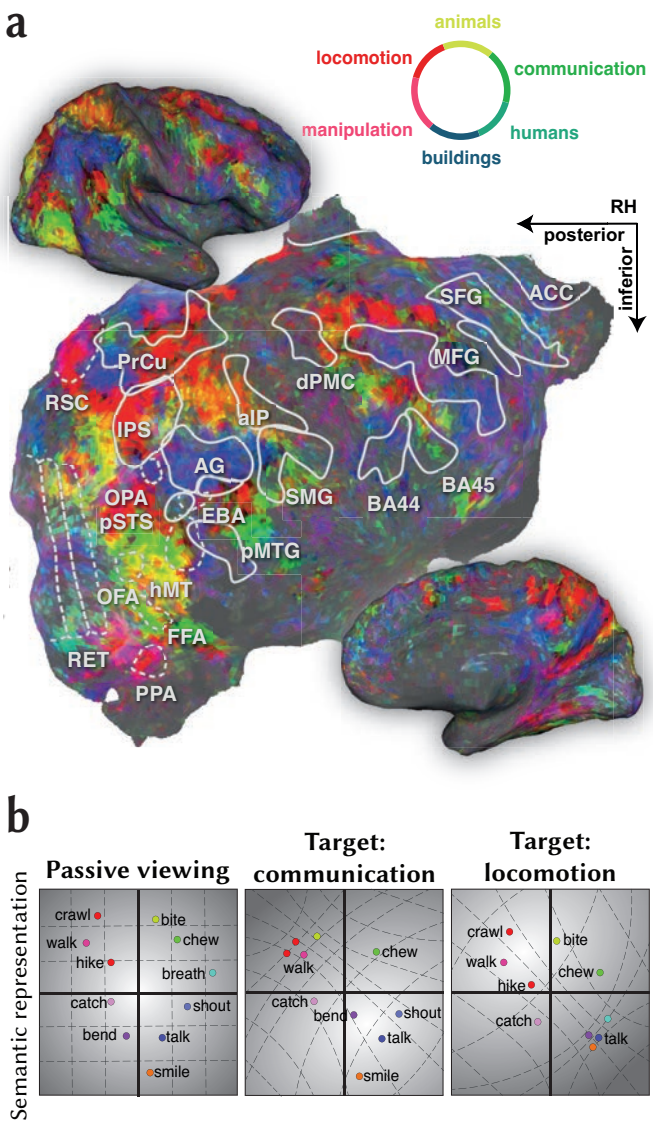
Figure 11-2. Cortical flat maps of TSI and PI for subject S2. a. Tuning shift index for all action categories (TSI_{all}), **b.** tuning shift for nontarget categories (TSI_{nt}), and **c.** preference index values (PI_{com}, PI_{loc}) were projected onto the semantic space in subject S2 (see legends in Fig. 11 and 14 for the colour map). Only significant voxels are shown (bootstrap test, $q(FDR)<0.05$). Formatting is identical to Fig. 6-1.

Figure 11-3. Cortical flat maps of TSI and PI for subject S3. a. Tuning shift index for all action categories (TSI_{all}), **b.** tuning shift for nontarget categories (TSI_{nt}), and **c.** preference index values (PI_{com}, PI_{loc}) were projected onto the semantic space in subject S3 (see legends in Fig. 11 and 14 for the colour map). Only significant voxels are shown (bootstrap test, $q(FDR)<0.05$). Formatting is identical to Fig. 6-1.

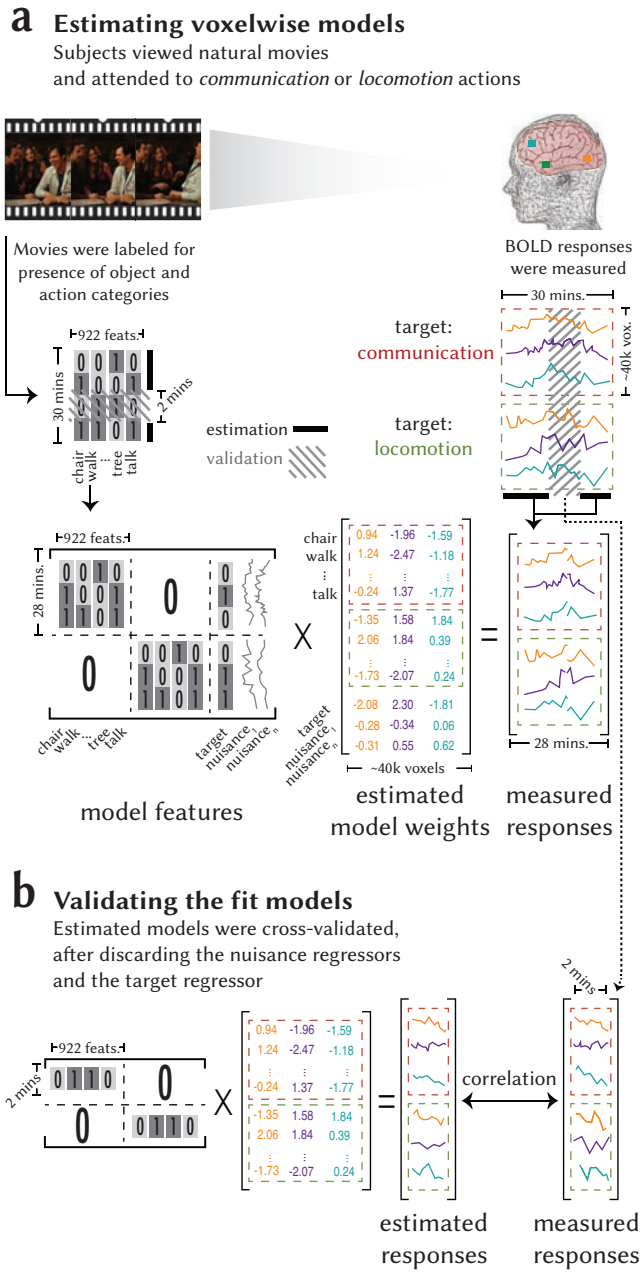
Figure 11-4. Cortical flat maps of TSI and PI for subject S4. a. Tuning shift index for all action categories (TSI_{all}), **b.** tuning shift for nontarget categories (TSI_{nt}), and **c.** preference index values (PI_{com}, PI_{loc}) were projected onto the semantic space in subject S4 (see legends in Fig. 11 and 14 for the colour map). Only significant voxels are shown (bootstrap test, $q(FDR)<0.05$). Formatting is identical to Fig. 6-1.

