Social-affective features drive human representations of observed actions

Diana C. Dima[1*], Tyler M. Tomita[2], Christopher J. Honey[2], Leyla Isik[1]

[1]Dept. of Cognitive Science, Johns Hopkins University, United States

[2]Dept. of Psychological and Brain Sciences, Johns Hopkins University, United States

Abstract

Humans observe actions performed by others in many different visual and social settings. What features do we extract and attend when we view such complex scenes, and how are they processed in the brain? To answer these questions, we curated two large-scale sets of naturalistic videos of everyday actions and estimated their perceived similarity in two behavioral experiments. We normed and quantified a large range of visual, action-related and social-affective features across the stimulus sets. Using a cross-validated variance partitioning analysis, we found that social-affective features predicted similarity judgments better than, and independently of, visual and action features in both behavioral experiments. Next, we conducted an electroencephalography (EEG) experiment, which revealed a sustained correlation between neural responses to videos and their behavioral similarity. Visual, action, and social-affective features predicted neural patterns at early, intermediate and late stages respectively during this behaviorally relevant time window. Together, these findings show that social-affective features are important for perceiving naturalistic actions, and are extracted at the final stage of a temporal gradient in the brain.

Introduction

In daily life, we rely on our ability to recognize a range of actions performed by others in a variety of different contexts. Our perception of others' actions is both efficient and flexible, enabling us to rapidly understand new actions no matter where they occur or who is performing them. This understanding plays a part in complex social computations about the mental states and intentions of others (Jamali et al., 2021; Spunt et al., 2011; Thornton et al., 2019; Thornton and Tamir, 2021; Weaverdyck et al., 2021). Visual action recognition also interacts cross-modally with language-based action understanding (Bedny and Caramazza, 2011; Humphreys et al., 2013). However, there are two important gaps in our understanding of action perception in realistic settings. First, we still don't know which features of the visual world underlie our representations of observed actions. Second, we do not know how different types of action-relevant features, ranging from visual to social, are processed in the brain, and especially how they unfold over time. Answering these questions can shed light on the computational mechanisms that support action perception. For example, are different semantic and social features extracted in parallel or sequentially?

Relatively few studies have investigated the temporal dynamics of neural responses to actions. During action observation, a distributed network of brain areas extracts action-related features ranging from visual to abstract, with viewpoint-invariant responses emerging as early as 200 ms (Isik et al., 2018). Visual features include the spatial scale of an action (i.e. fine-scale manipulations like knitting versus full-body movements like running) represented throughout visual cortex (Tarhan and Konkle, 2020), and information about biological motion, thought to be extracted within 200 ms in superior temporal cortex (Giese and Poggio, 2003; Hirai et al., 2003;

Hirai and Hiraki, 2006; Johansson, 1973; Jokisch et al., 2005; Vangeneugden et al., 2014). Responses in occipito-temporal areas have been shown to reflect semantic features like invariant action category (Hafri et al., 2017; Lingnau and Downing, 2015; Tucciarelli et al., 2019, 2015; Wurm and Caramazza, 2019; Wurm and Lingnau, 2015), as well as social features like the number of agents and sociality of actions (Tarhan and Konkle, 2020; Wurm et al., 2017; Wurm and Caramazza, 2019).

Among the visual, semantic, and social features thought to be processed during action observation, it is unclear which underlie our everyday perception in naturalistic settings. Mounting evidence suggests that naturalistic datasets are key to improving ecological validity and reliability in visual and social neuroscience (Haxby et al., 2020; Nastase et al., 2020; Redcay and Moraczewski, 2020). Most action recognition studies to date have used controlled images and videos showing actions in simple contexts (Isik et al., 2018; Wurm and Caramazza, 2019). However, presenting actions in natural contexts is critical, as stimulus-context interactions have been shown to modulate neural activity (Willems and Peelen, 2021). Recent attempts to understand naturalistic action perception, however, have yielded mixed results, particularly with regard to the role of social features. For example, one recent study concluded that sociality (i.e., presence of a social interaction) was the primary organizing dimension of action representations in the human brain (Tarhan and Konkle, 2020). Another, however, found that semantic action category explained the most variance in fMRI data, with little contribution from social features (Tucciarelli et al., 2019).

Here, we combined a new large-scale dataset of everyday actions with a priori feature labels to comprehensively sample the hypothesis space defined by previous work. This is essential in light of the conflicting results from previous studies, as it allowed us to disentangle the contributions

of distinct but correlated feature spaces. We used three-second videos of everyday actions from the "Moments in Time" dataset (Monfort et al., 2019) and replicated our results across two different stimulus sets. Action videos were sampled from different categories based on the American Time Use Survey (ATUS, 2019) and were highly diverse, depicting a variety of contexts and people. We quantified a wide range of visual, action-related, and social-affective features in the videos and, through careful curation, ensured that they were minimally confounded across our dataset.

We used this dataset to probe the behavioral and neural representational space of human action perception. To understand the features that support natural action viewing, we predicted behavioral similarity judgments using the visual, action-related, and social-affective feature sets. Next, to investigate the neural dynamics of action perception, we recorded electroencephalography (EEG) data while participants viewed the stimuli, and we used the three sets of features to predict time-resolved neural patterns.

We found that social-affective features predict action similarity judgments better than, and independently of, visual and action-related features. Visual and action-related features explained less variance in behavior, even though they included fundamental features such as the scene setting and the semantic category of each action. Neural patterns revealed that behaviorally-relevant features are automatically extracted by the brain in a progression from visual to action to social-affective features. Together, our results reveal the importance of social-affective features in how we represent other people's actions, and show that these representations emerge in the brain along a temporal gradient.

Results

Disentangling visual, action, and social-affective features in natural videos

We curated two sets of naturalistic three-second videos of everyday actions from the Moments in Time dataset (Monfort et al., 2019). The videos were selected from a larger set, ensuring that features of interest were minimally correlated. 18 common activities based on the National Bureau of Labor Statistics' American Time Use Survey (ATUS, 2019) were represented (Table 1; see Materials and methods, section *Behavior: Stimuli*). The two stimulus sets contained 152 videos (8 videos per activity and 8 additional videos with no agents, included to add variation in the dataset, see Materials and methods, section *Behavior: Stimuli*) and 65 videos (3-4 videos per activity) respectively. The second set was used to replicate behavioral results in a separate experiment with different stimuli and participants.

Naturalistic videos of actions can vary along numerous axes, including visual features (e.g. the setting in which the action takes place or objects in the scene), action-specific features (e.g. semantic action category), and social-affective features (e.g. the number of agents involved or perceived arousal). For example, an action like 'eating' may vary in terms of context (in the kitchen vs at a park), object (eating an apple vs a sandwich), and number of agents (eating alone vs together). Drawing these distinctions is crucial to disambiguate between context, actions, and agents in natural events. To evaluate these different axes, we quantified 17 visual, action-related and social-affective features using image properties, labels assigned by experimenters, and behavioral ratings collected in online experiments (Figure 1a). Visual features ranged from low-level (e.g. pixel values) to high-level features related to scenes and objects (e.g. activations from the final layer of a pretrained neural network). Action-related features included transitivity

(object-relatedness), activity (the amount of activity in a video), effectors (body parts involved), and action category based on the American Time Use Survey (ATUS, 2019). Finally, social-affective features included sociality, valence, arousal and number of agents (see Materials and methods, section *Representational similarity analysis*). Representational dissimilarity matrices (RDM) were created for each feature by calculating pairwise Euclidean distances between all videos.

In both video sets there were only weak correlations between visual features and the higher-level action/social-affective features (Figure 1a). The highest correlations were those within each of the three sets of features, including visual features (Exp 1: Conv1 and image saturation/gist, τ_A=0.29; Exp 2: Conv1 and image hue, τ_A=0.32), action features (Exp 1: arousal and activity, τ_A=0.31; Exp 2: activity and effectors, τ_A=0.33) and social features (sociality and number of agents; Exp 1: τ_A=0.31, Exp 2: τ_A=0.3).

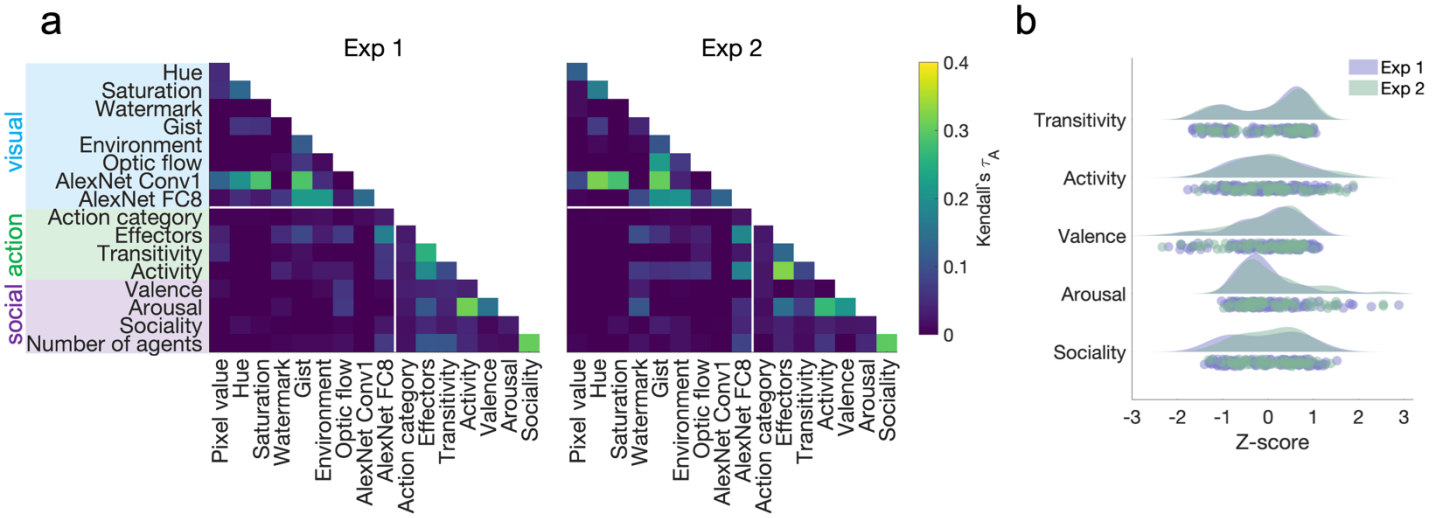

Figure 1. Quantifying visual, social-affective and action features in the two stimulus sets. **a.** Correlations between feature RDMs. Note the low correlations between visual features and action/social-affective features (white rectangle). **b.** Behavioral rating distributions in the two stimulus sets. The z-scored ratings were visualized as raincloud plots showing the individual data points, as well as probability density estimates computed using Matlab's *ksdensity* function (Allen et al., 2019).

The distributions of action and social-affective features were not significantly different between the two stimulus sets (all Mann-Whitney $z<1.08$, $P>0.28$). The width of these distributions suggests that the stimuli spanned a wide range along each feature (Figure 1b). In both experiments, transitivity was notable through its bimodal distribution, likely reflecting the presence or absence of objects in scenes, while other features had largely unimodal distributions.

Behaviorally rated features differed in reliability in Experiment 1 ($F(4,819) = 22.35$, $P<0.001$), with sociality being the most reliable and arousal the least reliable (Figure 3 – figure supplement 1). In Experiment 2, however, there was no difference in reliability ($F(4,619) = 0.76$, $P=0.55$). Differences in reliability were mitigated by our use of feature averages to generate feature RDMs.

