Shared representations of human actions across vision and language

Diana C. Dima[1,2*], Sugitha Janarthanan[3], Jody C. Culham[4], Yalda Mohsenzadeh[1,2]

[1]Dept of Computer Science, Western University, London, Ontario, Canada [2]Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada [3]Dept of Biochemistry, Western University, London, Ontario, Canada [4]Dept of Psychology, Western University, London, Ontario, Canada

Abstract

Humans can recognize and communicate about many actions performed by others. How are actions organized in the mind, and is this organization shared across vision and language? We collected similarity judgments of human actions depicted through naturalistic videos and sentences, and tested four models of action categorization, defining actions at different levels of abstraction ranging from specific (action verb) to broad (action target: whether an action is directed towards an object, another person, or the self). The similarity judgments reflected a shared semantic organization across videos and sentences, determined mainly by the target of actions, even after accounting for other semantic features. Large language model features predicted the behavioral similarity of action videos and sentences, and captured information about the target of actions alongside unique semantic information. Together, our results show how modality-invariant action concepts are organized in the human mind and in large language model representations.

Introduction

Daily life requires us to understand a wide range of actions performed by others, as well as communicate about them. Yet actions are an incredibly heterogeneous category, linking multiple visual, semantic and social domains[1–4] and varying along multiple axes. How does the mind organize this rich semantic space, and does this organization differ when actions are communicated visually or verbally?

Actions can be defined at multiple levels of abstraction and can only be understood in relation to agents, objects, and contexts, a complexity that still challenges machine learning algorithms for action recognition. Different brain areas have been shown to represent action categories with different degrees of specificity[5–7]. Despite this, previous work on semantic action representations has tended to either define action categories *a priori* (e.g. using action verbs or broad activity types)[8–10], or to employ a behaviorally estimated semantic space[11,12]. Recent work has investigated the taxonomy of actions using text stimuli of varying levels of abstraction[13]; but it is unclear whether such a taxonomy would generalize to visual actions and different categorization tasks.

Behavioral studies using visual actions and data-driven analyses have mapped representational spaces along broad axes such as *communication*, *locomotion*, *food* etc.[11,14,15] or axes related to facial traits, emotions, and action planning[16]. In a large-scale text analysis, six dimensions organized action representations[17], some of them notably more abstract than their visual counterparts (e.g. *abstraction*, *tradition*, *spirituality*). In light of such results, it has been speculated that the dimensions organizing actions in the mind may be modality-dependent[14]; but to our knowledge, no study to date has investigated semantic action representations using both visual and text stimuli.

A growing body of research suggests that action representations share a neural substrate across vision and language. Adding to long-standing evidence of overlap in neural activations elicited by action images and action verbs[10,18–24], recent investigations have more directly revealed modality-invariant neural representations in human occipito-temporal and parietal cortices[25,26]. These areas have also been shown to encode invariant action category [1,11,27–33], hinting at a putative cross-modal semantic architecture.

Here, we employed semantic similarity to capture how action categories are organized in the mind[34–36]. To accurately represent the richness of a potential action space[37–39], we curated a large-scale naturalistic set of videos and sentences depicting everyday actions defined at four levels of abstraction based on prior research: specific action verbs[10,26] (e.g. *swimming*), everyday activities[8,9,40] (e.g. *sports*), broad action classes inspired by primate research[41,42] (e.g. *locomotion*), and action target types, reflecting whether an action is directed towards an object, another person, or the self[12,43].

To understand how actions are organized in the mind, we used these semantic features, together with a comprehensive set of contextual, perceptual, and computational features, to predict behavioral similarity judgments of the videos and sentences within and across modalities. We found a semantic architecture shared across vision and language, reflecting the target of actions and, to a lesser extent, their class. Features from OpenAI's GPT-3 large language model (LLM) predicted this shared semantic structure better than, and independently of, other semantic features, while also capturing information about the target of actions. Together, these results reveal the modality-invariant architecture of action concepts and its organizing principles in minds and machines.

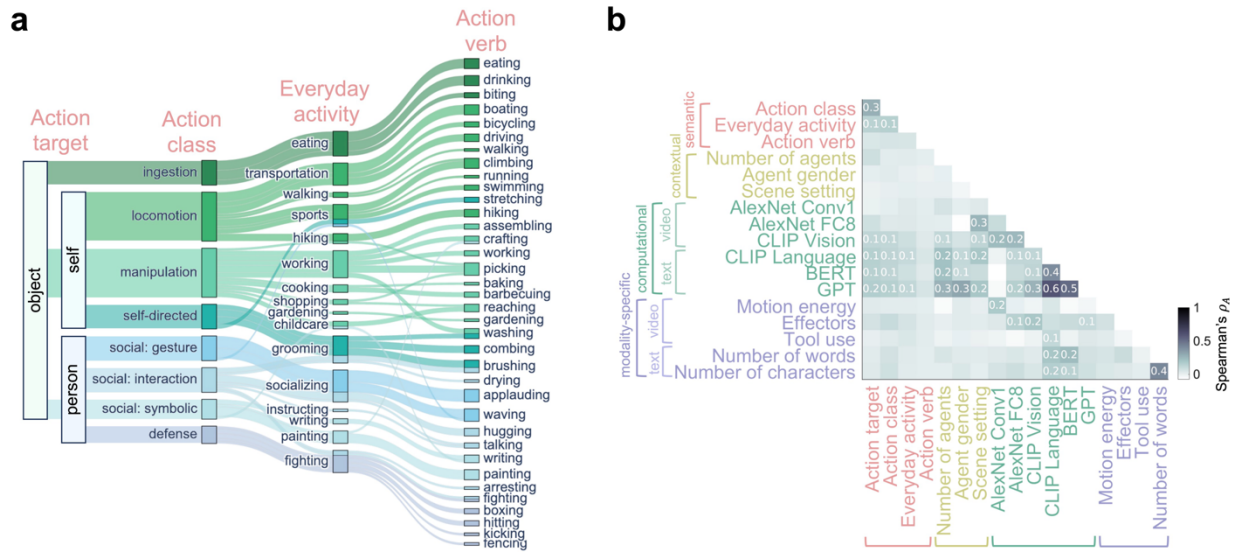

Figure 1. Action features across the stimulus set. **a,** Actions were categorized at different levels of abstraction, from specific to broad, according to their action verb, everyday activity category, action class, and action target. **b,** Correlations across features of interest were low, with the notable exception of more complex computational models (CLIP, BERT, GPT) that moderately correlated with most other features. Both video- and text-based features are included in the correlation matrix, and only correlations over 0.1 are displayed numerically.

Results

Features in naturalistic actions: from semantic to perceptual

We curated a set of 95 videos of natural actions from the Moments in Time dataset[44]. Videos were carefully selected to ensure that semantic, contextual, and perceptual features were minimally correlated across the dataset (**Figure 1b**). For each video, a corresponding naturalistic sentence was constructed following the structure *agent* + *action* (+ *object* if present) + *context*, e.g. *Women are swimming in a pool*. The sentences were displayed as images matched in size to the videos (see Methods: Stimuli).

We labeled the stimuli with a comprehensive set of 18 semantic, contextual, computational, and modality-specific features. First, actions were labeled with four semantic

features at different levels of abstraction based on previous research (**Figure 1a**): by their action verb[10,26,45] (e.g. *swimming*), everyday activity[15,46] (e.g. *sports*), action class[41] (e.g. *locomotion*), and action target[12,43] (e.g. *the self*). The first three roughly correspond to the subordinate, basic and superordinate levels of categorization identified in object research[47–49], while at the same time testing the main action definition criteria prevalent in the literature. The fourth feature (action target) includes information about the actions' goals and their degree of sociality (person-directedenss) and transitivity (object-directedness), features previously found to be essential in action processing[9,12,43].

Apart from their semantic content, naturalistic actions vary along numerous axes, including those related to the agents performing the action and the context in which the action takes place. To capture this information, we labeled each action with contextual features: the number and perceived gender of agents, as well as the scene setting (indoors or outdoors).