Figure 11-5. Cortical flat maps of TSI and PI for subject S5. a. Tuning shift index for all action categories (TSI_{all}), **b.** tuning shift for nontarget categories (TSI_{nt}), and **c.** preference index values (PI_{com}, PI_{loc}) were projected onto the semantic space in subject S5 (see legends in Fig. 11 and 14 the colour map). Only significant voxels are shown (bootstrap test, $q(FDR)<0.05$). Formatting is identical to Fig. 6-1.

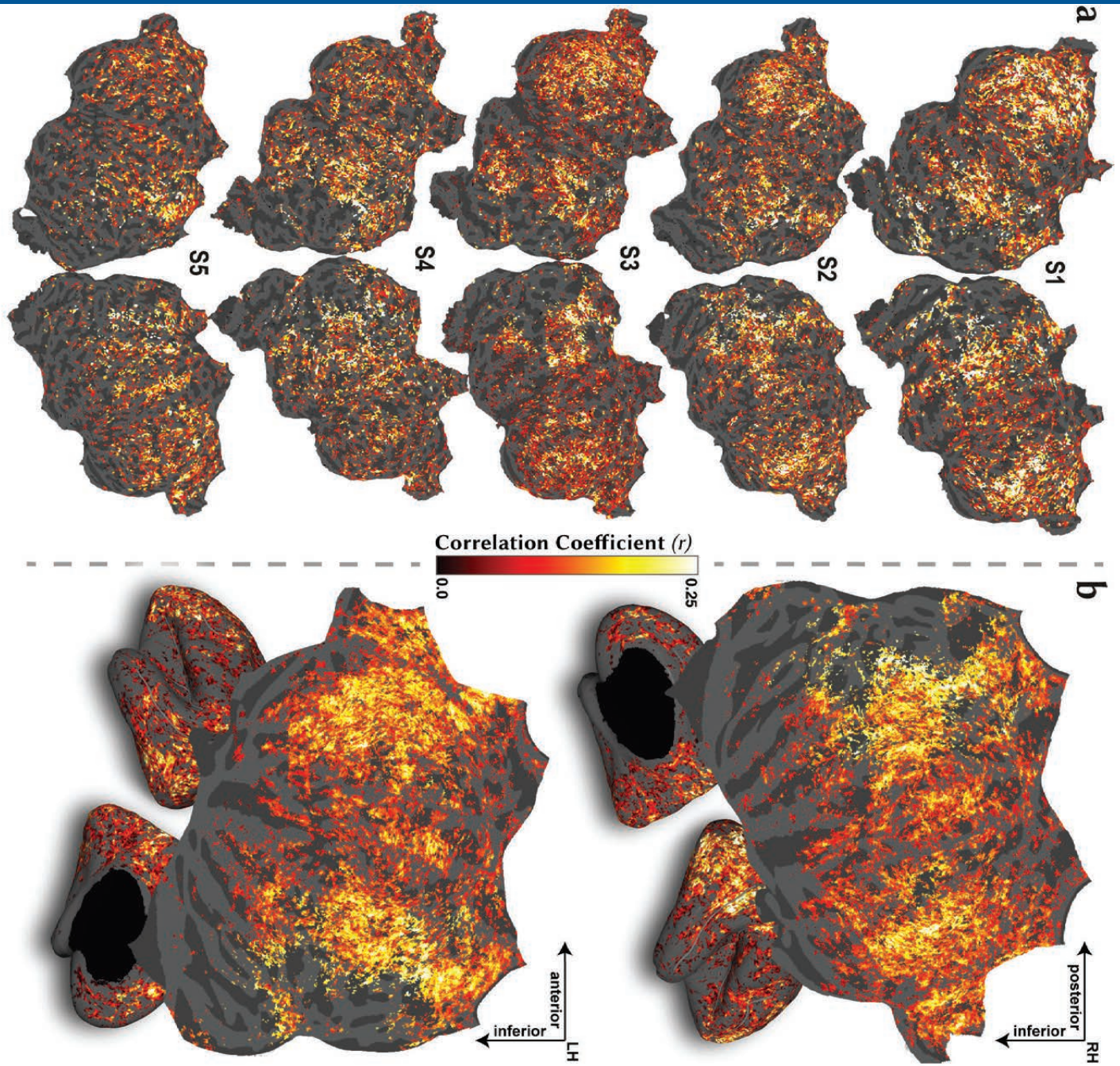
JNeurosci Accepted Manuscript



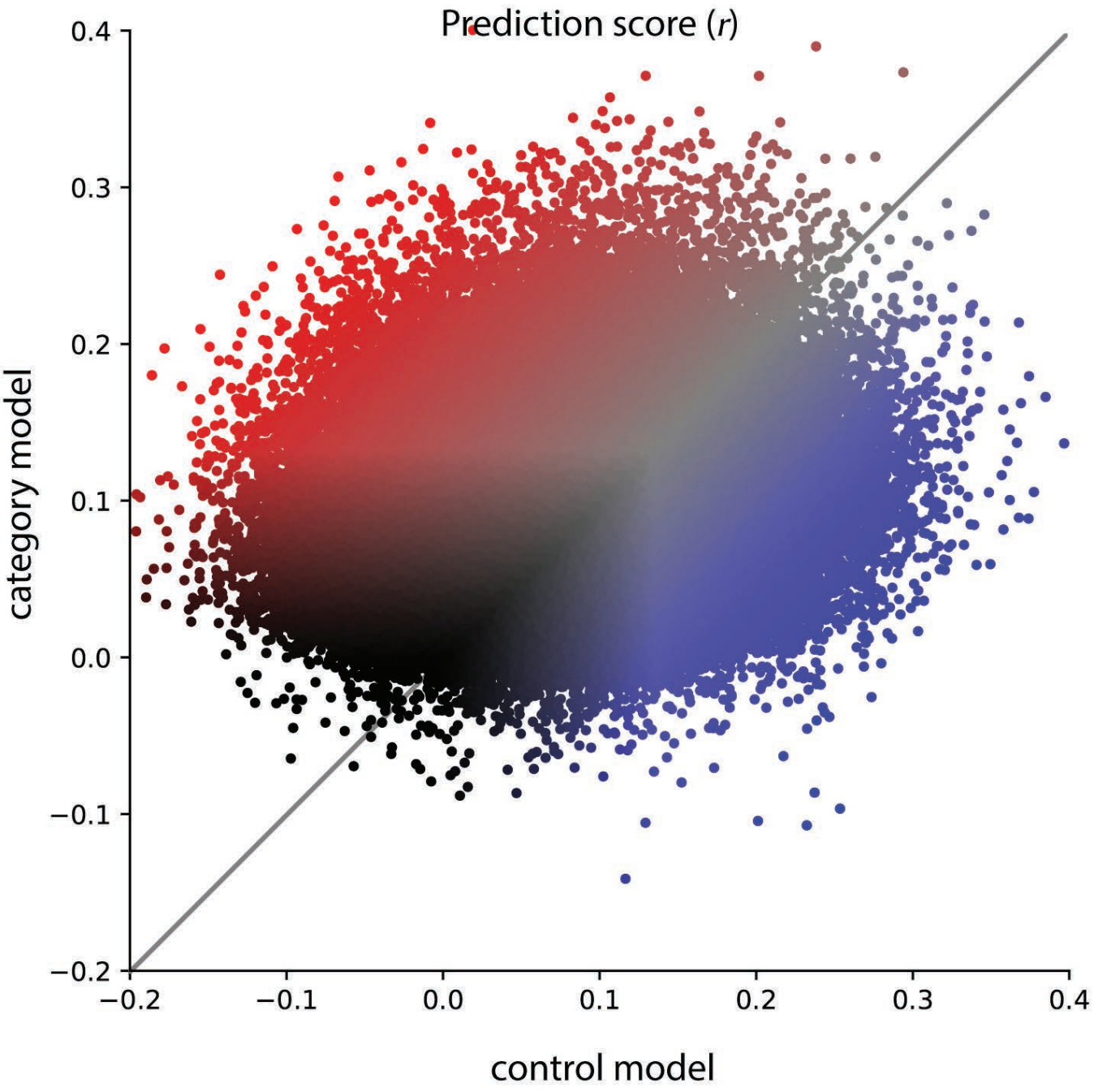
JNeurosci Accepted Manuscript



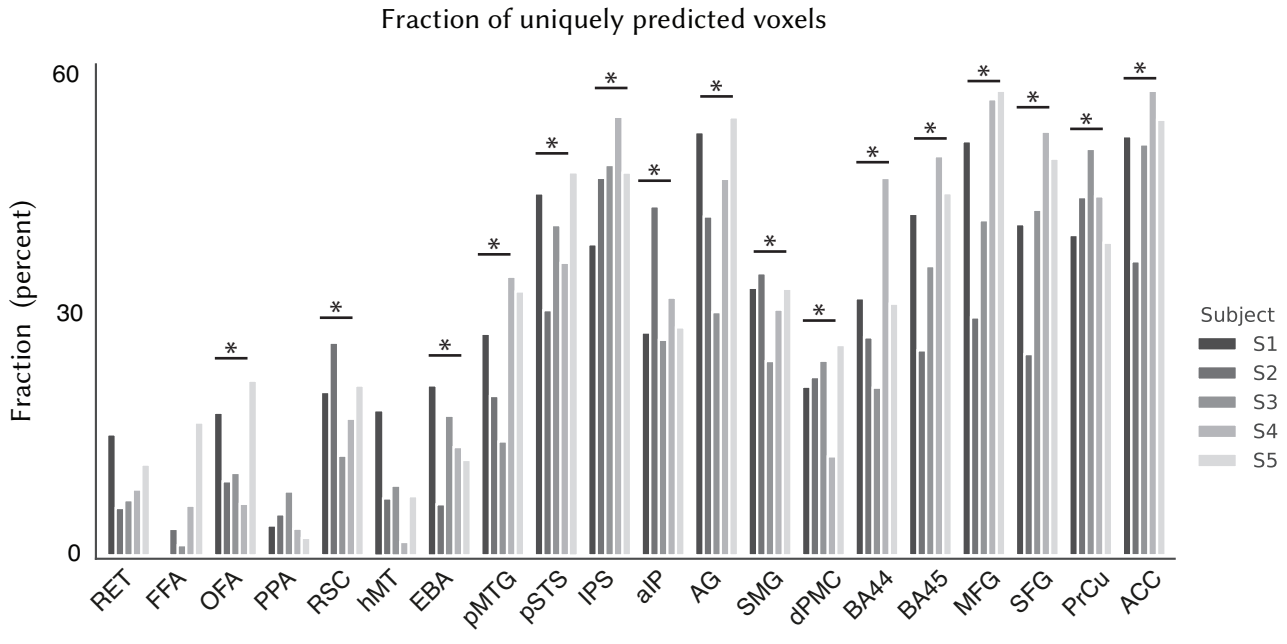