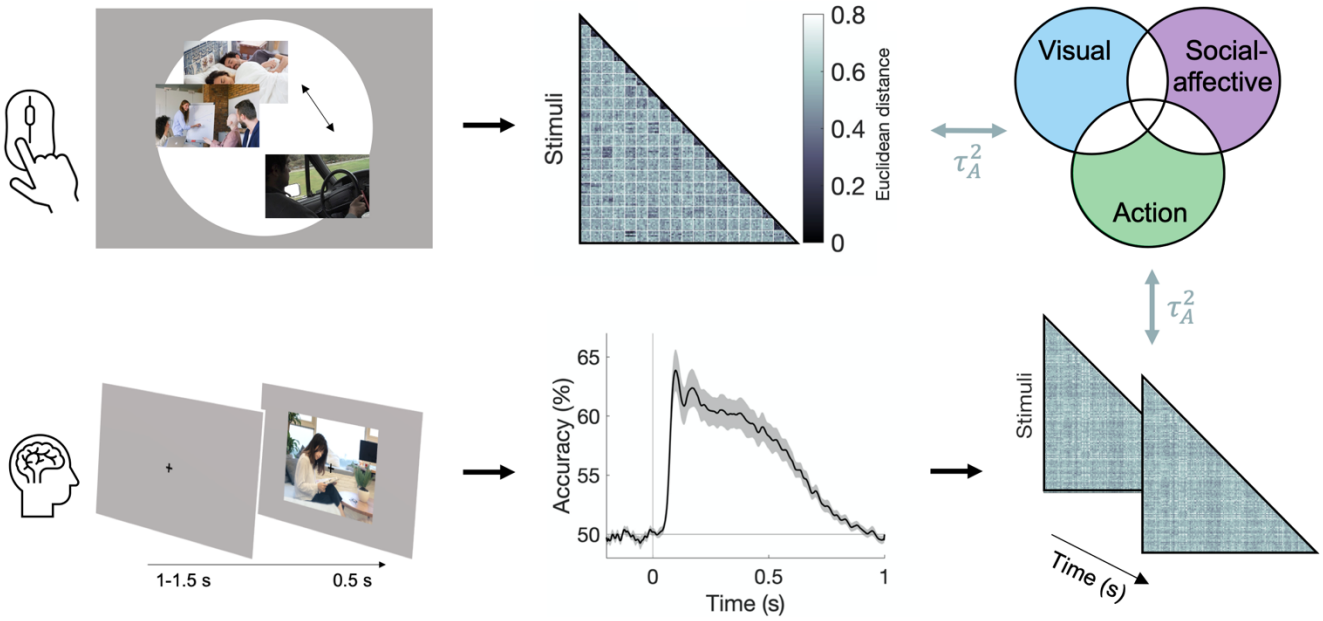

Figure 2. Experimental and analysis pipeline for evaluating the contribution of different features to action representations. Above: a multiple arrangement task was used to generate behavioral RDMs in the two behavioral experiments. Below: EEG data was recorded during a one-back task, and time-resolved neural RDMs were generated using pairwise decoding accuracies. Cross-validated variance partitioning was used to assess the unique contributions of visual, social-affective, and action features to the behavioral and neural RDMs, quantified as the predicted squared Kendall's τ_A. The stimuli in this figure are public domain images similar to the types of videos used in the experiments.

Individual feature contributions to behavioral similarity

To characterize human action representations, we measured behavioral similarity for all pairs of videos in each set in two multiple arrangement experiments (see Materials and methods, section *Multiple arrangement*). Participants arranged videos according to their similarity inside a circular arena (Figure 2). The task involved arranging different subsets of 3-8 videos until sufficiently reliable distance estimates were reached for all pairs of videos. Videos would play on hover, and

148 participants had to play and move each video to proceed to the next trial. In Experiment 1,
149 participants arranged different subsets of 30 videos out of the total 152, while in Experiment 2,
150 participants arranged all 65 videos. To emphasize natural behavior, participants were not given
151 specific criteria to use when judging similarity. Behavioral RDMs containing the Euclidean
152 distances between all pairs of stimuli were reconstructed from each participant's multiple
153 arrangement data using inverse MDS (Kriegeskorte and Mur, 2012).

154 The multiple arrangement task was unconstrained, which meant that participants could use
155 different criteria. Although this may have introduced some variability, the adaptive algorithm
156 used in the multiple arrangement task enabled us to capture a multidimensional representation of
157 how actions are intuitively organized in the mind, while at the same time ensuring sufficient data
158 quality. Data reliability was quantified using leave-one-subject-out correlations of the similarity
159 estimates and was above chance in both experiments (Kendall's $\tau_A = 0.13 \pm 0.08$ and 0.18 ± 0.08
160 respectively, both $P<0.001$, permutation testing; Figure 3 – figure supplement 1a). Reliability
161 was significantly higher in Experiment 2 than in Experiment 1 (Mann-Whitney $z=3.21$,
162 $P=0.0013$), potentially reflecting differences in both participant pools and sampling methods
163 (subsets of videos in Experiment 1 versus full video dataset in Experiment 2; see Materials and
164 methods, section *Multiple arrangement*).

165 We assessed the contribution of 17 different visual, social, and action features to behavior in
166 both experiments by correlating each feature RDM to each participant's behavioral RDM
167 (Supplementary File 1b). In Experiment 1 (Figure 3), only two visual features were significantly
168 correlated with the behavioral RDMs (environment and activations from the final fully-
169 connected layer FC8 of AlexNet). However, there were significant correlations between

behavioral RDMs and all action-related RDMs (action category, effectors, transitivity and activity), as well as all social-affective RDMs (valence, arousal, sociality and number of agents).

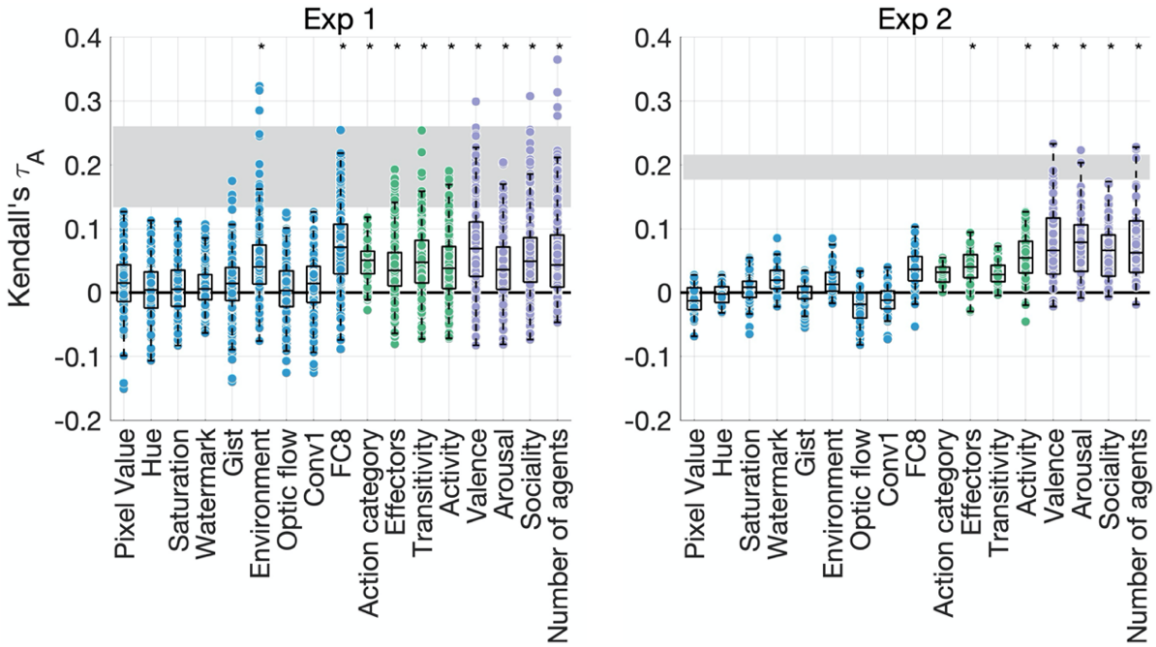

Figure 3. Feature contributions to behavioral similarity. The correlations between each feature RDM and the behavioral RDMs are plotted against the noise ceiling (gray). Each dot is an individual subject. Asterisks denote significance (P<0.005, sign permutation testing). The reliability of the data and feature ratings is presented in Figure 3 – figure supplement 1.

In Experiment 2, the only visual feature that moderately correlated with behavior was the final fully-connected layer of AlexNet (P=0.006, below our threshold for significance). Among action features, only effectors and activity were significantly correlated with the behavioral RDMs. However, we found significant correlations with all social-affective features. The results thus converge across both experiments in suggesting that social-affective and, to a lesser extent, action-related features, rather than visual properties, explain behavioral similarity.

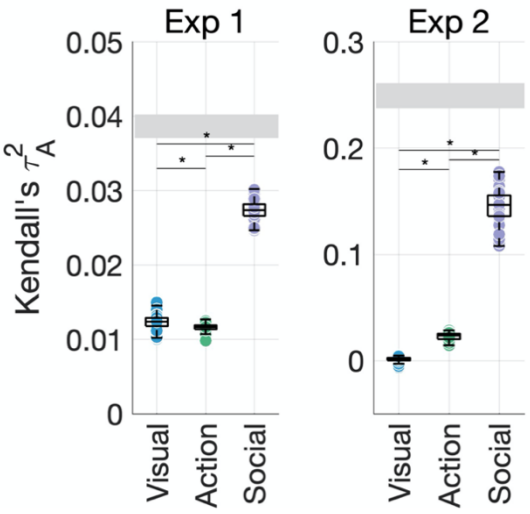

Figure 4. Social-affective features explain behavior better than visual and action features. The unique variance explained by visual, action, and social-affective features is plotted against the split-half reliability of the data (gray). Significant differences are marked with asterisks (all $P<0.001$, Wilcoxon signed-rank tests). For control analyses on how individual features (action category and the number of agents) and their assignment to groups affect the results, see Figure 4 – figure supplements 1-5.

Social-affective features explain the most unique variance in behavioral representations

We performed a cross-validated variance partitioning analysis (Groen et al., 2018; Lescroart et al., 2015; Tarhan et al., 2021) to determine which features contributed the most *unique* variance to behavior (see Materials and methods, section *Variance partitioning*). We selected the ten features that contributed significantly to behavior in either experiment, i.e. two visual features (environment and layer FC8 of AlexNet) and all action and social-affective features. To keep the analysis tractable and understand the contribution of each type of information, we grouped these features according to their type (visual, action and social-affective) and used them as predictors in a cross-validated hierarchical regression (Figure 4). Note that there was no collinearity among

198 the ten predictors, with an average variance inflation factor of 1.34 (Experiment 1) and 1.37
199 (Experiment 2).

200 Together, the ten predictors explained most of the systematic variance in behavior. In
201 Experiment 1, the predicted squared Kendall's τ_A of the full model (τ_A^2=0.06$\pm$0.001) was higher
202 on average than the true split-half squared correlation (τ_A^2=0.04$\pm$0.002). This is likely to be due
203 to the lower reliability of the behavioral similarity data in this experiment, and suggests that the
204 ten predictors are able to explain the data well despite the overall lower prediction accuracy. In
205 Experiment 2, the full model achieved a predicted τ_A^2 of 0.18$\pm$0.1 on average, compared to a true
206 squared correlation of 0.25$\pm$0.1, suggesting that the ten predictors explain most of the variance
207 (73.21%) in the behavioral data.

208 In both experiments, social-affective features contributed significantly more unique variance to
209 behavior than visual or action features (Figure 4, all Wilcoxon z>5.5, all P<0.001). While all
210 three groups of features contributed unique variance to behavior in Experiment 1 (all P<0.001,
211 randomization testing), in Experiment 2, only social-affective features contributed significantly
212 to behavior (P<0.001), while visual and action features did not (P=0.06 and P=0.47
213 respectively). Shared variance between feature groups was not a significant contributor in either
214 dataset. Although the effect sizes were relatively low, social-affective features explained more
215 than twice as much unique variance as either the visual or action features in Experiment 1, and
216 six times as much in Experiment 2. Furthermore, given the limits placed on predictivity by the
217 reliability of the behavioral data, affective features predicted a large portion of the explainable
218 variance in both experiments.

The semantic RDM included among the action features was a categorical model based on activity categories (ATUS, 2019). To assess whether a more detailed semantic model would explain more variance in behavior, we generated a feature RDM using WordNet similarities between the verb labels corresponding to the videos in the Moments in Time dataset. However, replacing the action category RDM with the WordNet RDM did not increase the variance explained by action features (Figure 4 – figure supplement 1).

Similarly, our decision to quantify motion and image properties separately by using an optic flow model may have reduced the explanatory power of motion features in our data. Indeed, a motion energy model (Adelson and Bergen, 1985; Nunez-Elizalde et al., 2021) significantly correlated with behavior in Experiment 1, but not in Experiment 2. However, the addition of this model did not change the pattern of unique feature contributions (Figure 4 – figure supplement 2).

Although the assignment of features to domains was not always straightforward, our results were robust to alternative assignment schemes. For example, high-level visual features can be seen as bordering the semantic domain, while features like the number of agents or the amount of activity can be seen as visual. However, feature assignment was not the main factor driving our results, which stayed the same even when the activity feature was assigned to the visual group. More strikingly, the social-affective feature group explained significantly more variance than all other features grouped together in both experiments (Figure 4 – figure supplement 3). This is a particularly stringent test as it pits the unique and shared contributions of all visual, semantic, and action features against the four social-affective features. In Experiment 1, the combined contribution of visual and action features approached that of social-affective features, while in Experiment 2 the difference was larger. Together with the larger contribution of the number of agents in Experiment 2 (Figure 4 – figure supplement 4 and Figure 4 – figure supplement 5), this

242 suggests that Experiment 2 may have captured more social information, potentially thanks to the
243 exhaustive sampling of the stimuli which allowed each participant to arrange the videos
244 according to different criteria.