Next, to better characterize stimulus features and evaluate how well computational models can capture human action representations, we extracted video features from different vision-trained neural networks (the first convolutional and last fully connected layers of AlexNet[50], the vision module of CLIP[51]) and sentence features (embeddings) from large language models (LLM: CLIP's language module, BERT[52], GPT[53]).

Finally, we accounted for additional sources of variance by extracting modality-specific features encoding perceptual and action information: motion energy, effectors (body parts involved in each action), and tool use (for videos) and the number of words and characters (for sentences).

We generated representational dissimilarity matrices (RDMs) quantifying the pairwise distances between all videos or sentences along each of the 18 semantic, contextual, computational, and modality-specific features (see Methods: Feature definitions).

Correlations among different groups of features were low across both videos and sentences (**Figure 1b**). There were moderate correlations among semantic features, particularly between action target and class (ρ_A=0.27). Features extracted from computational models (particularly LLMs) moderately correlated with semantic, contextual, and perceptual features, pointing to the rich, diverse information captured by these models.

Video and sentence semantic similarity judgments

To investigate how humans represent the meaning of actions, we conducted two multiple arrangement experiments with either the video or sentence stimuli. Participants arranged the videos or sentences inside a circular arena and were instructed to place similar actions closer together and different actions further apart, focusing on the meaning of the actions. Different sets of 3-12 videos were displayed during the task, until all pairwise distances were reliably estimated. This method efficiently captures multi-dimensional representations by presenting the stimuli multiple times in different contexts. Finally, on-screen distances were used to generate behavioral RDMs for each participant using inverse MDS[54].

The resulting dissimilarity matrices reflected clustering of action categories when inspected using t-distributed Stochastic Neighbor Embedding (tSNE) plots (**Figure 2a**). The reliability of the data, quantified by subsampling two sets of 16 participants and correlating their RDMs (1000 iterations), was higher in the video set (ρ_A=0.45±0.01) than the sentence set (ρ_A=0.29±0.01; Mann–Whitney z=38.72, p<0.001).

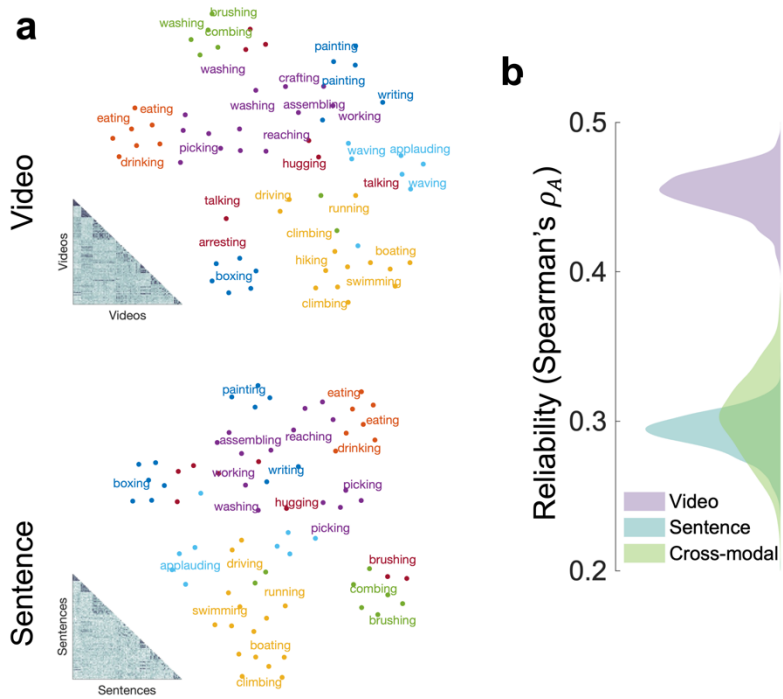

Figure 2. Multiple arrangement results. **a,** Participants arranged the videos and sentences inside a circular arena according to the actions' semantic similarity. Representational dissimilarity matrices (RDMs) were generated from each participant's arrangement. Here, the average RDM is shown in the lower-left corner, together with a t-distributed Stochastic Neighbor Embedding (tSNE) representation color-coded by action class and labeled with action verbs. **b,** Reliability of similarity judgments within and across modalities (correlation between subsets), pointing to a shared structure across modalities. Distributions are based on 1000 data splits with N = 16 across all datasets.

Across modalities, the average behavioral RDMs were substantially correlated (ρ_A=0.4). In a subsampling analysis drawing 16 participants from each modality 1000 times, the cross-modal correlation was ρ_A=0.3±0.02, slightly higher than the internal reliability of the sentence dataset (Mann-Whitney z=8.42, p <0.001; **Figure 2b**) and pointing at a strong cross-modal overlap.

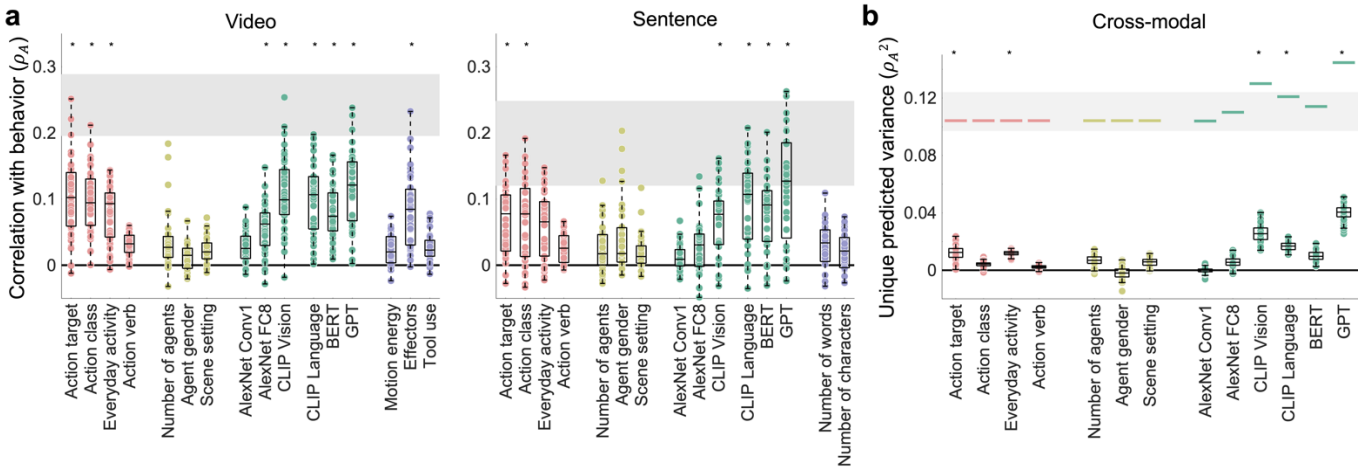

Figure 3. Broad semantic features and computational features best explain behavioral similarity within and across modalities. **a,** Individual feature-behavior correlations are plotted against the noise ceiling (in gray). Each dot is an individual subject. Asterisks denote correlations significantly greater than zero (p<0.005). **b,** Cross-modal analysis: dots are each of 100 subsampling iterations in which 20 participants' data from each modality were used for training and testing a linear regression model. We calculated the unique contributions of (1) each semantic and contextual feature, compared with the total variance explained by all semantic and contextual features; (2) each computational feature, compared with the total variance explained by all semantic and contextual features together with that computational feature (see Methods). The horizontal bars above each feature mark the total variance explained by the combined predictors. The noise ceiling is shown in gray and asterisks denote significance (p<0.005).

A shared organization across vision and language

We first evaluated the contribution of each feature to behavior by correlating each feature RDM to each participant's behavioral RDM (**Figure 3a**). Among the semantic features, the broadest (action target and class) correlated significantly with behavioral similarity across videos and sentences. In the video set, the everyday activity category also correlated with behavioral similarity, suggesting that this feature may be more salient in visual actions. None of the contextual and modality-specific features contributed to behavior, except for the effectors feature in the video set coding the body parts involved in each action.