JNeurosci Accepted Manuscript



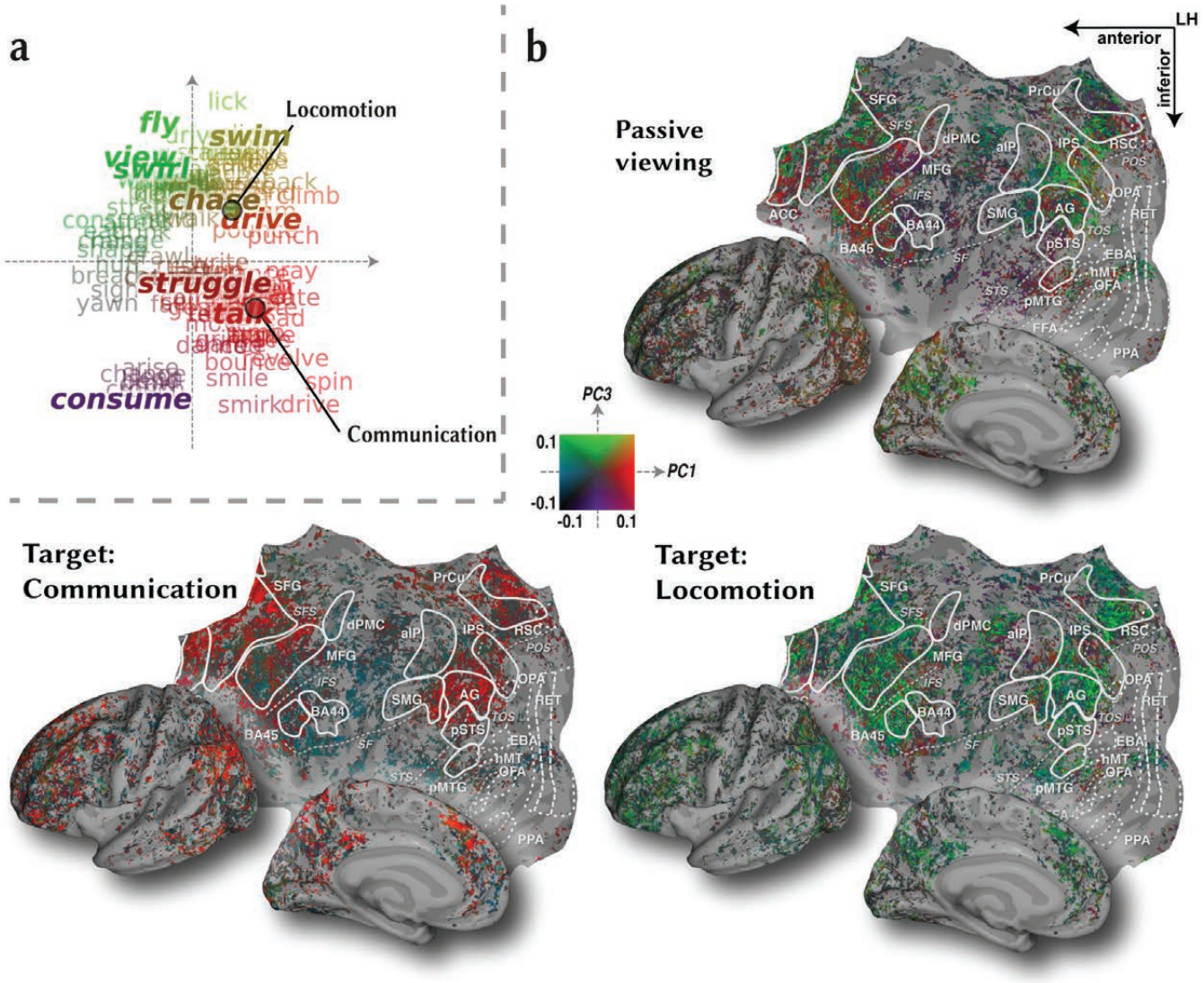
JNeurosci Accepted Manuscript



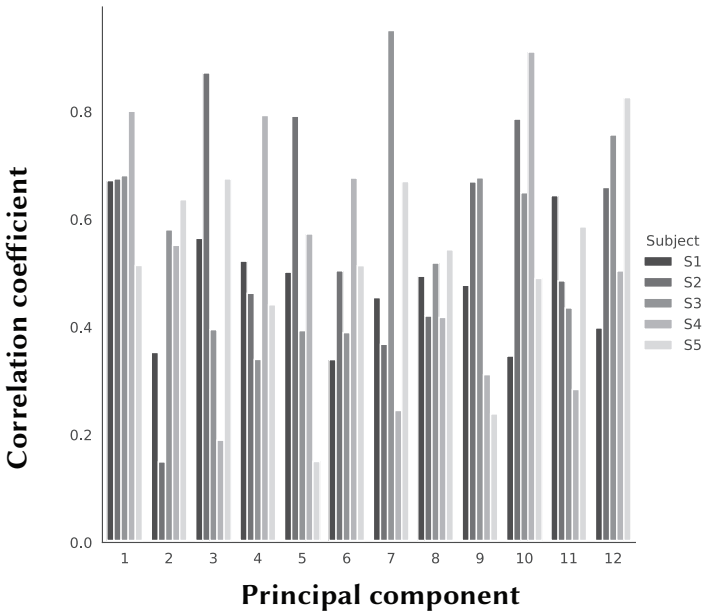
JNeurosci Accepted Manuscript