245 Among the social-affective features we tested, the number of agents could be seen as straddling
246 the visual and social domains. To assess whether our results were driven by this feature, we
247 performed a control variance partitioning analysis pitting the number of agents against the other,
248 higher-level social-affective features (Figure 4 – figure supplement 3). In both experiments, the
249 higher-level features (sociality, valence and arousal) contributed more unique variance than the
250 number of agents, suggesting that our results are not explained by purely visual factors.

251 Furthermore, an additional analysis looking at the separate contributions of the number of agents,
252 sociality, and affective features (valence and arousal) found that the affective features
253 contributed the greatest variance in both experiments (Figure 4 – Figure supplement 5). For
254 technical reasons, this analysis compared the joint contribution of both affective features to each
255 single social feature and did not discount the impact of variance shared with visual or action-
256 related features. Despite these limitations, the results suggest that the contribution of the social-
257 affective feature group is not driven by the number of agents or the variance it shares with
258 sociality, and highlight the role of affective features (valence and arousal) in explaining
259 behavior.

260 **EEG patterns reflect behavioral similarity**

261 We performed an EEG experiment to investigate how action-relevant features are processed over
262 time. Participants viewed 500 ms segments of the 152 videos from Experiment 1 and performed
263 a one-back action task in which they detected repetitions of the action category (see Materials

and methods, section *EEG: Experimental procedure*). To relate neural patterns to behavioral and feature RDMs, we computed time-resolved neural RDMs for each participant using decoding accuracies between all pairs of videos (Figures 2, 5a). The time-course of decoding performance was similar to that observed in previous E/MEG studies using still visual stimuli (Carlson et al., 2013; Cichy et al., 2014; Dima et al., 2018; Greene and Hansen, 2018; Isik et al., 2014). Decoding accuracy rose above chance at 50 ms after video onset, reached its maximum at 98 ms (63.88±6.82% accuracy), and remained above chance until 852 ms after video onset (cluster-corrected $P<0.05$, sign permutation testing).

To assess brain-behavior correlations, we related the average behavioral RDM obtained in Experiment 1 to the time-resolved neural RDMs (Kendall's τ_A). The behavioral RDM correlated significantly with neural patterns during a cluster between 62 and 766 ms after video onset (Figure 5b), suggesting that the features guiding the intuitive categorization of naturalistic actions also underlie their neural organization.

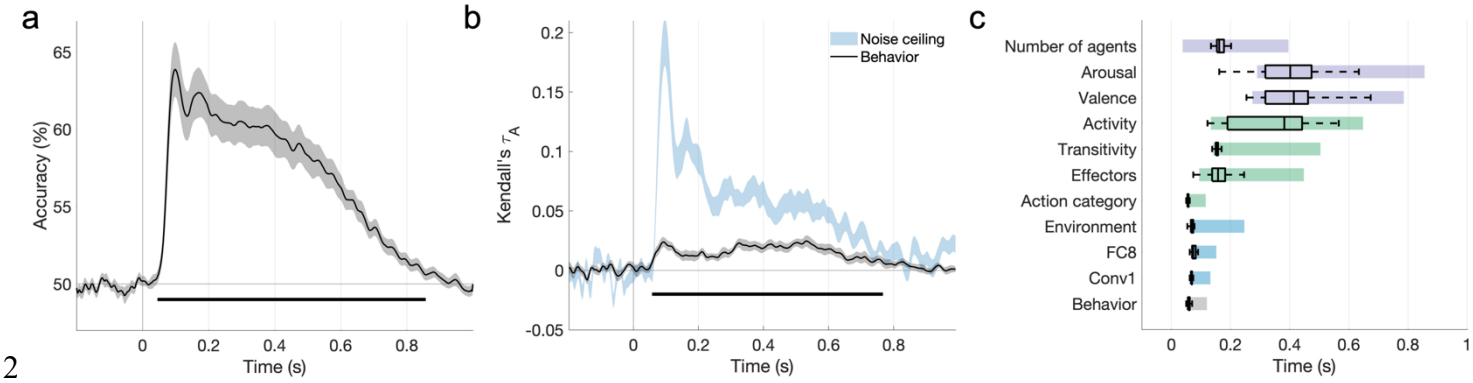

Figure 5. The features that explain behavioral action representations also contribute to neural representations. **a.** Time-course of video decoding accuracy, averaged across all pairs of videos and participants (in gray: SEM across participants). The horizontal line marks above-chance performance (sign permutation testing, cluster-corrected $P<0.05$). **b.** Behavioral similarity correlates with the neural RDM. The noise ceiling is shown in light blue (leave-

one-subject out correlation, mean±SEM). Horizontal lines mark significant time windows (sign permutation testing, cluster-corrected $P<0.05$). **c.** The distribution of significant correlation onsets for each feature model across 1000 bootstrapping iterations (sign permutation testing, cluster-corrected $P<0.05$). Color rectangles show 90% confidence intervals. The timecourses of all feature correlations are shown in Figure 5 – figure supplement 1. The average EEG evoked response is visualized in Figure 5 - figure supplement 2.

Neural timescale of individual feature representations

We assessed the correlations between EEG patterns and the ten feature RDMs found to contribute to behavior in Experiment 1. We also included an additional feature RDM based on the first convolutional layer of AlexNet, which best captures early visual neural responses (Figure 5 – figure supplement 1; see Materials and methods, section *Multivariate analysis*). The feature RDMs that contributed to behavioral similarity also correlated with the EEG patterns (Figure 5 – figure supplement 1), with a single exception (sociality).

A bootstrapping analysis of the cluster onsets of these correlations (Figure 5c) suggests a progression from visual to action and social-affective features. Visual predictors correlated with the neural patterns between 65±15 ms (mean ± SD, Conv1) and 84±62 ms (Environment), while action category also had an early onset (58±9 ms). Other action-related features, however, emerged later (transitivity: 170±67 ms, effectors: 192±94 ms, activity: 345±133 ms). Among social-affective features, the number of agents had the earliest correlation onset (178±81 ms), while valence and arousal emerged later (395±81 and 404±91 ms respectively). Importantly, these features are spontaneously extracted in the brain, as none of them, with the exception of

303 action category, were directly probed in the one-back task performed by participants. In addition,
304 all features were extracted during behaviorally relevant time windows (Figure 5b).

305 **A temporal hierarchy in action perception**

306 A cross-validated variance partitioning analysis revealed different stages in the processing of
307 naturalistic actions (Figure 6). Visual features dominated the early time windows (66-138 ms
308 after video onset). Action features also contributed a significant amount of unique variance (162-
309 598 ms), as well as variance shared with social-affective features (354-598 ms; Figure 6 – figure
310 supplement 1). Finally, social-affective features independently predicted late neural responses
311 (446-782 ms). Importantly, visual features did not share a significant amount of variance with
312 either action or social-affective features.

313 An analysis of effect onsets across 100 split-half iterations points to the hierarchical processing
314 of these features, with a progression from visual to action to social-affective features. Social-
315 affective features (mean onset 418$\pm$89 ms) contributed unique variance significantly later than
316 other feature sets, while action features (245$\pm$104 ms) came online later than visual features

(65±8 ms; all Wilcoxon z>7.27, P<0.001; Figure 6b). A fixed-effects analysis revealed the same order of feature information with larger effect sizes (Figure 6 – figure supplement 2).

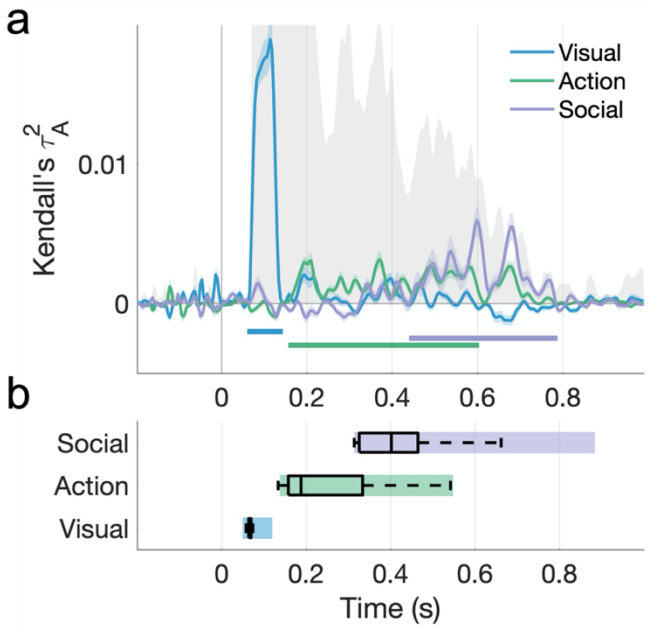

Figure 6. Hierarchical processing of visual, action, and social-affective features. **a.** Unique variance explained by each group of features over time. The split-half reliability of the data is shown in gray (shaded area; see also Figure 5b). Horizontal lines mark significant time windows (sign permutation testing, cluster-corrected P<0.05). The timecourse of shared variance is displayed in Figure 6 – figure supplement 1. See Figure 6 – figure supplement 2 for the results of a fixed-effects analysis. Figure 6 – figure supplement 3 shows how the addition of a motion energy model affects these results. **b.** The distribution of effect onsets across 100 split-half iterations (sign permutation testing, cluster-corrected P<0.05). Color rectangles show 90% confidence intervals.

Motion has been shown to drive the response of visual areas to naturalistic stimuli (Nishimoto et al., 2011; Russ and Leopold, 2015). To better assess the effect of motion on EEG responses, we performed an additional analysis including the motion energy model. There was a sustained correlation between motion energy and EEG patterns beginning at 62 ms (Figure 6 – figure supplement 3). In the variance partitioning analysis, the addition of motion energy increased the unique contribution of visual features and decreased that of action features, indicating that the

action features share variance with motion energy. However, the three stages of temporal processing were preserved in the fixed-effects analysis even with the addition of motion energy, suggesting that the three feature groups made distinct contributions to the neural patterns. Importantly, the unique contribution of social-affective features was unchanged in both analyses by the addition of the motion energy model.

Discussion

Here, we used a large-scale naturalistic stimulus set to disentangle the roles of different features in action perception. Two novel findings emerge from our study. First, our behavioral results suggest that social-affective features play the most important role in how we organize naturalistic everyday actions, above and beyond fundamental visual and action features like scene setting or action category. Second, these behaviorally-relevant features are spontaneously extracted in the brain, and follow a hierarchical sequence from visual to action-related and culminating with social-affective features. These results offer an account of how internal representations of everyday actions emerge in the mind and brain.

Behavioral representations: what features support action perception?

Across two separate multiple arrangement experiments with large-scale naturalistic stimulus sets, we found that social-affective features predicted similarity judgments better than, and independently of, visual and action-related features. By sampling a comprehensive feature space ranging from low-level to conceptual, we were able to distinguish between components that often covary, such as scene setting and action category or sociality and transitivity. Previous studies have operationalized features in different ways, and an exhaustive investigation is thus

difficult; however, our approach of including several important features from each group mitigated this, as suggested by the high amount of variance in behavior collectively explained by our features.

Our work adds to a growing body of evidence for the importance of social-affective features in action perception, and extends it by disentangling the contributions of specific social and semantic features. Previous work has highlighted sociality as an essential feature in neural action representations (Tarhan and Konkle, 2020; Wurm et al., 2017; Wurm and Caramazza, 2019) and a recent study (Tarhan et al., 2021) found that behavioral action similarity judgments were better explained by similarity in actors' goals than by visual similarity. In line with this work, we found a minimal contribution of visual features to action similarity judgments. In contrast, all of our social-affective features – the number of agents, sociality, valence, and arousal – were significantly correlated with behavioral similarity. Furthermore, only two individual action-related features replicated across the two experiments: the amount of activity and the effector (body part) feature, the latter of which is highly relevant to the actors' goals. This could be interpreted as further evidence for the importance of socially-relevant features in our internal representations of actions, and identifies specific social and goal-related features that are important for action understanding.