Features extracted from computational models correlated with behavior across modalities. In the video set, this included higher-level features from visual models (the last layer of AlexNet and CLIP's vision module; see also **Supplementary Figure 1**), as well as all the language models (CLIP, BERT, and GPT). In the sentence set, all LLM features, as well as visual features from CLIP, correlated with behavior. This suggests that LLMs and the multimodal CLIP represent features abstract enough to generalize across modalities. GPT model embeddings were the only feature to reach the noise ceiling in the sentence set, and also reached the highest correlation with behavior in the video set.

Within-modality results were similar across the video and sentence sets. To more rigorously quantify the extent of this overlap, we performed a cross-modal prediction analysis. We calculated the unique contribution of each feature as a predictor in a linear regression trained and tested on different modalities by subsampling each dataset 100 times. Together, the semantic and contextual features reached the noise ceiling (true cross-modal ρ_A^2) in prediction (**Figure 3b**). Among semantic features, the action target and everyday activity were the best unique predictors of cross-modal variance. Among computational features, GPT embeddings predicted more unique variance than any other feature, followed by visual features from CLIP.

Disentangling semantic features in action categorization

Even moderate correlations between related features (**Figure 1b**) could affect their contributions to the behavioral data, making their independent roles difficult to assess. We used cross-validated variance partitioning[12,55] to evaluate the unique and shared variance in behavior contributed by different features. Using 100 subsampling iterations, we trained and tested hierarchical regression models to predict behavior within the video and sentence sets, as well as across modalities.

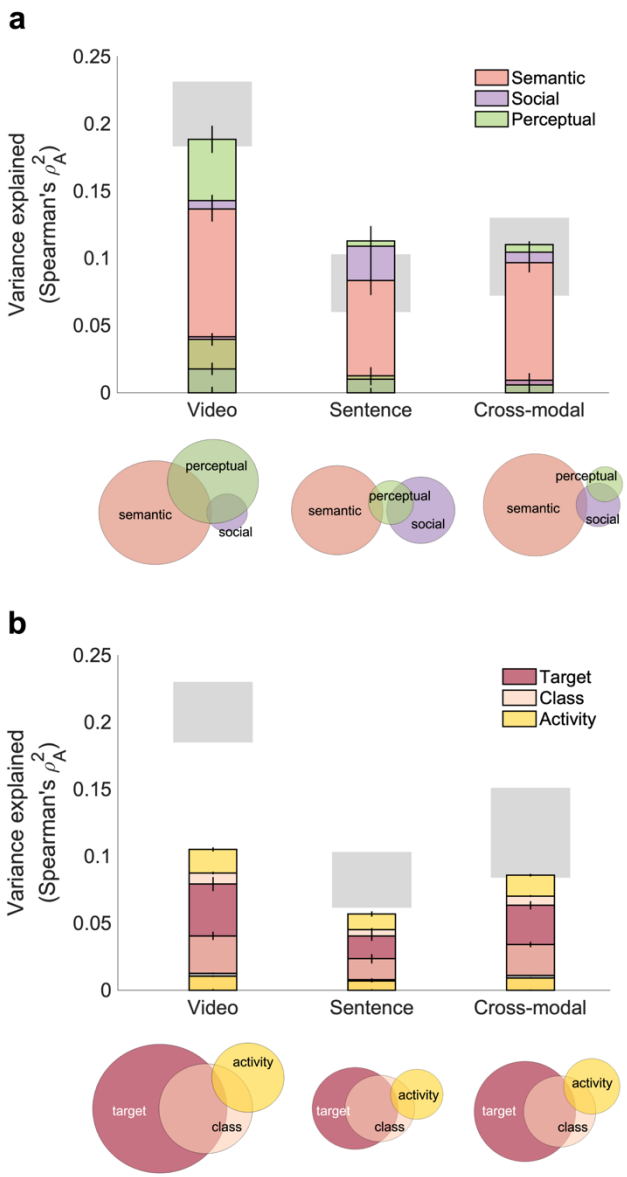

Figure 4. Variance partitioning: semantic features, particularly the action target and class, predict the most variance in behavior within and across modalities. **a,** Stacked plots and Venn diagrams showing the unique and shared contributions of semantic, social, and perceptual features to behavior (see Methods). Error bars are ±SD and the noise ceiling (true ρ_A^2) is shown in gray. As expected, semantic features explain the most variance. Together, the feature set reaches the noise ceiling. **b,** Unique and shared contributions of the three most predictive semantic features. The action target explains the most variance, including variance shared with the action class.

First, to evaluate to what degree the behavioral datasets captured semantic information, we grouped our feature set into three categories: semantic (action target, class, activity, and verb); social (number of agents and agent gender); and perceptual (scene settings and selected modality-specific features; see Methods: Variance partitioning).

Together, the three groups of features reached the noise ceiling in predicting variance within and across modalities (**Figure 4a**). In all cases, semantic features explained the most unique variance ($p<0.001$, randomization testing). Social features explained a significant amount of unique variance in sentence similarity ($p<0.001$), but not video similarity ($p=0.25$) or across modalities ($p=0.19$). Perceptual features explained unique variance in video similarity ($p<0.001$), but not sentence similarity ($p=0.3$) or across modalities ($p=0.26$). This points to differences in the features aiding the semantic categorization of videos and sentences, and shows that only semantic features are shared across modalities.

Next, we disentangled the broad semantic features (action target, action class, and everyday activity) that best correlated with behavior (**Figure 4b**). Both within and across modalities, the action target explained the most unique variance ($p<0.001$), followed by everyday activity ($p<0.001$) and action class (videos: $p=0.02$; sentences: $p=0.004$; cross-modal: $p=0.01$). A significant portion of variance in both datasets was explained jointly by the action target and class ($p<0.001$). This is in line with work suggesting that action goals are essential in action categorization[12,43], and highlights the overlap between different action taxonomies.

Large language model features predict human similarity judgments

Given how well GPT embeddings correlated with behavioral judgments in both modalities, we conducted an additional analysis to assess whether GPT embeddings shared

information with the semantic features. To this end, we entered the GPT embeddings into a variance partitioning analysis together with the action target and class (**Figure 5a**).