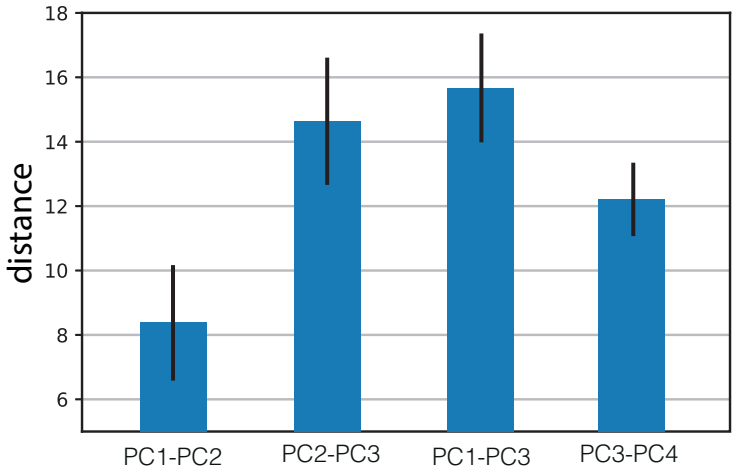
JNeurosci Accepted Manuscript



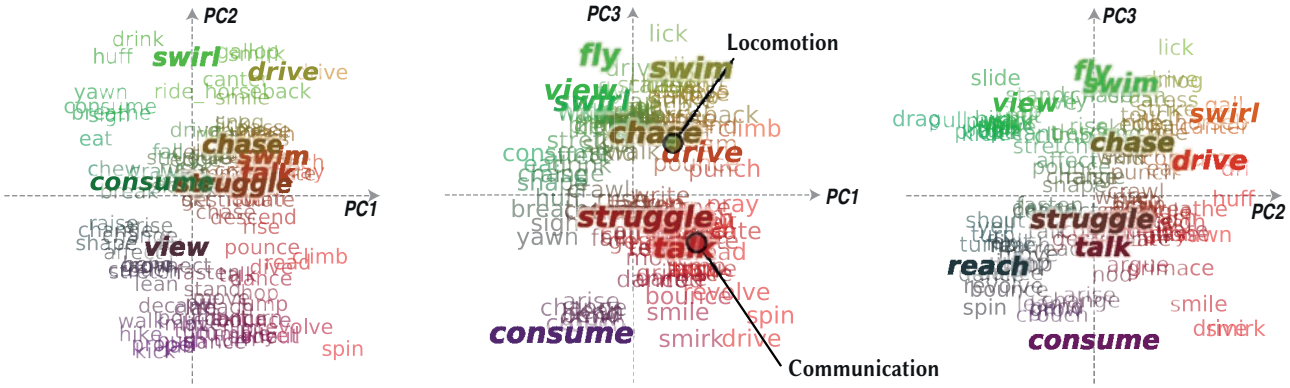
JNeurosci Accepted Manuscript



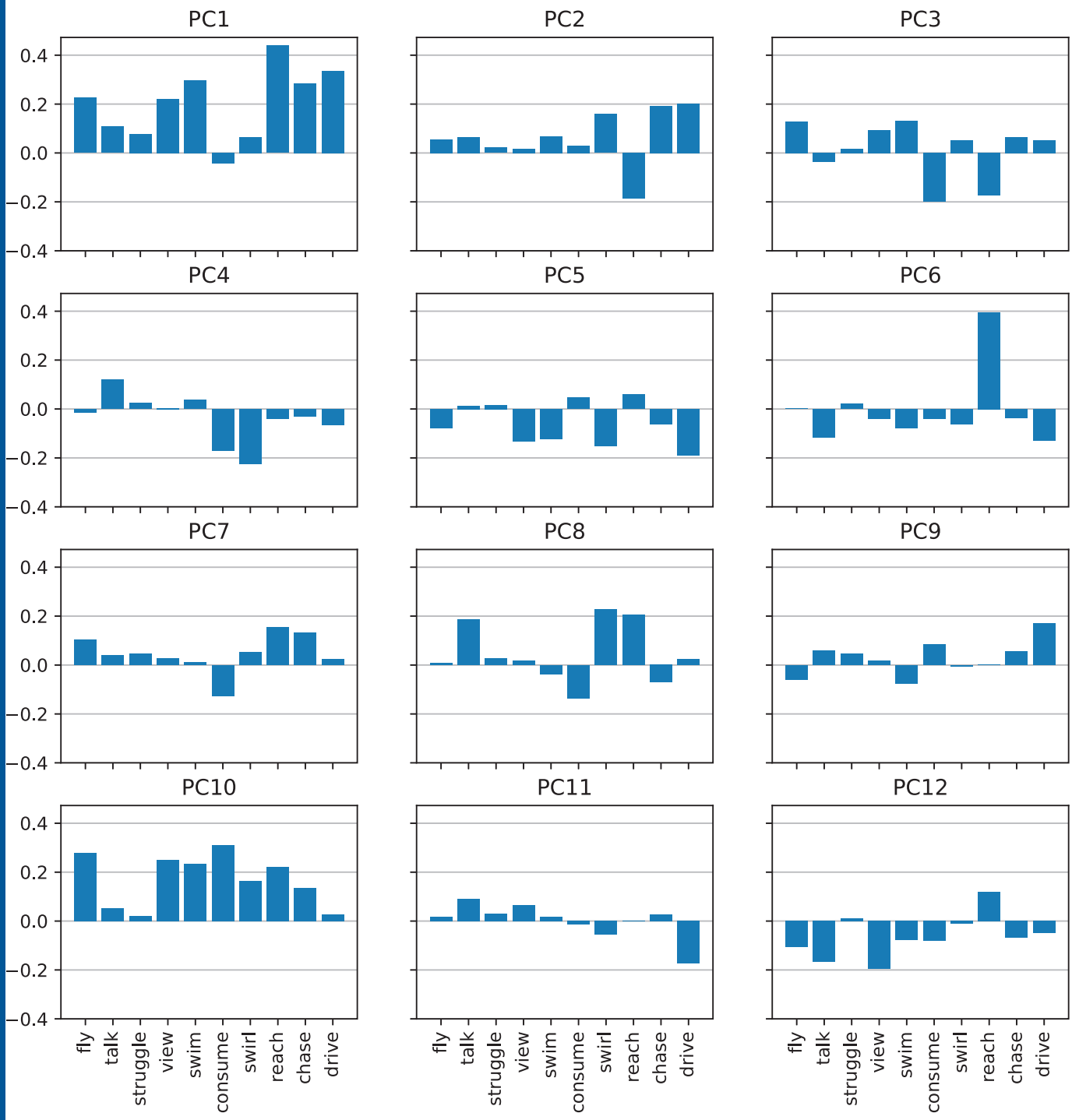
JNeurosci Accepted Manuscript