A hypothesis-driven approach will always pose challenges, due to practical limitations in the number of feature spaces one can feasibly test. Our approach of grouping predictors together based on theoretical distinctions made it possible to rigorously evaluate the unique contributions of different types of features, which is an essential first step in understanding naturalistic action representations. This analysis revealed that social-affective features contributed the most unique variance in both experiments, suggesting that they robustly predict behavioral similarity

judgments, while visual and action features explained little unique variance in either experiment (Figure 4). An exploratory follow-up analysis showed that this effect was primarily driven by affective features (valence and arousal), with the number of agents as a secondary contributor. Recent work found that affective features drive the perceived similarity of memories of real-life events (Tomita et al., 2021), suggesting that these features bridge the action, event and memory domains in organizing mental representations.

Among our social-affective features, the number of agents could be construed as a perceptual precursor to sociality. Indeed, previous fMRI work has suggested that neural representations of actions in the visual system reflect perceptual precursors of social features, rather than higher-level social features (Wurm and Caramazza, 2019). Here, we found that high-level social-affective features (particularly valence and arousal) contributed significantly to behavior independently of the number of agents. Further, affective features explained significantly more unique variance in behavior than the number of agents in both experiments (Figure 4 – figure supplements 4, 5). Our findings suggest that high-level social-affective features, and in particular valence and arousal, uniquely drive human action representations.

Neural representations: how does action perception unfold over time?

Using EEG, we tracked the temporal dynamics of naturalistic action perception. Using naturalistic stimuli and a rich feature space enabled us to disentangle the contributions of different features and investigate their relative timing. Visual, action, and social-affective features made unique contributions to the EEG patterns at different processing stages, revealing a representational hierarchy of spontaneously-extracted features.

Almost all behaviorally-relevant features correlated with the EEG patterns, with action-related and social-affective features emerging later than visual features (Figure 5c). Most action-related features emerged within 200 ms, on the timescale of feedforward processing, which is consistent with prior work showing invariant responses to actions as early as 200 ms (Isik et al., 2018; Tucciarelli et al., 2015), and action transitivity processing as early as 250 ms (Wamain et al., 2014). Among social-affective features, the number of agents emerged earliest (162 ms), pointing to the role of this feature as a perceptual precursor in social perception (Papeo, 2020; Wurm and Caramazza, 2019). Valence and arousal emerged later, around 400 ms after video onset. Interestingly, sociality, which has been highlighted as an important dimension in previous fMRI work on action perception (Tarhan and Konkle, 2020; Wurm et al., 2017), did not correlate with the EEG patterns. This effect was not likely to be driven by a lower reliability in the measurement of this feature, as sociality was more reliable than all other behaviorally rated features in Experiment 1 (Figure 1 – figure supplement 1). While the absence of an effect does not preclude the possibility that this feature is being processed, it is possible that prior work has confounded sociality with other correlated social-affective features (such as the number of agents, or arousal). Alternatively, our operationalization of this feature (which was broader than in some previous studies, e.g. Tucciarelli et al., 2019; Wurm et al., 2017) may have led to differences in the information captured. Note that this finding is mirrored in our behavioral results, where we observed larger unique contributions from valence, arousal, and the number of agents than sociality (Figure 4 – figure supplement 5).

Importantly, these features emerged spontaneously, as the one-back task performed during the EEG recordings only related to action category. However, the semantic processing required to perform the task may have contributed to these computations. The emergence of features irrelevant to the task at hand (action category is not correlated with any other features in the dataset) suggests that this temporal

420 hierarchy would also emerge in the absence of a task; however, future work can more directly test the
421 impact of implicit and explicit (e.g. social-affective) processing on these neural dynamics.

422 Variance partitioning revealed a clear temporal progression from visual features (~100 ms) to
423 action features (~150-600 ms) to social-affective features (~400-800 ms). Importantly, these
424 processing stages emerged after partialling out the contributions of other groups of predictors in
425 a cross-validated analysis, validating our a priori distinctions between feature classes. These
426 findings suggest that the extraction of visual features occurs rapidly, within 200 ms, and is likely
427 supported by feedforward computations. The social-affective features that support behavioral
428 representations, however, were extracted last. This is consistent with theories suggesting that
429 internal visual experience reverses the stages of perceptual processing (Dijkstra et al., 2020;
430 Hochstein and Ahissar, 2002). Specifically, it was the final, social-affective stage of neural
431 processing that was reflected in the intuitive behavioral representations, and not the initially
432 extracted visual features. Furthermore, action-related features were extracted significantly before
433 social-affective features, suggesting the two are not extracted in parallel, but instead pointing to a
434 hierarchy in which both visual and action-related features may contribute to socially relevant
435 computations. Given the short duration of our videos and the relatively long timescale of neural
436 feature processing, it is possible that social-affective features are the result of ongoing processing
437 relying on temporal integration of the previously extracted features. However, more research is
438 needed to understand how these temporal dynamics change with continuous visual input (e.g. a
439 natural movie), and whether social-affective features rely on previously extracted information.

440 Our results add temporal characterization to previous fMRI findings, suggesting that the
441 seemingly conflicting features revealed by previous studies, like the number of agents (Wurm
442 and Caramazza, 2019), sociality (Tarhan and Konkle, 2020), or semantic action category

443 (Tucciarelli et al., 2019), emerge at different stages during action observation. Thus, the
444 existence of different organizing dimensions can be explained not just through spatial
445 segregation within and across brain areas, but also through a temporal gradient starting with
446 visual features and concluding with behaviorally-relevant social and affective representations.
447 More work is needed to understand where these dynamic representations emerge in the brain,
448 and whether they are supported by overlapping or distinct networks. Future research could test
449 this using EEG-fMRI fusion to track the spatiotemporal dynamics of action representations.

450 **Actions in context**

451 As real-world actions tend to occur in a rich social context, studies of action perception should
452 consider socially relevant features and the interactions between different systems for perceiving
453 actions, agents and their mental states (Quadflieg and Koldewyn, 2017). Recent works suggests
454 that social perception enhances visual processing (Bellot et al., 2021; Papeo, 2020) and recruits
455 dedicated neural circuits (Isik et al., 2017; Pitcher and Ungerleider, 2021). Our findings open
456 exciting new avenues for connecting these areas of research. For example, future studies could
457 more explicitly disentangle the perceptual and conceptual building blocks of social and affective
458 features, such as body posture or facial expression, and their roles in action and interaction
459 perception.

460 One fundamental question that lies at the root of this work is how actions should be defined and
461 studied. Here, we adopted a broad definition of the term, focusing on activities as described in
462 the American Use Survey (ATUS, 2019). Although our stimuli were selected to clearly depict
463 short, continuous actions performed by visible agents, their naturalistic and context-rich nature

464 means that they could be understood as "events", encompassing elements that are not singularly
465 specific to actions. A wealth of evidence has shown that context changes visual processing in a
466 non-additive way (Bar, 2004; Willems and Peelen, 2021), and emerging evidence suggests that
467 the same is true for actions (Wurm et al., 2012). Studying actions in context holds promise for
468 understanding how semantically rich representations emerge in naturalistic vision. This, in turn,
469 will pave the way towards a computational understanding of the neural processes that link
470 perception and cognition.

471 Materials and methods

472 Behavior: Stimuli

473 We curated two stimulus sets containing three-second videos of everyday actions from the
474 Moments in Time dataset (Monfort et al., 2019). To broadly sample the space of everyday
475 actions, we first identified the most common activities from the National Bureau of Labor
476 Statistics' American Time Use Survey (ATUS, 2019). Our final dataset included 18 social and
477 non-social activities that lend themselves to visual representation (Table 1), to ensure a diverse
478 and balanced stimulus set representative of the human everyday action space. We note that the
479 ATUS distinctions are based on performed rather than observed actions. While imperfect, they
480 provide an ecologically-relevant and objective criteria with which to define our action space.

481 Action categories were selected from the second-level activities identified in the ATUS. We used
482 a minimum cutoff of 0.14 hours/day to select common actions (Table 1). To diversify our
483 dataset, we added a 'hiking' category (to increase variability in scene setting) and a "fighting"
484 category (for variability along affective dimensions). In addition, "driving" was selected as a

more specific instance of the "travel" categories in ATUS, as it is the most common form of transportation in the US. Some adjustments were also made to the "relaxing and leisure" category, by selecting two specific activities that were easy to represent and distinguish visually, as well as above threshold ("reading" and "playing games"). In addition, our "reading" category included both "reading for personal interest" and "homework and research", as this distinction is difficult to convey visually. We omitted three leisure categories that were difficult to represent in brief videos ("watching TV", "relaxing and thinking", and "computer use for leisure, excluding games"), as well as the "consumer goods purchases" category.

Activity	**Hrs**	**Verb labels (Moments in Time)**
Childcare / taking care of children	0.37	crying, cuddling, feeding, giggling, socializing
Driving	1.17	driving, socializing
Eating	1.06	chewing, eating
Fighting		fighting
Gardening	0.17	gardening, mowing, planting, shoveling, weeding
Grooming	0.68	bathing, brushing, combing, trimming, washing
Hiking		hiking
Housework	0.53	cleaning, dusting, repairing, scrubbing, vacuuming
Instructing and attending class	0.22	instructing, teaching
Playing games	0.26	gambling, playing+fun, playing+videogames, socializing
Preparing food	0.60	barbecuing, boiling, chopping, cooking, frying, grilling, rinsing, stirring
Reading	0.46	reading
Religious activities	0.14	praying, preaching
Sleeping	8.84	resting, sleeping
Socializing and social events	0.64	celebrating, dancing, marrying, singing, socializing, talking
Sports	0.34	exercising, playing+sports, swimming, throwing
Telephoning	0.16	calling, telephoning
Working	3.26	working
Control videos		blowing, floating, raining, shaking

Table 1. Activities from the ATUS included in each of the two stimulus sets, with the amount of daily hours spent performing each activity and the corresponding verb labels from the Moments in Time dataset. Note that control videos were only included in the first dataset. Fighting and hiking were added for variation in valence and action setting.

We curated an initial set of 544 videos from the Moments in Time dataset by identifying the verb labels relevant to our chosen activities. Videos were chosen that were horizontally-oriented (landscape), of reasonable image quality, that clearly represented the activity in question, that clearly depicted an agent performing the activity, and that varied in terms of number of agents, gender and ethnicity of agents, and scene setting. Although some categories were represented by fewer verb labels than others in the final set, our curation procedure aimed to balance important features within and across action categories. We also curated a set of control videos depicting natural and indoors scenes.

We then selected two subsets of videos (1) that sampled all activities in a balanced manner, and (2) where sociality (as assessed through behavioral ratings, see below Section 2.3) was minimally correlated to the number of agents (experimenter-labeled). This was done by randomly drawing 10,000 subsets of videos that sampled all activities equally, and selecting the video set with the lowest correlation between sociality and the number of agents. These two features are difficult to disentangle in naturalistic stimulus sets, and we were able to minimize, though not fully eliminate, this correlation (Figure 1a).

The first stimulus set contained 152 videos (8 videos per activity and 8 additional videos with no agents) and was used in Experiment 1. The videos with no agents were included to provide variation in the dataset along visual properties that did not pertain to actions or agents, as well as variation in the overall dataset in terms of number of agents per video. From the remaining videos, a second set of 76 videos was sampled and manually adjusted to remove videos without agents (in the interest of experimental time) and any videos that were too similar visually to other videos in the same category (e.g., involving a similar number of agents in similar postures). The second stimulus set thus contained 65 videos (3-4 videos per activity) and was used in

Experiment 2. The videos were preprocessed to a framerate of 24 frames-per-second using the *videoWriter* object in MATLAB and resized to 600 x 400 pixels. This was done by first resizing the videos in order to meet the dimension criteria (using MATLAB's *imresize* function with bicubic interpolation). The videos were then cropped to the correct action ratio either centrally or using manually determined coordinates to make sure the action remained clear after cropping.

Behavior: Participants

Behavioral ratings

A total of 256 workers (202 after exclusions, located in the United States, worker age and gender not recorded) from the online platform Amazon Mechanical Turk provided sociality, valence, arousal, and activity ratings of the video stimuli, and 43 workers (35 after exclusions) provided transitivity ratings.

Multiple arrangement

Two separate online multiple arrangement experiments were performed on each of the two stimulus sets. A total of 374 workers from Amazon Mechanical Turk took part in Experiment 1 (300 after exclusions, located in the United States, worker age and gender not recorded). Experiment 2 involved 58 participants (53 after exclusions, 31 female, 20 male, 1 non-binary, 1 not reported, mean age 19.38±1.09) recruited through the Department of Psychological and Brain Sciences Research Portal at Johns Hopkins University.

All procedures for online data collection were approved by the Johns Hopkins University Institutional Review Board (protocol number HIRB00009730) and informed consent was obtained from all participants.