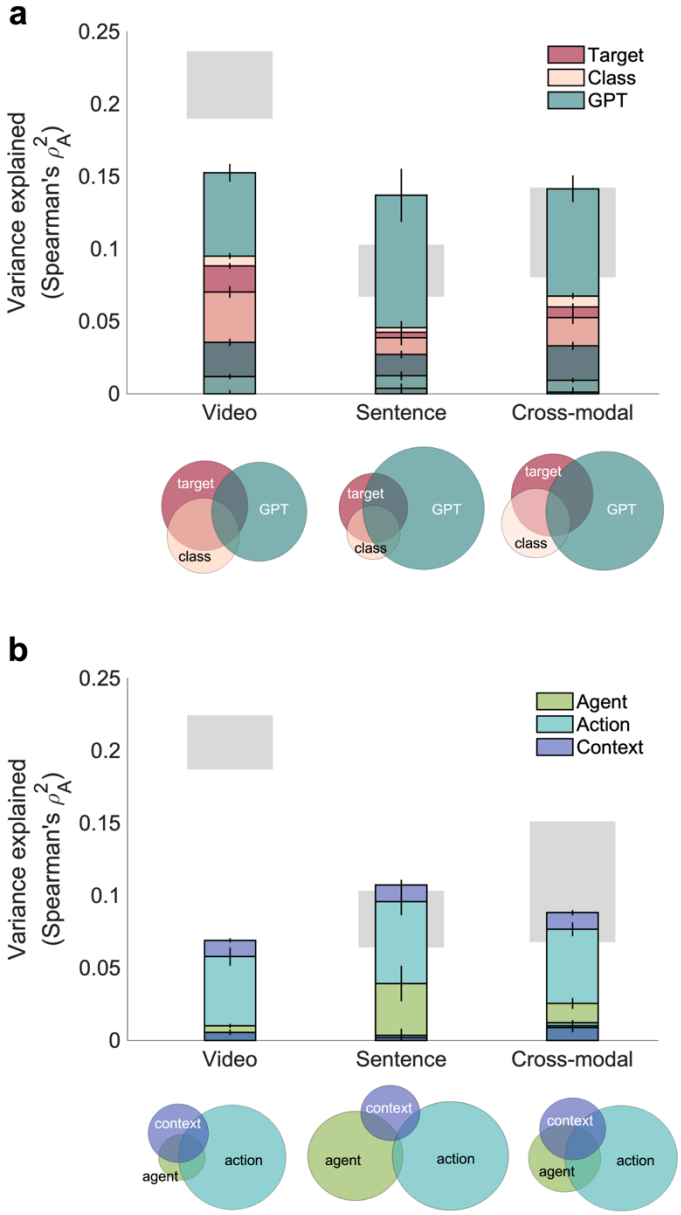

Figure 5. Variance partitioning: understanding GPT representations. **a,** Unique and shared variance in behavior explained by GPT embeddings, action target, and action class. GPT captures more, and different, variance in behavior than the two main semantic features, alongside shared variance with the action target. **b,** Unique and shared contributions of GPT embeddings based on the agent, action, and context parts of sentences. Most of the explanatory power of GPT reflects action information, as well as agent information when predicting sentence similarity.

Both within and across modalities, GPT uniquely explained a large portion of the variance in behavior ($p<0.001$), suggesting that GPT captures more information relevant to human categorization than individual semantic features. In the video dataset and across modalities, there was also significant shared variance between action target and class ($p<0.001$ and $p<0.005$ respectively), as well as between action target and GPT embeddings ($p<0.001$).

As naturalistic sentences contain rich information, it is difficult to know what features are captured by the sentence-level GPT embeddings. To assess this, we extracted separate GPT embeddings for the three sections of each sentence: the agent (e.g. *women*), the action (e.g. *are swimming*), and the context (e.g. *in a pool*). We used these embeddings to predict behavior in a variance partitioning analysis (**Figure 5b**), and found that action embeddings explained most of the unique variance both within and across modalities ($p<0.001$). However, in the sentence dataset, agent embeddings also contributed unique variance ($p<0.001$). This suggests that the categorization of text stimuli relies on social, as well as semantic information, and that GPT embeddings capture both types of information.

Discussion

Using a naturalistic, multimodal dataset and semantic similarity judgments, we interrogated how action concepts are organized in the mind across vision and language. We reveal a shared organization along broad semantic axes reflecting the target and, to a lesser extent, the class of actions. Furthermore, we show that language model embeddings accurately capture this semantic structure across modalities.

212 **A cross-modal organization of actions in the mind**

213 Using a semantic similarity task and a naturalistic set of action videos and sentences, we
214 found a shared semantic organization of actions in the mind across vision and language. This was
215 reflected in the similar categorization of videos and sentences and the high reliability of
216 similarity judgments across modalities (**Figure 1c**), as well as the successful prediction of
217 behavioral judgments across modalities (**Figure 3b**).

218 Previous behavioral studies have focused on either visual stimuli[11,14,15] or text data[17,49] to
219 understand how actions are organized in the mind. However, neural investigations using both
220 visual and text stimuli have found modality-invariant encoding of actions in the brain[25,26]. Our
221 work links these findings to a shared representational space across vision and language, and
222 offers the first behavioral evidence for it.

223 Although semantic features explained most of the variance in behavior across modalities,
224 the contribution of contextual features was different for videos and sentences (**Figure 4a**). In the
225 video arrangement task, perceptual features (scene setting, motion energy, AlexNet final layer
226 activations, tool and effector features) made a significant unique contribution, while social
227 features (number and perceived gender of agents) explained little variance. Conversely, sentence
228 similarity reflected social features, but not perceptual features, alongside semantic information.
229 These differences open exciting avenues for identifying the modality-specific building blocks of
230 action categorization.

231 Previous work has highlighted the important role of social features, particularly sociality
232 (often defined as person-directedness), in intuitive visual action processing[8,9,43,56]. However,
233 sociality can be difficult to disentangle from other semantic and perceptual features, such as
234 action category and number of agents[8,11,57]. Here, the action target feature, which included

aspects of both sociality (person-directedness) and transitivity (object-directedness), explained behavioral judgments best. In contrast, lower-level agent-related features did not capture much variance in the video behavioral data.

Actions are categorized according to their target

Among semantic features, action target, action class, and activity category correlated best with the individual similarity judgments (**Figure 3a**). As these features were moderately correlated (particularly action target and class: $\rho_A = 0.27$), we performed a variance partitioning analysis to assess their unique and shared contributions (**Figure 4b**). Action target uniquely predicted behavioral judgments within and across modalities, in line with previous research highlighting the importance of action goals in organizing action representations in brain and behavior[12,32,58]. Action class, a feature inspired by a proposed human motor repertoire distinguishing between behaviorally relevant action categories[31,41,42], shared variance with the action target. This likely reflects the fact that motor goals are generally shared by exemplars within each proposed action class[41]. It is interesting, however, that action class contributed little unique variance in behavior, suggesting that we should look to broader behaviorally relevant features as organizing axes in action understanding. Indeed, the primacy of goals can arguably allow for more flexible handling of the vast human action space.

Video similarity judgments also correlated with the effectors feature describing what body parts were involved in each action. This information, although highly diagnostic of action goals and types[9,33,41], was not present in the sentence versions of the stimuli.

Language models capture human action categorization

Recently, neural networks have made significant strides in predicting human behavioral judgments[59–62] and neural responses to naturalistic stimuli across vision[63–68] and language[69–72].

However, actions remain a difficult domain to tackle for even state-of-the-art algorithms. We assessed a comprehensive set of computer vision models (AlexNet, CLIP, and others; see **Supplementary Figure 1**) and natural language processing models (CLIP, BERT, and GPT-3) in their ability to capture the semantic organization of action similarity judgments. Features extracted from both types of models were used to explain behavioral data across modalities, to assess how well they capture the covariance between language and the visual world[73–76].

Visual and textual features extracted from CLIP, a multimodal model trained on image and text pairs, captured behavioral representations well within and across modalities, but were surpassed by GPT embeddings (**Figures 3, 5**). GPT explained behavioral similarity better than all other language models, and better than individual semantic features. It captured video similarity better than vision models and uniquely explained a large portion of the variance both within and across modalities. It also shared significant variance with the action target, highlighting the importance of this feature in structuring action representations.

Furthermore, an analysis comparing GPT embeddings extracted from the agent, action, and context parts of each sentence shows that most of the explanatory power of GPT originated from the action itself (**Figure 5b**). In the video dataset, context-related GPT embeddings did not capture perceptual variance as successfully as image features (**Figure 4a**), and thus GPT did not explain all the variance in similarity. However, in the sentence dataset, both action and agent embeddings explained significant variance. Together, the GPT embeddings reached the noise ceiling in both sentence and cross-modal prediction, pointing at the strength of this model in capturing the conceptual organization of actions. Collectively, these results demonstrate the generalizability of LLM representations across tasks and modalities, as well as the potential for using human behavioral data and insights to better understand these models.

Features and categories in minds and machines

Since the discovery of functional selectivity for object categories in the human brain[77–79], a growing body of work has investigated the role of visual features in explaining or supporting categorization[61,80–82]. In action perception, evidence for categorical, invariant representations [1,6,11,28,83,84] has grown alongside evidence for featural representations along the ventral and dorsal streams: from bodies and their motion[79,85–89] to information about tools and effectors [2,27,33,43] and complex social information[8,9,41,90–92]. Importantly, the use of naturalistic stimuli has been essential in uncovering these rich representations, adding to growing evidence of the impact of context in visual processing[37,93–96], and pointing to the importance of actions as a step towards natural event understanding.