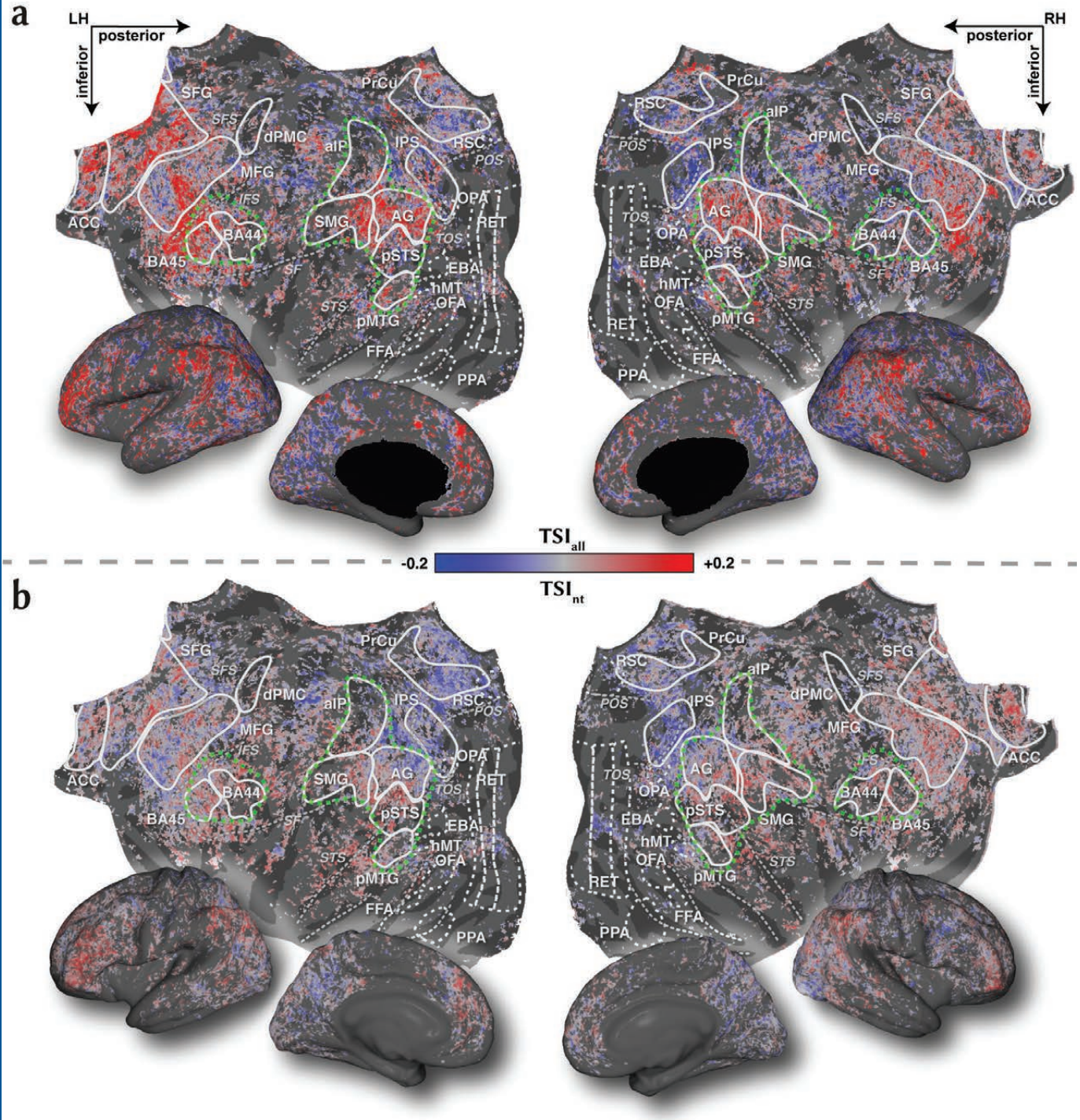
JNeurosci Accepted Manuscript



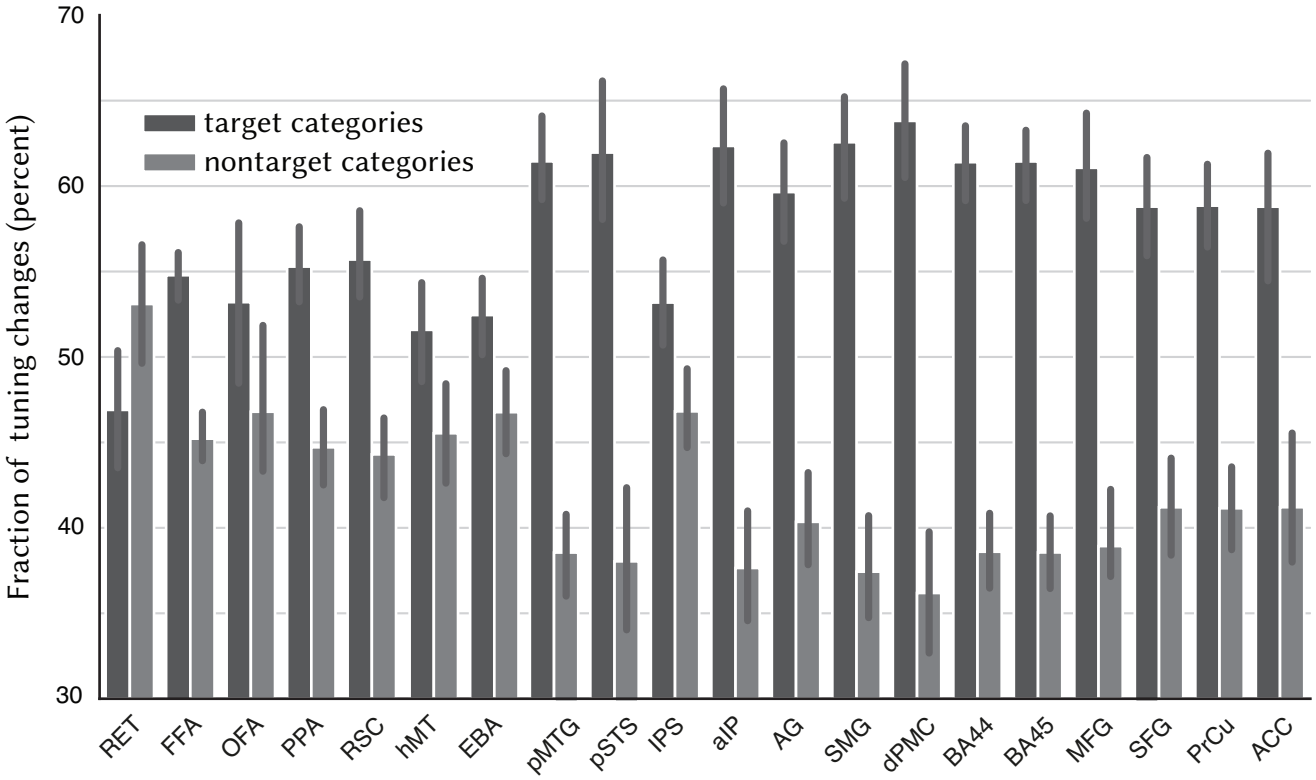
JNeurosci Accepted Manuscript



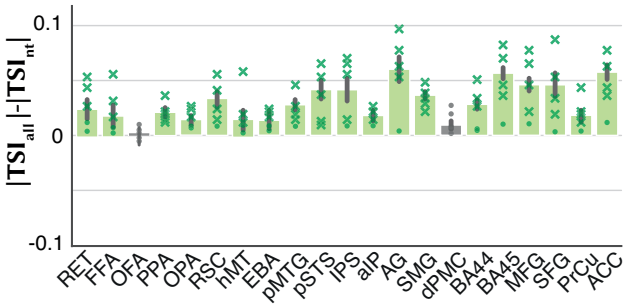