Behavior: Experimental procedure

Behavioral ratings

Participants viewed subsets of 30-60 videos from the initially curated large-scale set and rated the events depicted on a five-point scale. In a first set of experiments, the dimensions rated were: sociality (how social the events were, from 1 – not at all to 5 – very social); valence (how pleasant the events were, from 1 – very unpleasant to 5 – very pleasant); arousal (how intense the events were, from 1 – very calm to 5 – very intense); and activity (how active they were, from 1 – no action to 5 – very active). In separate experiments, participants provided transitivity ratings for the two final stimulus sets (i.e. to what extent the actions involved a person or people interacting with an object, from 1 – not at all to 5 – very much). Participants were excluded if they responded incorrectly to catch trials (approximately 10% of trials) requiring them to label the action shown in the prior video, or if they provided overly repetitive ratings (e.g. using only two unique values or fewer out of five possible ratings throughout the entire experiment). This amounted to an average of 17.46±2.14 ratings per video (Experiment 1) and 18.22±2.09 ratings per video (Experiment 2). The experiments were implemented in JavaScript using the jsPsych library (de Leeuw, 2015).

Multiple arrangement

To characterize human action representations, we measured behavioral similarity using two multiple arrangement experiments. The experiments were conducted on the Meadows platform (www.meadows-research.com) and required participants to arrange the videos according to their similarity inside a circular arena. Participants were free to use their own criteria to determine similarity, so as to encourage natural behavior.

Each trial started with the videos arranged around the circular arena. The videos would start playing on hover, and the trial would not end until all videos were played and dragged-and-dropped inside the arena (Figure 2). Different sets of videos were presented in different trials. An adaptive "lift-the-weakest" algorithm was used to resample the video pairs placed closest together, so as to gather sufficient evidence (or improve the signal-to-noise ratio) for each pair. This procedure was repeated until an evidence criterion of 0.5 was reached for each pair, or until the experiment timed out (Experiment 1: 90 minutes, Experiment 2: 120 minutes). By asking participants to zoom into the subsets previously judged as similar, the task required the use of different contexts and criteria to judge relative similarities. Compared to other methods of measuring similarity, multiple arrangement thus combines efficient sampling of a large stimulus set with adaptive behavior that can recover a multi-dimensional similarity structure (Kriegeskorte and Mur, 2012).

In Experiment 1, participants arranged different subsets of 30 videos from the 152-video set, with a maximum of 7 videos shown in any one trial. The stimuli were sampled in a balanced manner across participants. The task took on average 32±14.4 minutes and 86.8±22.6 trials.

In Experiment 2, all participants arranged the same 65 videos (entire 65-video set), with a maximum of 8 videos shown in any one trial. The task took on average 87.5±24.6 minutes, including breaks, and 289.7±57.3 trials.

The experiments included a training trial in which participants arranged the same seven videos (in Experiment 1) or eight videos (in Experiment 2) before beginning the main task. In both experiments, these videos were hand selected to represent clear examples from four categories. Participants were excluded from further analysis if there was a low correlation between their

training data and the average of all other participants' data (over 2 standard deviations below the mean). They were also excluded if they responded incorrectly to a catch trial requiring them to label the action in previously seen videos.

Inverse MDS was used to construct behavioral dissimilarity matrices containing normalized Euclidean distances between all pairs of videos (Kriegeskorte and Mur, 2012). In Experiment 1, the behavioral RDM contained 11476 pairs with an average of 11.37±3.08 ratings per pair; in Experiment 2, there were 2080 pairs rated by all 53 participants.

Behavior: Data analysis

Representational similarity analysis

Everyday actions can be differentiated along numerous axes. Perceptually, they can differ in terms of visual properties, like the setting in which they take place. They can also be characterized through action-related features like semantic action category, or through social features, like the number of agents involved. Understanding how these features contribute to natural behavior can shed light on how naturalistic action representations are organized. Here, we used representational similarity analysis (RSA) to assess the contribution of visual, action, and social-affective features to the behavioral similarity data.

We quantified features of interest using image properties, labels assigned by experimenters (Supplementary File 1a), and behavioral ratings (provided by participants, see above Section 2.3). We calculated the Euclidean distances between all pairs of stimuli in each feature space, thus generating 17 feature RDMs.

605 To quantify visual features, image properties were extracted separately for each frame of each
606 video and averaged across frames. These included pixel value (luminance), hue, saturation, optic
607 flow (the magnitude of the optic flow estimated using the Horn-Schunck method), and the spatial
608 envelope of each image quantified using GIST (Oliva and Torralba, 2001). We also extracted
609 activations from the first convolutional layer and last fully-connected layer of a pre-trained
610 feedforward convolutional neural network (AlexNet; Krizhevsky et al., 2012). These features
611 were vectorized prior to computing Euclidean distances between them (see Supplementary File
612 1b for the dimensionality of each feature). Two additional experimenter-labeled features were
613 included: scene setting (indoors/outdoors) and the presence of a watermark. To assess whether a
614 motion energy model (Adelson and Bergen, 1985; Nishimoto et al., 2011; Watson and Ahumada,
615 1985) would better capture the impact of motion, we performed control analyses by computing
616 motion energy features for each video using a pyramid of spatio-temporal Gabor filters with the
617 *pymoten* package (Nunez-Elizalde et al., 2021).

618 Action feature RDMs were based on transitivity and activity ratings (provided by participants,
619 see above), as well as action category (a binary RDM clustering the stimuli into activity
620 categories based on the initial dataset designations) and effectors (experimenter-labeled). The
621 latter consisted of binary vectors indicating the involvement of body parts in each action
622 (face/head, hands, arms, legs, and torso). To assess whether a more detailed semantic model
623 would capture more information, we also performed a control analysis using a feature RDM
624 based on WordNet similarities between the verb labels in the "Moments in Time" dataset (Figure
625 4 – figure supplement 1).

626 Social-affective feature RDMs were based on sociality, valence, and arousal ratings (all provided
627 by participants, see Behavioral Ratings above) and the number of agents in each video, which

628 was labeled by experimenters on a four-point scale (from 0, no agent present, to 3, three or more
629 agents present).

630 Each subject's behavioral RDM was correlated to the feature RDMs, and the resulting Kendall's
631 τ_A values were tested against chance using one-tailed sign permutation testing (5000 iterations).
632 P-values were omnibus-corrected for multiple comparisons using a maximum correlation
633 threshold across all models (Nichols and Holmes, 2001).

634 A noise ceiling was calculated by correlating each subject's RDM to the average RDM (upper
635 bound), as well as to the average RDM excluding the left-out subject (lower bound; Nili et al.,
636 2014).

637 **Variance partitioning**

638 Despite low correlations between features of interest in both stimulus sets (Figure 1a), shared
639 variance could still contribute to the representational similarity analysis results. To estimate the
640 unique contributions of the three primary groups of features, we performed a cross-validated
641 variance partitioning analysis, excluding individual features that did not correlate with the
642 behavioral data in the above RSA analysis. The three groups included: visual features (scene
643 setting and the last fully-connected layer of AlexNet), action features (action category, effectors,
644 transitivity, action), and social-affective features (number of agents, sociality, valence, arousal).

645 The behavioral data were randomly split into training and test sets (100 iterations) by leaving out
646 half of the individual similarity estimates for each pair of videos in Experiment 1 (since different
647 participants saw different subsets of videos) or half of the participants in Experiment 2. We fit
648 seven different regression models using the average training RDM (with every possible

combination of the three groups of features), and we calculated the squared Kendall's τ_A between the predicted responses and the average test RDM. These values were then used to calculate the unique and shared portions of variance contributed by the predictors (Groen et al., 2018; Lescroart et al., 2015; Tarhan et al., 2021).

The resulting values were tested against chance using one-tailed sign permutation testing (5000 iterations, omnibus-corrected for multiple comparisons). Differences between groups of features were assessed with two-sided Wilcoxon signed-rank tests.

EEG: Stimuli

The stimulus set from behavioral Experiment 1 was used in the EEG experiment, containing 152 videos from 18 categories, as well as control videos. The three-second stimuli were trimmed to a duration of 0.5 seconds centered around the action as determined by visual inspection, to ensure that the shorter videos were easily understandable. This helped improve time-locking to the EEG signals and allowed for a condition-rich experimental design. An additional 50 videos were included as catch stimuli (25 easily identifiable pairs depicting the same action, manually chosen from the larger stimulus set).

EEG: Participants

Fifteen participants (6 female, 9 male, mean age 25.13±6.81) took part in the EEG experiment. All participants were right-handed and had normal or corrected-to-normal vision. Informed consent was obtained in accordance with the Declaration of Helsinki and all procedures were approved by the Johns Hopkins University Institutional Review Board (protocol number HIRB00009835).

670 EEG: Experimental procedure

671 Continuous EEG recordings with a sampling rate of 1000 Hz were made with a 64-channel Brain
672 Products ActiCHamp system using actiCAP electrode caps in a Faraday chamber. Electrode
673 impedances were kept below 25 kΩ when possible and the Cz electrode was used as an online
674 reference.

675 Participants were seated upright while viewing the videos on a back-projector screen situated
676 approximately 45 cm away. The 152 videos were shown in pseudorandom order in each of 10
677 blocks with no consecutive repetition allowed. In addition, four repetitions of the 25 catch video
678 pairs were presented at random times during the experiment. The video pairs were presented in
679 different orders to minimize learning effects, so that for each video pair (V1, V2), half of the
680 presentations were in the order V1-V2 and half of them were in the order V2-V1. Participants
681 performed a one-back task and were asked to press a button on a Logitech game controller when
682 they detected two consecutive videos showing the same action. Participants were not instructed
683 on what constituted an action, beyond being given "eating" as a simple example. There was a
684 break every 150 trials and participants could continue the experiment by pressing a button. In
685 total, the experiment consisted of 1720 trials (1520 experimental trials and 200 catch trials) and
686 took approximately 45 minutes.

687 The stimuli were presented using an Epson PowerLite Home Cinema 3000 projector with a 60
688 Hz refresh rate. Each trial started with a black fixation cross presented on a gray screen for a
689 duration chosen from a uniform distribution between 1 and 1.5 s, followed by a 0.5 s video. The
690 stimuli were presented on the same gray background and subtended approximately 15 x 13
691 degrees of visual angle. The fixation cross remained on screen and participants were asked to

fixate throughout the experiment. A photodiode was used to accurately track on-screen stimulus presentation times and account for projector lag. The paradigm was implemented in MATLAB R2019a using the Psychophysics Toolbox (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997).

EEG: Data analysis

Preprocessing

EEG data preprocessing was performed using MATLAB R2020b and the FieldTrip toolbox (Oostenveld et al., 2011). First, the EEG data were aligned to stimulus onset using the photodiode data to correct for any lag between stimulus triggers and on-screen presentation. The aligned data were segmented into 1.2 s epochs (0.2 s pre-stimulus to 1 s post-stimulus onset), baseline-corrected using the 0.2 s prior to stimulus onset, and high-pass filtered at 0.1 Hz.

Artefact rejection was performed using a semi-automated pipeline. First, the data were filtered between 110 and 140 Hz and Hilbert-transformed to detect muscle artefacts; segments with a z-value cutoff above 15 were removed. Next, channels and trials with high variance were manually rejected based on visual inspection of a summary plot generated using the *ft_rejectvisual* function in FieldTrip. Finally, independent component analysis (ICA) was performed to identify and remove eye movement components from the data.

Catch trials were removed from the data together with any trials that elicited a button response (13.74%±1.82% of all trials). Of the remaining trials, 8.36%±5.01% (ranging between 25 and 275 trials) were removed during the artefact rejection procedure. A maximum of two noisy electrodes were removed from eight participants' datasets.

712 Prior to further analysis, the data were re-referenced to the median across all electrodes, low-pass
713 filtered at 30 Hz to investigate evoked responses and downsampled to 500 Hz.

714 **Multivariate analysis**

715 We performed multivariate analyses to investigate (1) whether EEG patterns reflected behavioral
716 similarity, and (2) whether different visual, action, and social-affective features explained
717 variance in the neural data.