Here, we focused on categorical models of semantic organization, grouping actions into classes at different levels of abstraction or according to their targets, effectors, or the involvement of tools. Our results show that a broad categorical model (i.e., whether an action is directed towards a person, an object, or the self) determines much of the semantic organization of actions in the mind, which is consistent across vision and language. However, it remains to be seen whether this organizing feature should be thought of as truly categorical or continuous, mapping actions along a spectrum according to whether they aim to instill change in the self, the environment, or other people.

Furthermore, the growing capabilities of artificial intelligence models, particularly language transformers, raise new questions about their internal representations and their alignment with the human mind. Our results highlight the ability of language models to capture human-like action representations, adding to recent work that suggests a key role for cognitive science in measuring and narrowing the gap between model and human representations[62,97–99].

Together, our results characterize the semantic organization of naturalistic actions across vision and language, and open exciting avenues for investigating conceptual representations and their building blocks in minds and machines.

Methods

Stimuli

Videos

The stimulus set consisted of 95 short videos and sentences depicting everyday actions. Actions were defined at four levels of abstraction from specific (action verb) to broad (action target; **Figure 1a**; see Methods: Feature definitions).

Videos were selected from the Moments in Time dataset[44]. First, we inspected the Moments in Time verb labels to identify actions belonging to eight main classes (defense, ingestion, locomotion, manipulation, self-directed actions, gestures, social interactions, and symbolic actions)[41,42]. We then curated an initial set of 249 videos with the aim of representing each class through a variety of verbs and everyday activities (see Methods: Feature definitions). Videos were chosen that were horizontally oriented (landscape), shot in color, and of reasonable image quality. Each video depicted a single, continuous human action. We excluded videos where agents interacted with the camera. The selected videos varied along multiple axes related to the action (action target, tool use), agents (number, gender), and scene setting, which were manually labeled by the experimenters across the video set.

Since features can covary in naturalistic stimulus sets, we performed further stimulus selection to reduce feature correlations. To this end, we randomly drew 10,000 samples of 100 videos from the initial stimulus set. Each sample included 10 videos per action class, except for

326 locomotion and manipulation, which were represented by 20 videos each to reflect their action
327 diversity.

328 Out of all samples, we selected the set of stimuli with the lowest overall correlation
329 across a range of visual, semantic, agent and action features (mean Spearman's ρ=0.07,
330 SD=0.13). We manually adjusted this set to remove videos that were too visually similar to
331 others in the same category. The final set consisted of 95 videos, with 7-20 videos per action
332 class.

333 All videos were resized to 600 x 400 pixels. To achieve this, we first applied the *imresize*
334 function in MATLAB R2021b using bicubic interpolation. Next, the videos were cropped to the
335 correct aspect ratio either centrally or using manually determined coordinates to ensure the
336 action remained clearly visible. Using the *videoWriter* object in MATLAB R2021b, the videos
337 were trimmed to 2 s centered around the action and resampled to 20 fps.

338 *Sentences*

339 Sentences were generated to succintly describe the action in each video, using a
340 consistent format: Agent + Action [+ Object, if present] + Context; for example: *A girl and a boy*
341 *are eating burgers in a kitchen. Two men are talking in a living room.* (see **Appendix 1** for the
342 full list of sentences).

343 The number and perceived gender of agents were specified where clearly visible. Actions
344 were accompanied by their objects where applicable. There was no difference in overall length
345 between sentences that included an object and those that did not (t(82.7)=1.22, p=0.22). The
346 context was described using a single word or phrase. The most specific wording possible was
347 used to reflect the level of detail present in the video. Although detail was necessarily lost in the
348 transition between modalities, we emphasized natural phrasing and included information about

agents, actions, objects, and contexts to ensure a clear one-to-one correspondence between actions and videos.

Sentence stimuli were 400x600 pixel images with gray borders, ensuring visual consistency with the video stimuli. Each image displayed a sentence in black Arial font on a white background. For readability, each sentence was split into three lines corresponding to the three sentence components (agent, action, and context).

Participants

A sample of 41 participants completed the video arrangement experiment (39 after exclusions, mean age 18.08, SD 0.84, 22 female, 16 male, 1 non-binary). A different cohort of 35 participants completed the sentence arrangement experiment (32 after exclusions, mean age 18.16, SD 0.45, 19 female, 13 male). Participants were recruited through the Department of Psychology Research Participation Pool at Western University. All procedures for online data collection were approved by the Western University Research Ethics Board, and informed consent was obtained from all participants.

Multiple arrangement

We conducted two multiple arrangement experiments to understand semantic action representations across vision (Experiment 1, using videos) and language (Experiment 2, using sentences).

The experiments were conducted on the Meadows platform (https://meadows-research.com/). Participants were instructed to arrange the videos or sentences according to the actions' semantic similarity, placing similar actions closer together and different actions further apart.

371 The task took 90 minutes to complete and an average of 186 (videos) or 215 (sentences)
372 trials. The difference between the numbers of trials in the two tasks was not significant
373 ($t(60)=0.97$, $p=0.34$). Each trial started with a maximum of 12 videos or sentences arranged
374 around a circular arena. The videos played automatically on hover. All stimuli had to be
375 dragged-and-dropped inside the arena before the trial ended. Participants viewed different
376 subsets of stimuli during each trial. Stimulus selection was guided by a 'lift-the-weakest'
377 algorithm so as to optimize the signal-to-noise ratio of the distance estimates[54].
378 Before the main task, participants completed a training trial in which they arranged the
379 same three videos (in Experiment 1) or sentences (in Experiment 2). Participants were excluded
380 if their training data correlated poorly to the average of all other participants' data (over 2 SDs
381 below the mean). Participants were also excluded if they responded incorrectly to a catch trial
382 requiring them to label the action in previously seen stimuli.
383 Using inverse MDS, participants' arrangements were used to generate behavioral
384 dissimilarity matrices quantifying the Euclidean distances between all pairs of videos. We
385 visualized the clustering of action categories using t-SNE representations of the average RDMs,
386 which display the stimuli as points in a 2D space.

387 **Feature definitions**

388 *Semantic features*

389 Actions can be defined at several levels of abstraction, from specific verbs to broad
390 classes. Which of these guides human judgments about the meaning of actions? We tested
391 several features to understand the structure of semantic action representations in the mind. Each
392 feature was converted into a representational dissimilarity matrix (RDM) quantifying the
393 dissimilarity between stimuli along that axis.

394 Our four main hypothesis-driven semantic features were: the action target (whether an
395 action was directed towards an object, a person, or the self); action class (based on a putative
396 human motor repertoire[41]); everyday activity (based on activity categories from the American
397 Time Use Survey[40]); and action verb (based on the Moments in Time verb labels; see **Figure 1a**
398 for a complete list of semantic categories). These were binary RDMs, in which distances
399 between stimuli belonging to the same category were set to 0, and distances between stimuli
400 belonging to different categories were set to 1.

401 *Contextual features*

402 To capture information about the agents (social features), we included the number of
403 agents (labeled on a three-point scale from 1, one agent present, to 3, three or more agents
404 present) and their perceived genders. To capture action context, we included scene setting as a
405 perceptual feature (indoors/outdoors). Contextual features were labeled by experimenters and
406 converted to binary or Euclidean distance RDMs.

407 *Computational features*

408 To assess whether computational models can capture human semantic judgments, we
409 extracted activations from computational models trained on visual and text stimuli.

410 As the videos were short and did not contain major visual changes across frames, we
411 extracted activations using the first frame of each video. We used a pretrained AlexNet
412 feedforward neural network trained on ImageNet, and extracted features from its first
413 convolutional layer (Conv1, to capture low-level visual properties), and its last fully connected
414 layer (FC8, to capture higher-level features). We also extracted features from CLIP, a
415 multimodal neural network trained on publicly available image-text pairs[51]. As CLIP contains
416 both image and text encoders, we used it to extract features from both videos and sentences.