JNeurosci Accepted Manuscript



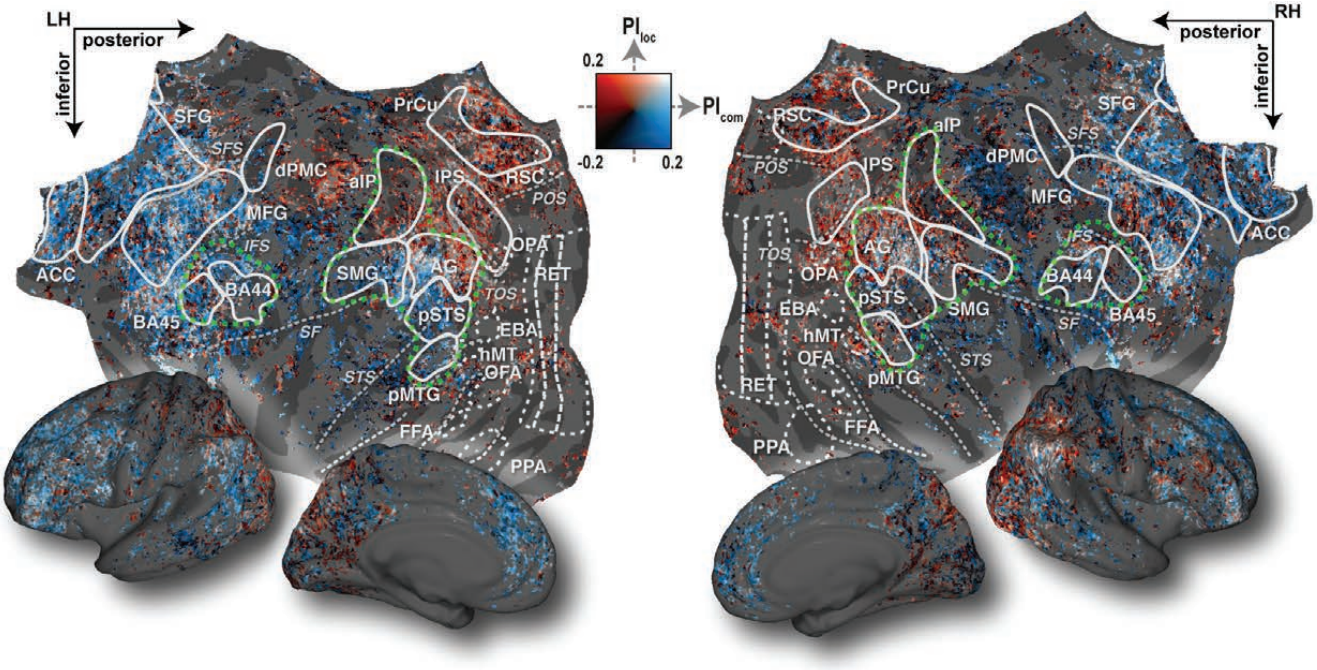
JNeurosci Accepted Manuscript



JNeurosci Accepted Manuscript



JNeurosci Accepted Manuscript



JNeurosci Accepted Manuscript