718 First, time-resolved decoding of every pair of videos was performed using a linear Support
719 Vector Machine classifier as implemented in the LibSVM library (Chang and Lin, 2011). Split-
720 half cross-validation was used to classify each pair of videos in each participant's data. To do
721 this, the single-trial data was divided into two halves for training and testing, whilst ensuring that
722 each condition was represented equally. To improve SNR, we combined multiple trials
723 corresponding to the same video into pseudotrials via averaging. The creation of pseudotrials
724 was performed separately within the training and test sets. As each video was shown 10 times,
725 this resulted in a maximum of 5 trials being averaged to create a pseudotrial. Multivariate noise
726 normalization was performed using the covariance matrix of the training data (Guggenmos et al.,
727 2018). Classification between all pairs of videos was performed separately for each time-point.
728 Data were sampled at 500 Hz and so each time point corresponded to non-overlapping 2 ms of
729 data. Voltage values from all EEG channels were entered as features to the classification model.

730 The entire procedure, from dataset splitting to classification, was repeated 10 times with different
731 data splits. The average decoding accuracies between all pairs of videos were then used to

732 generate a neural RDM at each time point for each participant. To generate the RDM, the
733 dissimilarity between each pair of videos was determined by their decoding accuracy (increased
734 accuracy representing increased dissimilarity at that time point).

735 Next, we evaluated the correlations between each participant's time-resolved neural RDM and
736 the feature RDMs found to correlate with behavioral similarity (Experiment 1). To investigate
737 the link between behavioral and neural representations, we also correlated neural RDMs with the
738 average behavioral RDM obtained from the multiple arrangement task in Experiment 1. This
739 analysis was performed using 10 ms sliding windows with an overlap of 6 ms. The resulting
740 Kendall's τ_A values were tested against chance using one-tailed sign permutation testing (5000
741 iterations, cluster-corrected for multiple comparisons across time using the maximum cluster
742 sum, $\alpha = 0.05$, cluster-setting $\alpha = 0.05$). A noise ceiling was calculated using the same procedure
743 as in the behavioral RSA (see above Section 2.4). Effect latencies were assessed by
744 bootstrapping the individual correlations 1000 times with replacement to calculate 90%
745 confidence intervals around effect onsets.

746 To quantify the contributions of visual, social-affective, and action features to the neural RDMs,
747 a time-resolved cross-validated variance partitioning procedure was performed. Using 100 split-
748 half cross-validation iterations, the neural RDM was entered as a response variable in a
749 hierarchical regression with three groups of feature RDMs (visual, social-affective, and action)
750 as predictors. This analysis employed the same ten feature RDMs used in the behavioral variance
751 partitioning (see above Variance partitioning), with the addition of activations from the first
752 convolutional layer of AlexNet (Conv1). As Conv1 best captures early visual responses (Figure 5
753 – figure supplement 1), its inclusion ensured that we did not underestimate the role of visual
754 features in explaining neural variance. We did not use frame-wise RDMs to model these visual

features; however, our approach of averaging features across video frames was justified by the short duration of our videos and the high correlation of CNN features across frames (Conv1: Pearson's $\rho = 0.89 \pm 0.09$; FC8: $\rho = 0.98 \pm 0.03$).

The analysis was carried out using 10 ms sliding windows with an overlap of 6 ms. The resulting predicted Kendall's τ_A values were tested against chance using one-tailed sign permutation testing (5000 iterations, cluster-corrected for multiple comparisons using the maximum cluster sum across time windows and regressions performed, $\alpha = 0.05$, cluster-setting $\alpha = 0.05$). The distributions of effect onsets across the 100 split-half iterations were compared using two-sided Wilcoxon signed-rank tests.

Data availability

Behavioral and EEG data and results have been archived as an Open Science Framework repository (https://osf.io/hrmxn/). Analysis code is available on GitHub (https://github.com/dianadima/mot_action).

Acknowledgments

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. The authors wish to thank Tara Ghazi, Seah Chang, Alyssandra Valenzuela, Melody Lee, Cora Mentor Roy, Haemy Lee Masson, and Lucy Chang for their help with the EEG data collection, Dimitrios Pantazis for pairwise decoding code, and Emalie McMahon for comments on the manuscript.

Competing interests

The authors declare no competing interests.

References

Adelson, E.H., Bergen, J.R., 1985. Spatiotemporal energy models for the perception of motion. J. Opt. Soc. Am. A 2, 284. https://doi.org/10.1364/josaa.2.000284

Allen, M., Poggiali, D., Whitaker, K., Marshall, T.R., Kievit, R.A., 2019. Raincloud plots: A multi-platform tool for robust data visualization [version 1; peer review: 2 approved]. Wellcome Open Res. 4. https://doi.org/10.12688/wellcomeopenres.15191.1

ATUS, 2019. American Time Use Survey, United States Department of Labor. Bureau of Labor Statistics. Washington, DC.

Bar, M., 2004. Visual objects in context. Nat. Rev. Neurosci. 5, 617–629. https://doi.org/10.1038/nrn1476

Bedny, M., Caramazza, A., 2011. Perception, action, and word meanings in the human brain: The case from action verbs. Ann. N. Y. Acad. Sci. https://doi.org/10.1111/j.1749-6632.2011.06013.x

Bellot, E., Abassi, E., Papeo, L., 2021. Moving Toward versus Away from Another: How Body Motion Direction Changes the Representation of Bodies and Actions in the Visual Cortex. Cereb. Cortex 1–16. https://doi.org/10.1093/cercor/bhaa382

Brainard, D.H., 1997. The Psychophysics Toolbox. Spat. Vis. 10, 433–436. https://doi.org/http://dx.doi.org/10.1163/156856897X00357

Carlson, T., Tovar, D.A., Alink, A., Kriegeskorte, N., 2013. Representational dynamics of object vision: The first 1000 ms. J. Vis. 13, 1–19. https://doi.org/10.1167/13.10.1

Chang, C.C., Lin, C.J., 2011. LIBSVM: A Library for support vector machines. ACM Trans.

Intell. Syst. Technol. 2. https://doi.org/10.1145/1961189.1961199

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

de Leeuw, J.R., 2015. jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. Behav. Res. Methods 47, 1–12. https://doi.org/10.3758/s13428-014-0458-y

Dijkstra, N., Ambrogioni, L., Vidaurre, D., van Gerven, M., 2020. Neural dynamics of perceptual inference and its reversal during imagery. Elife 9, 1–19. https://doi.org/10.7554/eLife.53588

Dima, D.C., Perry, G., Messaritaki, E., Zhang, J., Singh, K.D., 2018. Spatiotemporal dynamics in human visual cortex rapidly encode the emotional content of faces. Hum. Brain Mapp. 39, 3993–4006. https://doi.org/10.1002/hbm.24226

Giese, M.A., Poggio, T., 2003. Cognitive neuroscience: Neural mechanisms for the recognition of biological movements. Nat. Rev. Neurosci. 4, 179–192. https://doi.org/10.1038/nrn1057

Greene, M.R., Hansen, B.C., 2018. Shared spatiotemporal category representations in biological and artificial deep neural networks. PLOS Comput. Biol. 14, e1006327. https://doi.org/10.1371/journal.pcbi.1006327

Groen, I.I., Greene, M.R., Baldassano, C., Fei-Fei, L., Beck, D.M., Baker, C.I., 2018. Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. Elife 7, e32962. https://doi.org/10.7554/eLife.32962

Guggenmos, M., Sterzer, P., Cichy, R.M., 2018. Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. Neuroimage 173, 434–447. https://doi.org/10.1016/J.NEUROIMAGE.2018.02.044

Hafri, A., Trueswell, J.C., Epstein, R.A., 2017. Neural representations of observed actions generalize across static and dynamic visual input. J. Neurosci. 37, 3056–3071. https://doi.org/10.1523/JNEUROSCI.2496-16.2017

822 Haxby, J. V., Gobbini, M.I., Nastase, S.A., 2020. Naturalistic stimuli reveal a dominant role for
823 agentic action in visual representation. Neuroimage 216, 116561.
824 https://doi.org/10.1016/j.neuroimage.2020.116561

825 Hirai, M., Fukushima, H., Hiraki, K., 2003. An event-related potentials study of biological
826 motion perception in humans. Neurosci. Lett. 344, 41–44. https://doi.org/10.1016/S0304-
827 3940(03)00413-0

828 Hirai, M., Hiraki, K., 2006. The relative importance of spatial versus temporal structure in the
829 perception of biological motion: An event-related potential study. Cognition 99, 15–29.
830 https://doi.org/10.1016/j.cognition.2005.05.003

831 Hochstein, S., Ahissar, M., 2002. View from the Top: Hierarchies and Reverse Hierarchies in the
832 Visual System. Neuron 36, 791–804.

833 Humphreys, G.F., Newling, K., Jennings, C., Gennari, S.P., 2013. Motion and actions in
834 language: Semantic representations in occipito-temporal cortex. Brain Lang. 125, 94–105.
835 https://doi.org/10.1016/j.bandl.2013.01.008

836 Isik, L., Koldewyn, K., Beeler, D., Kanwisher, N., 2017. Perceiving social interactions in the
837 posterior superior temporal sulcus. Proc. Natl. Acad. Sci. U. S. A. 114, E9145–E9152.
838 https://doi.org/10.1073/pnas.1714471114

839 Isik, L., Meyers, E.M., Leibo, J.Z., Poggio, T., 2014. The dynamics of invariant object
840 recognition in the human visual system. J. Neurophysiol. 111, 91–102.
841 https://doi.org/10.1152/jn.00394.2013

842 Isik, L., Tacchetti, A., Poggio, T., 2018. A fast, invariant representation for human action in the
843 visual system. J. Neurophysiol. 119, 631–640. https://doi.org/10.1152/jn.00642.2017

844 Jamali, M., Grannan, B.L., Fedorenko, E., Saxe, R., Báez-Mendoza, R., Williams, Z.M., 2021.
845 Single-neuronal predictions of others' beliefs in humans. Nature 591, 610–614.
846 https://doi.org/10.1038/s41586-021-03184-0

847 Johansson, G., 1973. Visual perception of biological motion and a model for its analysis.

Percept. Psychophys. 14, 201–211. https://doi.org/10.3758/BF03212378

Jokisch, D., Daum, I., Suchan, B., Troje, N.F., 2005. Structural encoding and recognition of biological motion: Evidence from event-related potentials and source analysis. Behav. Brain Res. 157, 195–204. https://doi.org/10.1016/j.bbr.2004.06.025

Kleiner, M., Brainard, D.H., Pelli, D.G., Broussard, C., Wolf, T., Niehorster, D., 2007. What's new in Psychtoolbox-3? Perception 36, S14. https://doi.org/10.1068/v070821

Kriegeskorte, N., Mur, M., 2012. Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. Front. Psychol. 3, 1–13. https://doi.org/10.3389/fpsyg.2012.00245

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet Classification with Deep Convolutional Neural Networks. Adv. Neural Inf. Process. Syst. 1–9. https://doi.org/http://dx.doi.org/10.1016/j.protcy.2014.09.007

Lescroart, M.D., Stansbury, D.E., Gallant, J.L., 2015. Fourier power, subjective distance, and object categories all provide plausible models of BOLD responses in scene-selective visual areas. Front. Comput. Neurosci. 9, 1–20. https://doi.org/10.3389/fncom.2015.00135

Lingnau, A., Downing, P.E., 2015. The lateral occipitotemporal cortex in action. Trends Cogn. Sci. 19, 268–277. https://doi.org/10.1016/j.tics.2015.03.006

Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S.A., Yan, T., Brown, L., Fan, Q., Gutfruend, D., Vondrick, C., Oliva, A., 2019. Moments in Time Dataset: one million videos for event understanding. IEEE Trans. Pattern Anal. Mach. Intell. 1–8.