417 AlexNet features were extracted using the Deep Learning toolbox in MATLAB R2022a, while
418 CLIP features were extracted using the THINGSvision[100] toolbox in Python 3.10.9.
419 We also extracted embeddings (features measuring the relatedness of text strings) from
420 Google's BERT[52] and OpenAI's GPT[51,53] transformers. We used the latest version of GPT
421 embeddings at the time of writing, *text-embedding-ada-002*, which outperform previous GPT
422 embeddings despite being lower-dimensional. To assess the contributions of different sentence
423 components, we also extracted separate GPT embeddings for the agent, action, and context
424 portion of each sentence. All text features were extracted from the Hugging Face transformers
425 library[101] using PyTorch in Python 3.10.9. RDMs were created using the pairwise Euclidean
426 distances between feature vectors.
427 *Modality-specific features*
428 Finally, we quantified modality-specific features to capture additional sources of
429 variance. For videos, these included motion energy computed using a pyramid of spatio-temporal
430 Gabor filters as implemented in the *pymoten* package[102]. We also included two video-specific
431 action features which were not present or discernible in the sentences: tool use (binary feature:
432 whether the action was tool-mediated), and effectors (binary vectors indicating the body parts
433 involved in each action: face/head, hands, arms, legs, and torso).
434 For the sentence set, we created two low-level features reflecting the number of words
435 and characters in each sentence.
436 **Representational Similarity Analysis (RSA)**
437 To assess each feature's contribution to the human similarity judgments, we calculated
438 Spearman's ρ_A correlations between each feature RDM and each participant's behavioral RDM.

439 While Spearman's ρ can favour categorical models with many tied ranks, Spearman's ρ_A solves
440 this issue using random tiebreaking to compute the expected correlation[103].
441 We assessed the significance of the correlations using sign permutation testing (5000
442 iterations, one-tailed). P-values were omnibus-corrected for multiple comparisons using a
443 maximum correlation threshold across models[104,105].
444 A noise ceiling was calculated by correlating each participant's RDM to the average
445 RDM across participants (upper bound), as well as to the average RDM excluding the left-out
446 subject (lower bound)[106].
447 **Cross-modal prediction**
448 The average RDMs obtained from video and sentence arrangement were moderately
449 correlated (Spearman's ρ_A=0.4). To understand the semantic structure shared between the two
450 datasets, we performed a cross-modal prediction analysis. Across 100 subsampling iterations, we
451 randomly selected 20 participants' RDMs from each dataset and averaged them.
452 Using the four semantic features and three contextual features as predictors, we trained an
453 ordinary least-squares linear regression model on the video data and tested it on the sentence data
454 (and viceversa). Results were averaged across both directions of training and testing.
455 We used the Spearman's ρ_A^2 correlation between the predicted and test RDMs as a
456 measure of total explained variance. To determine the unique contribution of each feature f_i, we
457 performed a leave-one-feature-out analysis, and subtracted the variance explained by all other
458 features from the total variance, e.g. $\rho_{A_{f_1}}^2 = \rho_{A_{f_1+f_2+f_3}}^2 - \rho_{A_{f_2+f_3}}^2$.
459 To assess the cross-modal predictive power of each computational model, we subtracted
460 the variance explained by the semantic and contextual features from the total variance explained
461 by the semantic and contextual features together with each computational feature. Because the

computational features correlated with each other, we did not group them together (**Figure 1b**). This analysis allowed us to assess whether our set of features predicted behavioral judgments across modalities, as well as the unique contribution of each individual feature.

We assessed the unique variance contributed by each feature against chance using sign permutation testing (5000 iterations, one-tailed). P-values were omnibus-corrected for multiple comparisons using a maximum correlation threshold across models[104,105].

Variance partitioning

To disentangle the contributions of specific sets of features, including their unique and shared contributions to the behavioral similarity judgments, we performed cross-validated variance partitioning both within and across modalities.

In the within-modality analysis, we performed 100 random splits of each behavioral dataset into training and test sets of approximately equal sizes. For the cross-modal analysis, we drew 100 random subsets of 20 participants from each behavioral dataset. We used video data for training and sentence data for testing (and vice versa). We averaged results across both directions of training and testing.

We first evaluated whether the behavioral data reflected semantic information more than social and perceptual information. To this end, we selected three groups of features: semantic (action target, action class, everyday activity, and action verb); social (number of agents and agent gender); and perceptual (for videos: scene setting, tool use, effectors, motion energy, and features from the FC8 layer of AlexNet; for sentences: scene setting and number of words; cross-modal: scene setting). Each group of features was considered a separate predictor in this analysis.

At each iteration, we fit seven regression models using every possible subset of the three predictors and the averaged training data. As a measure of prediction accuracy, we calculated the squared Spearman's ρ_A correlation between the predicted data and the averaged test data.

We then used this to calculate the unique and shared variance contributed by each group of features[8,55,107]. We tested the results against chance using one-tailed sign permutation testing (5000 iterations, omnibus-corrected for multiple comparisons).

We repeated this analysis with different sets of predictors to evaluate: (1) the variance explained by each semantic feature that significantly correlated with the behavioral similarity data (action target, action class, and everyday activity); (2) the variance explained by GPT embeddings compared to the most predictive semantic features (action target and action class); (3) the variance explained by GPT embeddings separately extracted from the agent, action, and context portion of each sentence.

Acknowledgements

The authors would like to thank Leyla Isik for helpful comments on the manuscript and Martin Hebart for code calculating the Spearman's ρ_A. This material is based upon work supported by BrainsCAN at Western University through the Canada First Research Excellence Fund (CFREF) to Y.M., a Western Interdisciplinary Development Initiatives Grant to Y.M. and J.C.C., a Vector Institute for Artificial Intelligence Research Grant to Y.M. and D.C.D., and a Western Postdoctoral Fellowship to D.C.D.

Data availability

The data and results have been archived as an Open Science Framework repository (https://osf.io/c5bdv/). Analysis code is available on GitHub (https://github.com/dianadima/multi-action).

References

1. Lingnau, A. & Downing, P. E. The lateral occipitotemporal cortex in action. *Trends Cogn Sci* **19**, 268–277 (2015).
2. Leshinskaya, A., Wurm, M. F. & Caramazza, A. Concepts of Actions and their Objects. in *The Cognitive Neurosciences* (eds. Gazzaniga, M., Mangun, G. & Poeppel, D.) 757–765 (MIT Press, 2020).
3. Baker, C. L., Saxe, R. & Tenenbaum, J. B. Action understanding as inverse planning. *Cognition* **113**, 329–349 (2009).
4. Thornton, M. A. & Tamir, D. I. The brain represents situations and mental states as sums of their action affordances. *PsyArXiv* 1–52 (2023).
5. Spunt, R. P., Kemmerer, D. & Adolphs, R. The neural basis of conceptualizing the same action at different levels of abstraction. *Soc Cogn Affect Neurosci* **11**, 1141–1151 (2016).
6. Wurm, X. M. F. & Lingnau, A. Decoding Actions at Different Levels of Abstraction. *The Journal of Neuroscience* **35**, 7727–7735 (2015).
7. Zhuang, T., Kabulska, Z. & Lingnau, A. The representation of observed actions at the subordinate, basic and superordinate level. *The Journal of Neuroscience* JN-RM-0700-22 (2023) doi:10.1523/JNEUROSCI.0700-22.2023.
8. Dima, D. C., Tomita, T. M., Honey, C. J. & Isik, L. Social-affective features drive human representations of observed actions. *Elife* **11**, (2022).
9. Tarhan, L. & Konkle, T. Sociality and interaction envelope organize visual action representations. *Nat Commun* **11**, 1–11 (2020).
10. Bedny, M. & Caramazza, A. Perception, action, and word meanings in the human brain: The case from action verbs. *Annals of the New York Academy of Sciences* vol. 1224 81–95 Preprint at https://doi.org/10.1111/j.1749-6632.2011.06013.x (2011).
11. Tucciarelli, R., Wurm, M., Baccolo, E. & Lingnau, A. The representational space of observed actions. *Elife* **8**, 1–24 (2019).
12. Tarhan, L., De Freitas, J. & Konkle, T. Behavioral and Neural Representations en route to Intuitive Action Understanding. *Neuropsychologia* **163**, 108048 (2021).
13. Zhuang, T. & Lingnau, A. The characterization of actions at the superordinate, basic and subordinate level. *Psychol Res* (2021) doi:10.1007/s00426-021-01624-0.
14. Kabulska, Z. & Lingnau, A. The cognitive structure underlying the organization of observed actions. *Behav Res Methods* (2022) doi:10.3758/s13428-022-01894-5.
15. Dima, D. C., Hebart, M. N. & Isik, L. A data - driven investigation of human action representations. *Sci Rep* 1–11 (2023) doi:10.1038/s41598-023-32192-5.