Nastase, S.A., Goldstein, A., Hasson, U., 2020. Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. Neuroimage 222, 117254. https://doi.org/10.1016/j.neuroimage.2020.117254

Nichols, T.E., Holmes, A.P., 2001. Nonparametric Permutation Tests For Functional Neuroimaging : A Primer with Examples. Hum. Brain Mapp. 25, 1–25. https://doi.org/10.1002/hbm.1058

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., Kriegeskorte, N., 2014. A Toolbox for Representational Similarity Analysis. PLoS Comput. Biol. 10, e1003553. https://doi.org/10.1371/journal.pcbi.1003553

Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J.L., 2011. Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. Curr. Biol. 21, 1641–1646. https://doi.org/10.1016/j.cub.2011.08.031

Nunez-Elizalde, A., Deniz, F., Dupré la Tour, T., Visconti di Oleggio Castello, M., Gallant, J., 2021. pymoten: scientific python package for computing motion energy features from video. https://doi.org/https://doi.org/10.5281/zenodo.6349625

Oliva, A., Torralba, A., 2001. Modeling the Shape of the Scene : A Holistic Representation of the Spatial Envelope. Int. J. Comput. Vis. 42, 145–175. https://doi.org/10.1023/A:1011139631724

Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.M., 2011. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput. Intell. Neurosci. 2011, 156869. https://doi.org/10.1155/2011/156869

Papeo, L., 2020. Twos in human visual perception. Cortex 132, 473–478. https://doi.org/10.1016/j.cortex.2020.06.005

Pelli, D.G., 1997. The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spat. Vis. 10, 437–442. https://doi.org/10.1163/156856897X00366

Pitcher, D., Ungerleider, L.G., 2021. Evidence for a Third Visual Pathway Specialized for Social Perception. Trends Cogn. Sci. https://doi.org/10.1016/j.tics.2020.11.006

Quadflieg, S., Koldewyn, K., 2017. The neuroscience of people watching: How the human brain makes sense of other people's encounters. Ann. N. Y. Acad. Sci. 1396, 166–182. https://doi.org/10.1111/nyas.13331

Redcay, E., Moraczewski, D., 2020. Social cognition in context: A naturalistic imaging approach. Neuroimage 216, 116392. https://doi.org/10.1016/j.neuroimage.2019.116392

Russ, B.E., Leopold, D.A., 2015. Functional MRI mapping of dynamic visual features during natural viewing in the macaque. Neuroimage 109, 84–94. https://doi.org/10.1016/j.neuroimage.2015.01.012

Spunt, R.P., Satpute, A.B., Lieberman, M.D., 2011. Identifying the what, why, and how of an observed action: An fMRI study of mentalizing and mechanizing during action observation. J. Cogn. Neurosci. 23, 63–74. https://doi.org/10.1162/jocn.2010.21446

Tarhan, L., De Freitas, J., Konkle, T., 2021. Behavioral and Neural Representations en route to Intuitive Action Understanding. Neuropsychologia 163, 108048. https://doi.org/10.1016/j.neuropsychologia.2021.108048

Tarhan, L., Konkle, T., 2020. Sociality and interaction envelope organize visual action representations. Nat. Commun. 11, 1–11. https://doi.org/10.1038/s41467-020-16846-w

Thornton, M.A., Tamir, D.I., 2021. People accurately predict the transition probabilities between actions. Sci. Adv. 7, eabd4995. https://doi.org/10.1126/sciadv.abd4995

Thornton, M.A., Weaverdyck, M.E., Tamir, D.I., 2019. The brain represents people as the mental states they habitually experience. Nat. Commun. 10, 1–10. https://doi.org/10.1038/s41467-019-10309-7

Tomita, T.M., Barense, M.D., Honey, C.J., 2021. The Similarity Structure of Real-World Memories. bioRxiv 1–38.

Tucciarelli, R., Turella, L., Oosterhof, N.N., Weisz, N., Lingnau, A., 2015. MEG Multivariate Analysis Reveals Early Abstract Action Representations in the Lateral Occipitotemporal Cortex. J. Neurosci. 35, 16034–16045. https://doi.org/10.1523/JNEUROSCI.1422-15.2015

Tucciarelli, R., Wurm, M., Baccolo, E., Lingnau, A., 2019. The representational space of observed actions. Elife 8, 1–24. https://doi.org/10.7554/eLife.47686

Vangeneugden, J., Peelen, M. V., Tadin, D., Battelli, L., 2014. Distinct neural mechanisms for body form and body motion discriminations. J. Neurosci. 34, 574–585. https://doi.org/10.1523/JNEUROSCI.4032-13.2014

Wamain, Y., Pluciennicka, E., Kalénine, S., 2014. Temporal dynamics of action perception: Differences on ERP evoked by object-related and non-object-related actions. Neuropsychologia 63, 249–258. https://doi.org/10.1016/j.neuropsychologia.2014.08.034

Watson, A.B., Ahumada, A.J., 1985. Model of human visual-motion sensing. J. Opt. Soc. Am. 2, 322–342.

Weaverdyck, M.E., Thornton, M.A., Tamir, D.I., 2021. The representational structure of mental states generalizes across target people and stimulus modalities. Neuroimage 238, 118258. https://doi.org/10.1016/j.neuroimage.2021.118258

Willems, R.M., Peelen, M. V., 2021. How context changes the neural basis of perception and language. iScience 24, 102392. https://doi.org/10.1016/j.isci.2021.102392

Wurm, M.F., Caramazza, A., 2019. Lateral occipitotemporal cortex encodes perceptual components of social actions rather than abstract representations of sociality. Neuroimage 202. https://doi.org/10.1016/j.neuroimage.2019.116153

Wurm, M.F., Caramazza, A., Lingnau, A., 2017. Action categories in lateral occipitotemporal cortex are organized along sociality and transitivity. J. Neurosci. 37, 562–575. https://doi.org/10.1523/JNEUROSCI.1717-16.2016

Wurm, M.F., Cramon, D.Y., Schubotz, R.I., 2012. The Context-Object-Manipulation triad: Cross Talk during action perception revealed by fMRI. J. Cogn. Neurosci. 24, 1548–1559. https://doi.org/10.1162/jocn_a_00232

Wurm, X.M.F., Lingnau, A., 2015. Decoding Actions at Different Levels of Abstraction. J. Neurosci. 35, 7727–7735. https://doi.org/10.1523/JNEUROSCI.0188-15.

Supplementary figures

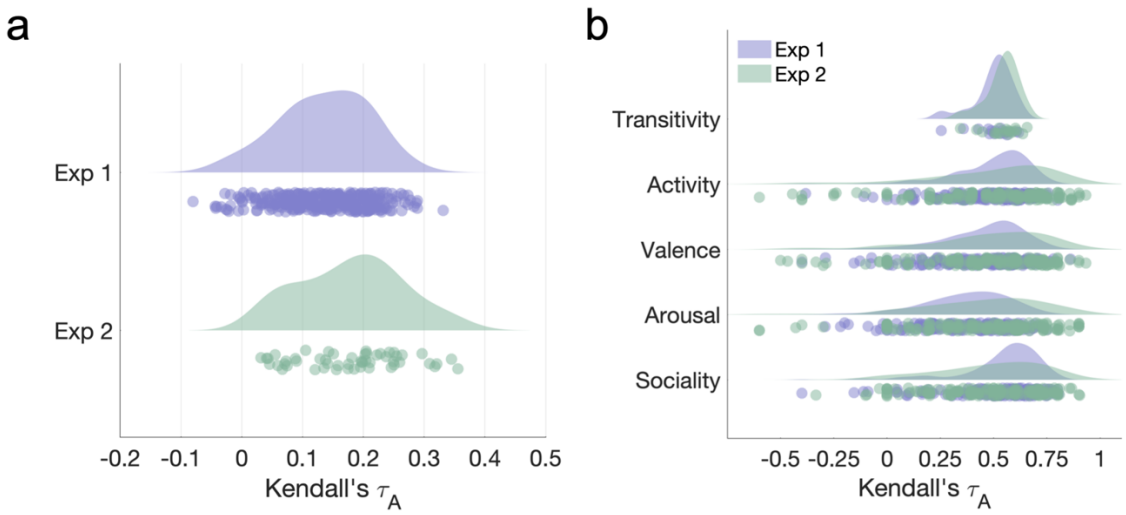

Figure 3 – figure supplement 1. **a.** Reliability of behavioral similarity estimates (leave-one-subject out correlations) in the two datasets. **b.** Reliability of features measured in behavioral experiments (leave-one-subject out correlations). Note that since the ratings were collected using the larger video set, the video sets rated by each participant differed both in number and content. Only participants who rated at least 5 videos from each set were included in the analysis.

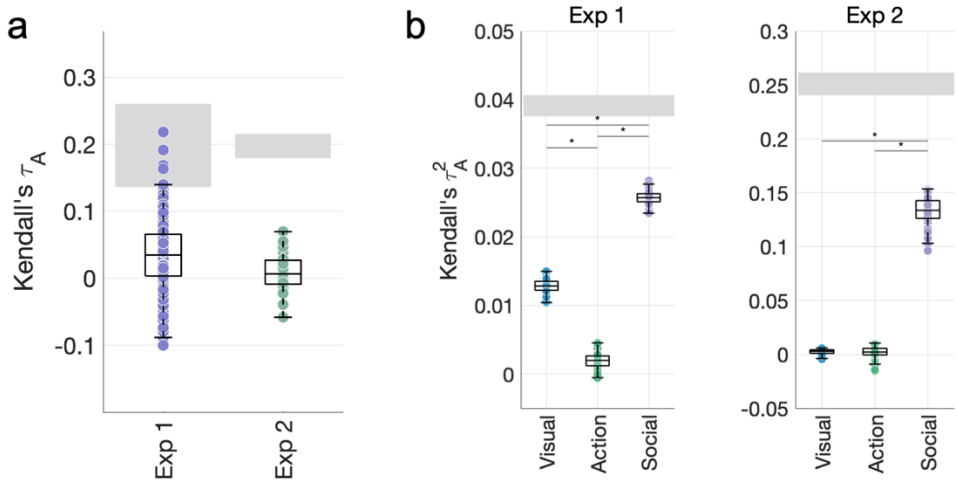

Figure 4 – figure supplement 1. Using a more detailed semantic feature RDM based on WordNet similarities between video labels does not increase the contribution of action features. **a.** Correlations between the WordNet RDM and the behavioral RDMs. **b.** Variance partitioning results obtained after replacing the action category RDM with the WordNet RDM.

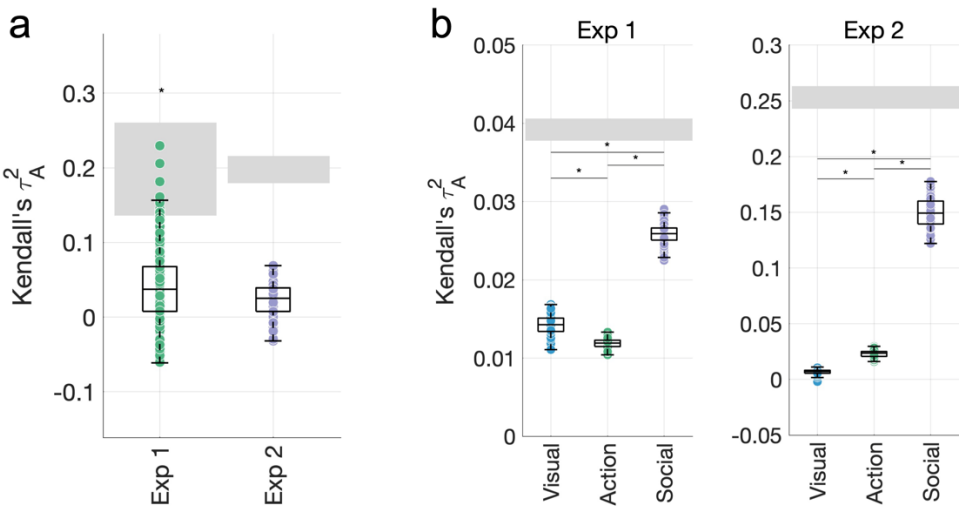

Figure 4 – figure supplement 2. **a.** A motion energy model correlates with the behavioral similarity data in Experiment 1, but not in Experiment 2. **b.** Adding the motion energy model to the group of visual features does not change the pattern of results in our variance partitioning analyses.

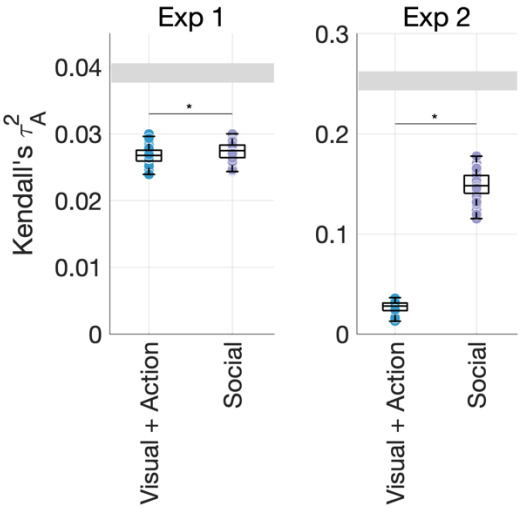

Figure 4 – figure supplement 3. Unique variance explained by visual and action features (environment, FC8, activity, transitivity, effectors, action category) and social-affective features (number of agents, sociality, valence, arousal) in the behavioral data. This shows a striking contribution from social-affective features even when pitted against all other features.

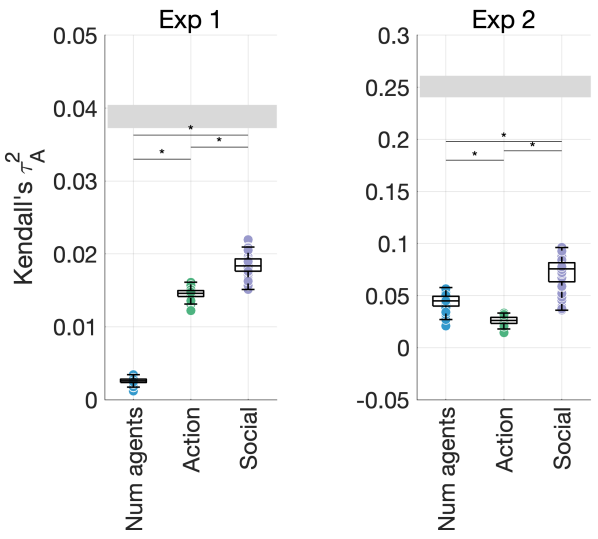

Figure 4 – figure supplement 4. Unique variance explained by the number of agents, action features (action category, effectors, transitivity, activity), and other social-affective features (sociality, valence and arousal) in the behavioral data. This control analysis shows that higher-level social-affective features explain most of the unique variance in our data, above and beyond the number of agents.

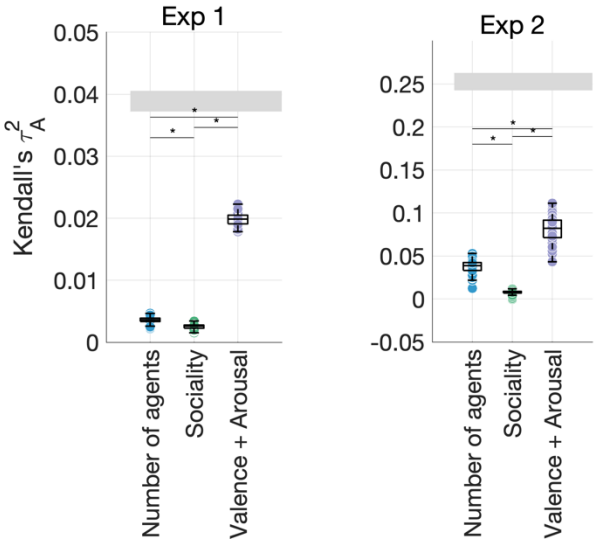

Figure 4 – figure supplement 5. Unique variance explained by the number of agents, sociality, and affective features (valence and arousal) in the behavioral data. This control analysis suggests that affective features explain most of the variance in the data, alongside the number of agents (particularly in Experiment 2). However, this analysis does not account for the contributions of other visual and action features.

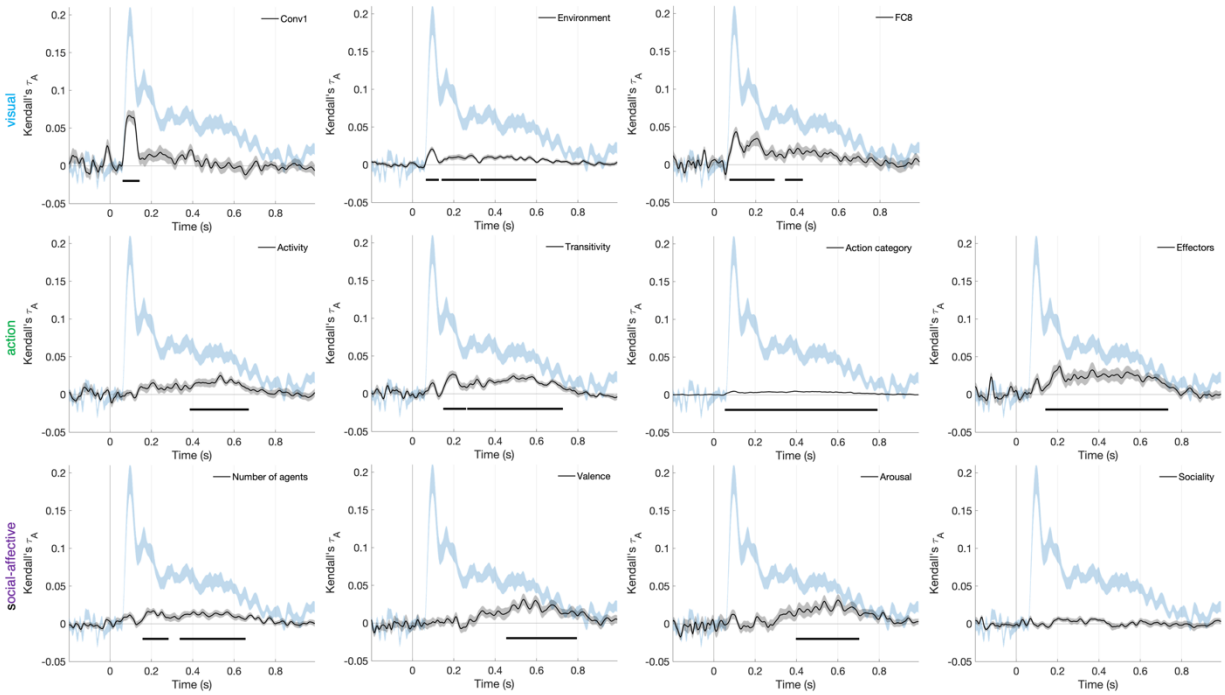

Figure 5 – figure supplement 1. Correlations between feature RDMs and the time-resolved neural RDMs. The first layer of AlexNet (Conv1) was included since it is known to provide the best match to early visual neural responses.

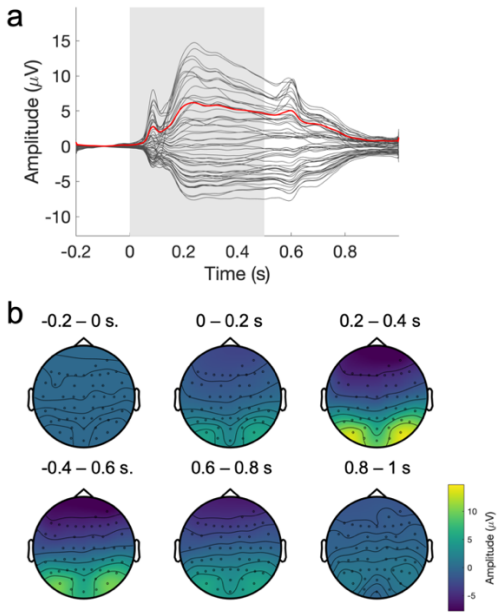

Figure 5 – figure supplement 2. Average EEG evoked response. **a.** Grand average evoked response across participants. The average timecourse of the amplitude at each electrode is shown in black, with the global field power shown in red and the stimulus duration in light grey. **b.** Average topography of the evoked response (re-referenced to the median), averaged across 0.2 s time windows.

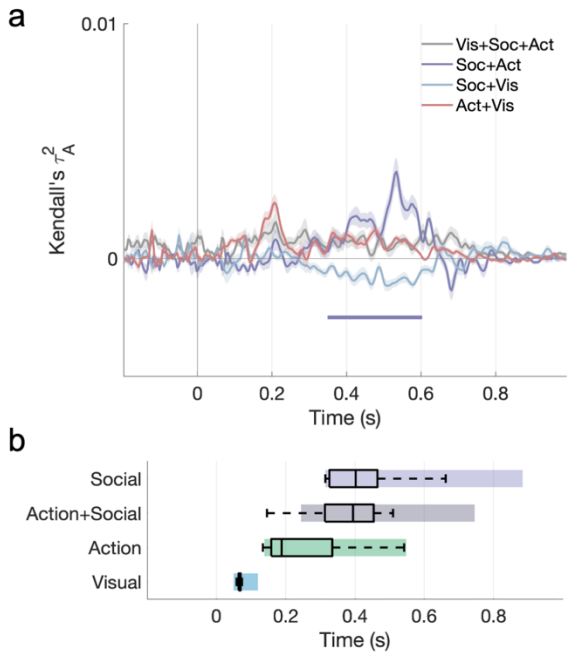

Figure 6 – figure supplement 1. Shared variance among visual, action, and social predictors in the cross-validated variance partitioning analysis. **a.** Timecourse of shared variance. **b.** Onsets (with bootstrapped 90% CIs) for all significant portions of unique and shared variance.

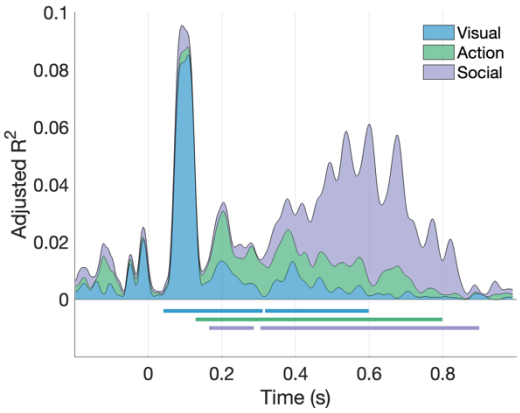

Figure 6 – figure supplement 2. Fixed-effects variance partitioning: results of a hierarchical regression with the average RDM as a response variable (stacked area plot). Only the unique contributions of the three predictor groups are shown. Horizontal lines mark significant time windows (cluster-corrected $P<0.001$, permutation testing).

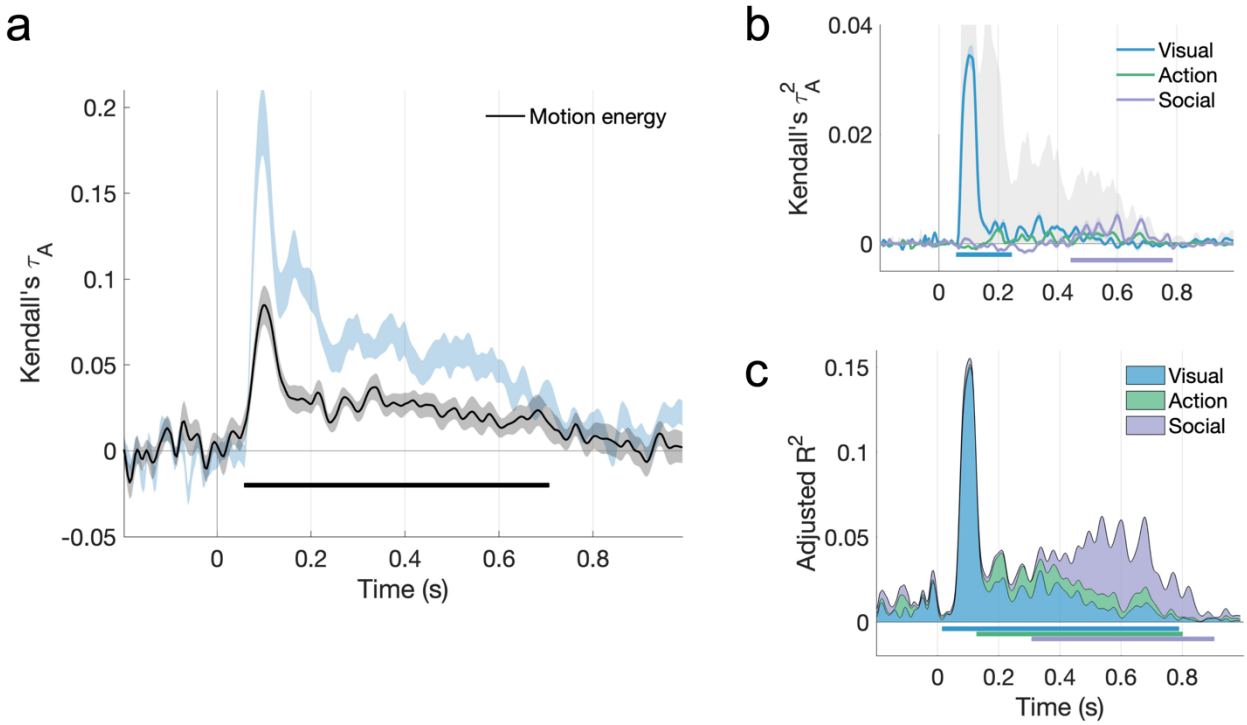

Figure 6 – figure supplement 3. a. The motion energy model correlates with EEG responses. **b.** Including the motion energy model among visual features reduces the unique contribution of action features, but not social features, in the cross-validated variance partitioning analysis. **c.** The temporal hierarchy, however, remains the same in the fixed-effects analysis.

Supplementary File 1. Additional information about the stimulus sets and features used in analysis. **1a.** Breakdown of scene setting and number of agents across the two final stimulus sets. **1b.** Features quantified in both stimulus sets and used to generate feature RDMs in the representational similarity analysis.