16. Vinton, L. C. *et al.* Four fundamental dimensions underlie the perception of human actions. *Atten Percept Psychophys* (2023) doi:10.3758/s13414-023-02709-1.
17. Thornton, M. A. & Tamir, D. I. Six Dimensions Describe Action Understanding: The ACT-FASTaxonomy. *J Pers Soc Psychol* **122**, 577–605 (2021).
18. Martin, A., Haxby, J. V, Lalonde, F. M., Wiggs, C. L. & Ungerleider, L. G. Discrete cortical regions associated with knowledge of color and knowledge of action. *Science (1979)* **270**, 102–105 (1995).
19. Spunt, R. P. & Lieberman, M. D. Dissociating modality-specific and supramodal neural systems for action understanding. *Journal of Neuroscience* **32**, 3575–3583 (2012).
20. Aziz-Zadeh, L., Wilson, S. M., Rizzolatti, G. & Iacoboni, M. Congruent Embodied Representations for Visually Presented Actions and Linguistic Phrases Describing Actions. *Current Biology* **16**, 1818–1823 (2006).
21. Bidet-Ildei, C., Beauprez, S. A. & Badets, A. A review of literature on the link between action observation and action language: advancing a shared semantic theory. *New Ideas Psychol* **58**, 100777 (2020).
22. Beauprez, S. A. & Bidet-Ildei, C. Perceiving a Biological Human Movement Facilitates Action Verb Processing. *Current Psychology* **38**, 1355–1359 (2019).
23. Humphreys, G. F., Newling, K., Jennings, C. & Gennari, S. P. Motion and actions in language: Semantic representations in occipito-temporal cortex. *Brain Lang* **125**, 94–105 (2013).
24. Quandt, L. C., Lee, Y. S. & Chatterjee, A. Neural bases of action abstraction. *Biol Psychol* **129**, 314–323 (2017).
25. Wurm, M. F. & Caramazza, A. Distinct roles of temporal and frontoparietal cortex in representing actions across vision and language. *Nat Commun* **10**, 1–10 (2019).
26. Aflalo, T. *et al.* A shared neural substrate for action verbs and observed actions in human posterior parietal cortex. *Sci Adv* **6**, (2020).
27. Wurm, M. F. & Caramazza, A. Two 'what' pathways for action and object recognition. *Trends Cogn Sci* **26**, 103–116 (2022).
28. Tucciarelli, R., Turella, L., Oosterhof, N. N., Weisz, N. & Lingnau, A. MEG Multivariate Analysis Reveals Early Abstract Action Representations in the Lateral Occipitotemporal Cortex. *Journal of Neuroscience* **35**, 16034–16045 (2015).
29. Hafri, A., Trueswell, J. C. & Epstein, R. A. Neural representations of observed actions generalize across static and dynamic visual input. *Journal of Neuroscience* **37**, 3056–3071 (2017).
30. Culham, J. C. & Valyear, K. F. Human parietal cortex in action. *Curr Opin Neurobiol* **16**, 205–212 (2006).
31. Urgen, B. A. & Orban, G. A. The unique role of parietal cortex in action observation: Functional organization for communicative and manipulative actions. *Neuroimage* **237**, 118220 (2021).
32. Urgen, B. A., Pehlivan, S. & Saygin, A. P. Distinct representations in occipito-temporal, parietal, and premotor cortex during action perception revealed by fMRI and computational modeling. *Neuropsychologia* **127**, 35–47 (2019).

33. Jastorff, J., Begliomini, C., Fabbri-Destro, M., Rizzolatti, G. & Orban, G. A. Coding observed motor acts: Different organizational principles in the parietal and premotor cortex of humans. *J Neurophysiol* **104**, 128–140 (2010).
34. Shepard, R. N. Towards a Universal Law of Generalization for Psychological Science. *Science (1979)* **237**, 1317–1323 (1987).
35. Edelman, S. Representation is representation of similarities. *Behav Brain Sci* **21**, 449–498 (1998).
36. Murphy, G. L. *The big book of concepts*. (MIT Press, 2002).
37. Haxby, J. V., Gobbini, M. I. & Nastase, S. A. Naturalistic stimuli reveal a dominant role for agentic action in visual representation. *Neuroimage* **216**, 116561 (2020).
38. Nastase, S. A., Goldstein, A. & Hasson, U. Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. *Neuroimage* **222**, 117254 (2020).
39. Redcay, E. & Moraczewski, D. Social cognition in context: A naturalistic imaging approach. *Neuroimage* **216**, 116392 (2020).
40. ATUS. *American Time Use Survey*. *United States Department of Labor. Bureau of Labor Statistics* (2019).
41. Orban, G. A., Lanzilotto, M. & Bonini, L. From Observed Action Identity to Social Affordances. *Trends Cogn Sci* **25**, 493–505 (2021).
42. Graziano, M. S. A. & Aflalo, T. N. Mapping behavioral repertoire onto the cortex. *Neuron* **56**, 239–251 (2007).
43. Wurm, M. F., Caramazza, A. & Lingnau, A. Action categories in lateral occipitotemporal cortex are organized along sociality and transitivity. *Journal of Neuroscience* **37**, 562–575 (2017).
44. Monfort, M. *et al.* Moments in Time Dataset: one million videos for event understanding. *IEEE Trans Pattern Anal Mach Intell* 1–8 (2019).
45. Peelen, M. V., Romagno, D. & Caramazza, A. Independent representations of verbs and actions in left lateral temporal cortex. *J Cogn Neurosci* **24**, 2096–2107 (2012).
46. Tarhan, L. & Konkle, T. Sociality and interaction envelope organize visual action representations. *Nat Commun* **11**, 1–11 (2020).
47. Rosch, E. Cognitive representations of semantic categories. *J Exp Psychol Gen* **104**, 192–233 (1975).
48. Iordan, M. C., Greene, M. R., Beck, D. M. & Fei-Fei, L. Basic Level Category Structure Emerges Gradually across Human Ventral Visual Cortex. *J Cogn Neurosci* **27**, 1427–1446 (2015).
49. Zhuang, T. & Lingnau, A. The characterization of actions at the superordinate, basic and subordinate level. *Psychol Res* **86**, 1871–1891 (2022).
50. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv Neural Inf Process Syst* 1–9 (2012) doi:http://dx.doi.org/10.1016/j.protcy.2014.09.007.
51. Radford, A. *et al.* Learning Transferable Visual Models From Natural Language Supervision. *Arxiv preprint* (2021).
52. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies - Proceedings of the Conference **1**, 4171–4186 (2019).

53. Brown, T. B. *et al.* Language models are few-shot learners. *Adv Neural Inf Process Syst* **2020-Decem**, (2020).
54. Kriegeskorte, N. & Mur, M. Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Front Psychol* **3**, 1–13 (2012).
55. Lescroart, M. D., Stansbury, D. E. & Gallant, J. L. 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 (2015).
56. Wurm, M. F. & Caramazza, A. Lateral occipitotemporal cortex encodes perceptual components of social actions rather than abstract representations of sociality. *Neuroimage* **202**, (2019).
57. McMahon, E. & Isik, L. Seeing social interactions. *Trends Cogn Sci* **xx**, (2023).
58. Gallivan, J. P. & Culham, J. C. Neural coding within human brain areas involved in actions. *Curr Opin Neurobiol* **33**, 141–149 (2015).
59. Muttenthaler, L., Dippel, J., Linhardt, L., Vandermeulen, R. A. & Kornblith, S. Human alignment of neural network representations. *ICLR 2023* (2022).
60. Jozwik, K. M. *et al.* Face dissimilarity judgments are predicted by representational distance in morphable and image-computable models. *Proc Natl Acad Sci U S A* **119**, 1–11 (2022).
61. Jozwik, K. M., Kriegeskorte, N., Storrs, K. R. & Mur, M. Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Front Psychol* **8**, 1726 (2017).
62. Golan, T., Siegelman, M., Kriegeskorte, N. & Baldassano, C. Testing the limits of natural language models for predicting human language judgments. *ArXiv* 1–39 (2023).
63. Kriegeskorte, N. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annu Rev Vis Sci* **1**, 417–446 (2015).
64. Cichy, R. M., Khosla, A., Pantazis, D. & Oliva, A. Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *Neuroimage* (2016) doi:10.1016/j.neuroimage.2016.03.063.
65. Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A. & Oliva, A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci Rep* **6**, 1–13 (2016).
66. Jozwik, K. M., Kietzmann, T. C., Cichy, R. M., Kriegeskorte, N. & Mur, M. Deep neural networks and visuo-semantic models explain complementary components of human ventral-stream representational dynamics. 1–10 (2021).
67. Al-Tahan, H. & Mohsenzadeh, Y. Reconstructing feedback representations in the ventral visual pathway with a generative adversarial autoencoder. *PLoS Comput Biol* **17**, 1–19 (2021).
68. Mohsenzadeh, Y., Mullin, C., Lahner, B. & Oliva, A. Emergence of Visual Center-Periphery Spatial Organization in Deep Convolutional Neural Networks. *Sci Rep* **10**, 4638 (2020).
69. Kumar, S. *et al.* Shared functional specialization in transformer-based language models and the human brain.

70. Hosseini, E. A. *et al.* Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training. *bioRxiv* 2022.10.04.510681 (2022).
71. Schrimpf, M. *et al.* The neural architecture of language: Integrative modeling converges on predictive processing. *Proc Natl Acad Sci U S A* **118**, e2105646118 (2021).
72. Caucheteux, C. & King, J. R. Brains and algorithms partially converge in natural language processing. *Commun Biol* **5**, (2022).
73. Conwell, C., Prince, J. S., Hamblin, C. & Alvarez, G. A. Controlled assessment of CLIP-style language-aligned vision models in prediction of brain and behavioral data. *ICLR 2023* 1–12 (2023).
74. Popham, S. F. *et al.* Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nat Neurosci* **24**, (2021).
75. Marjieh, R. *et al.* Words are all you need? Capturing human sensory similarity with textual descriptors. (2022).
76. Wang, A. Y., Kay, K., Naselaris, T. & Tarr, M. J. Natural language supervision with a large and diverse dataset builds better models of human high-level visual cortex. (2023).
77. Kanwisher, N. Domain specificity in face perception. *Nat Neurosci* **3**, 759–763 (2000).
78. Kanwisher, N., McDermott, J. & Chun, M. M. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience* **17**, 4302–11 (1997).
79. Downing, P. E., Jiang, Y., Shuman, M. & Kanwisher, N. A cortical area selective for visual processing of the human body. *Science (1979)* **293**, 2470–2473 (2001).
80. Bracci, S., Ritchie, J. B. & de Beeck, H. O. On the partnership between neural representations of object categories and visual features in the ventral visual pathway. *Neuropsychologia* **105**, 153–164 (2017).
81. Haxby, J. V *et al.* Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science (1979)* **293**, 2425–30 (2001).
82. Jozwik, K. M., Kriegeskorte, N. & Mur, M. Visual features as stepping stones toward semantics : Explaining object similarity in IT and perception with non-negative least squares. *Neuropsychologia* **83**, 201–226 (2016).
83. Tacchetti, A., Isik, L. & Poggio, T. Invariant recognition drives neural representations of action sequences. *PLoS Comput Biol* **13**, 1–20 (2017).
84. Isik, L., Tacchetti, A. & Poggio, T. A fast, invariant representation for human action in the visual system. *J Neurophysiol* **119**, 631–640 (2018).
85. Felleman, D. J. & Van Essen, D. C. Distributed hierachical processing in the primate cerebral cortex. *Cerebral Cortex* **1**, 1–47 (1991).
86. Dittrich, W. H. Action categories and the perception of biological motion. *Perception* **22**, 15–22 (1993).
87. Neri, P., Morrone, M. C. & Burr, D. C. Seeing biological motion. *Nature* **395**, 894–896 (1998).
88. Grossman, E. D. & Blake, R. Brain areas active during visual perception of biological motion. *Neuron* **35**, 1167–1175 (2002).
89. Giese, M. A. & Poggio, T. Cognitive neuroscience: Neural mechanisms for the recognition of biological movements. *Nat Rev Neurosci* **4**, 179–192 (2003).

90. Thornton, X. M. A., Weaverdyck, M. E. & Tamir, X. I. The social brain automatically predicts others' future mental states. *Journal of Neuroscience* **39**, 140–148 (2019).
91. Spunt, R. P., Satpute, A. B. & Lieberman, M. D. 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 (2011).
92. McMahon, E., Bonner, M. F. & Isik, L. Hierarchical organization of social action features along the lateral visual pathway. *Current Biology* **33**, 1–13 (2023).
93. Wurm, M. F., Cramon, D. Y. & Schubotz, R. I. The Context-Object-Manipulation triad: Cross Talk during action perception revealed by fMRI. *J Cogn Neurosci* **24**, 1548–1559 (2012).
94. Bar, M. Visual objects in context. *Nat Rev Neurosci* **5**, 617–629 (2004).
95. Willems, R. M. & Peelen, M. V. How context changes the neural basis of perception and language. *iScience* **24**, 102392 (2021).
96. Magri, C., Elmoznino, E. & Bonner, M. F. Context predicts commonsense notions of object similarity. (2020).
97. Manrique, N. F., Bao, W., Herbelot, A. & Hasson, U. Enhancing Interpretability using Human Similarity Judgements to Prune Word Embeddings. *ArXiv* (2023).
98. Schwartz, D., Toneva, M. & Wehbe, L. Inducing brain-relevant bias in natural language processing models. *Adv Neural Inf Process Syst* **32**, (2019).
99. Toneva, M. & Wehbe, L. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Adv Neural Inf Process Syst* **32**, (2019).
100. Muttenthaler, L. & Hebart, M. N. THINGSvision: A Python Toolbox for Streamlining the Extraction of Activations From Deep Neural Networks. *Front Neuroinform* **15**, 1–12 (2021).
101. Wolf, T. *et al.* Transformers: State-of-the-Art Natural Language Processing. *ArXiv* 38–45 (2020) doi:10.18653/v1/2020.emnlp-demos.6.
102. Nunez-Elizalde, A., Deniz, F., Dupré la Tour, T., Visconti di Oleggio Castello, M. & Gallant, J. pymoten: scientific python package for computing motion energy features from video. Preprint at https://doi.org/https://doi.org/10.5281/zenodo.6349625 (2021).
103. Schütt, H. H., Kipnis, A. D., Diedrichsen, J. & Kriegeskorte, N. Statistical inference on representational geometries. *Elife* **12**, 1–46 (2023).
104. Nichols, T. E. & Holmes, A. P. Nonparametric Permutation Tests For Functional Neuroimaging : A Primer with Examples. *Hum Brain Mapp* **25**, 1–25 (2001).
105. Singh, K. D., Barnes, G. R. & Hillebrand, A. Group imaging of task-related changes in cortical synchronisation using nonparametric permutation testing. *Neuroimage* **19**, 1589–1601 (2003).
106. Nili, H. *et al.* A Toolbox for Representational Similarity Analysis. *PLoS Comput Biol* **10**, e1003553 (2014).
107. Groen, I. I. *et al.* Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *Elife* **7**, e32962 (2018).