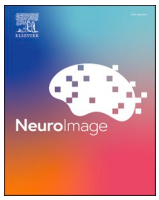
NeuroImage 319 (2025) 121439

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg



Neural representation of action features across sensory modalities: A multimodal fMRI study

Laura Marras [a,1], Lorenzo Teresi [a,1], Francesca Simonelli [a,1], Francesca Setti [a], Alessandro Ingenito [a,b], Giacomo Handjaras [a,*], Emiliano Ricciardi [a]

[a] MoMiLab, IMT School for Advanced Studies Lucca, Lucca, Italy
[b] International School of Advanced Studies, University of Camerino, Camerino, Italy

ARTICLE INFO

Keywords:
Action representation
fMRI
Canonical correlation analysis
Naturalistic stimulation
Sensory modality

ABSTRACT

Action representation and the sharing of feature coding within the Action Observation Network (AON) remain debated, and our understanding of how the brain consistently encodes action features across sensory modalities under variable, naturalistic conditions is still limited. Here, we introduce a theoretically-based taxonomic model of action representation that categorizes action-related features into six conceptual domains: Space, Effector, Agent & Object, Social, Emotion, and Linguistic. We assessed the predictive power of this model on human brain activity by acquiring functional MRI (fMRI) data from participants exposed to audiovisual, visual-only, or auditory-only versions of the same naturalistic movie. Using a multi-voxel encoding analysis and variance partitioning, we demonstrated that our model significantly predicts cortical activity within the AON, with a comparable effect size across modalities. The Effector and Social domains contributed most to the model predictions and domain-specific representations were largely stable across sensory modalities. This study elucidates how the human brain robustly encodes action-related information across different sensory modalities, revealing that certain action domains have a stronger influence on neural representation in a modality-general manner. Overall, this research enhances our understanding of how the brain integrates complex action information from multiple sensory inputs, offering insights into the generalized nature of action representation in human cognition and paving the way for further exploration into multisensory integration.

1. Introduction

Action processing in neuroscience explores how the brain encodes, processes, and retrieves information related to actions (Grafton and Hamilton, 2007; Kilner, 2011; Giese and Rizzolatti, 2015). In the last decades, particular interest has arisen around the role of the so-called action representation or action observation network (AON) which is believed to play a crucial role in understanding and interpreting observed actions (Decety and Grèzes, 1999; Gallese et al., 1996; Rizzolatti and Craighero, 2004). This network spans a wide extent of the cortical mantle and comprises distant, yet functionally interconnected regions, extending from the inferior ventral and dorsal premotor cortex (vPMC and dPMC) to the bilateral occipitotemporal (LOTC) and the parietal cortices.

Various theoretical frameworks have been proposed in the literature to explain the mechanisms underlying action representation in the human brain, each contributing unique insights into the cognitive and neural mechanisms involved. Evidence suggests that the human brain encodes the specific features contributing to action recognition (e.g., kinematics, object-centered goals, motor acts) through a hierarchical and distributed organization (Grafton and Hamilton, 2007; Kilner, 2011). This existing literature has primarily explored the representation of distinct sets of features - e.g., effector-target interaction (Beurze et al., 2007), target-agent identity (Chambon et al., 2014), social-emotional valence (De Gelder and Van den Stock, 2011). However, the emergence of high-level conceptual representations of action, and the degree to which feature coding is shared across constituent regions of the AON remain subjects of ongoing debate (Simonelli et al., 2024).

Furthermore, while much of the existing literature has focused on the visuomotor processing of actions, the ability of the AON to encode the properties of actions across different sensory modalities is still of particular interest, as it may reveal fundamental principles of neural

* Corresponding author at: Piazza San Francesco, 19, Lucca (LU), 55100, Italy.
E-mail address: giacomo.handjaras@imtlucca.it (G. Handjaras).
[1] these authors contributed equally

https://doi.org/10.1016/j.neuroimage.2025.121439
Received 13 March 2025; Received in revised form 31 July 2025; Accepted 30 August 2025
Available online 31 August 2025

representation and cognitive processing (Beauchamp, 2005). Recent studies have suggested that the AON exhibits a degree of invariance to the modality of stimulus presentation, indicating that the brain can extract and represent action-related information regardless of whether it is presented visually, auditorily, or through a combination of both (James et al., 2011; Kirsch and Cross, 2015; Copelli et al., 2022). In this regard, evidence exists that congruent auditory and visual stimuli enhance the activation of the AON compared to unimodal presentations (Bischoff et al., 2014) and that the integration of auditory and visual information leads to a more robust representation of actions in the brain. Although this invariance is crucial for understanding how individuals interpret and respond to actions in a dynamic environment where sensory inputs can vary widely, the extent to which the encoding of distinct action features is stable across sensory modalities remains poorly understood.

Here, we assessed whether and to what extent distinct features of action representation are processed under naturalistic conditions across different sensory modalities. Taking advantage of the rich literature based on behavioral and functional studies (e.g., Van Elk et al., 2014; Kemmerer, 2021; Kabulska and Lingnau, 2023), we purposely defined a comprehensive taxonomic model for action representation that allows a detailed categorization of action-related events even in naturalistic conditions, when action features are often interacting and overlapping. Specifically, a set of six domains of action features - Space, Effector, Agent & Object, Social, Emotion, and Linguistic - was defined a priori encompassing multiple action features in terms of their context, execution, and underlying dynamics (Fig. 1A and Table 1). Furthermore, we acquired brain responses using functional Magnetic Resonance Imaging (fMRI) in three groups of participants exposed to the multimodal audiovisual, the visual-only, or the auditory-only versions of the live-action movie 101 Dalmatians (Herek, 1996). Our aim was twofold: first, we tested the ability of our model to predict fMRI activity by measuring the impact of each action domain through a multi-voxel encoding analysis (Haynes, 2015; Naselaris et al., 2011); second, we determined whether the representation of action domains was comparable across sensory modalities within the AON. Importantly, here we adopted a variance partitioning approach to disentangle the unique contribution of action features, whereas previous research on action representation has largely focused on individual dimensions in isolation (Kemmerer, 2021), often overlooking potential confounds due to collinearity among action categories.

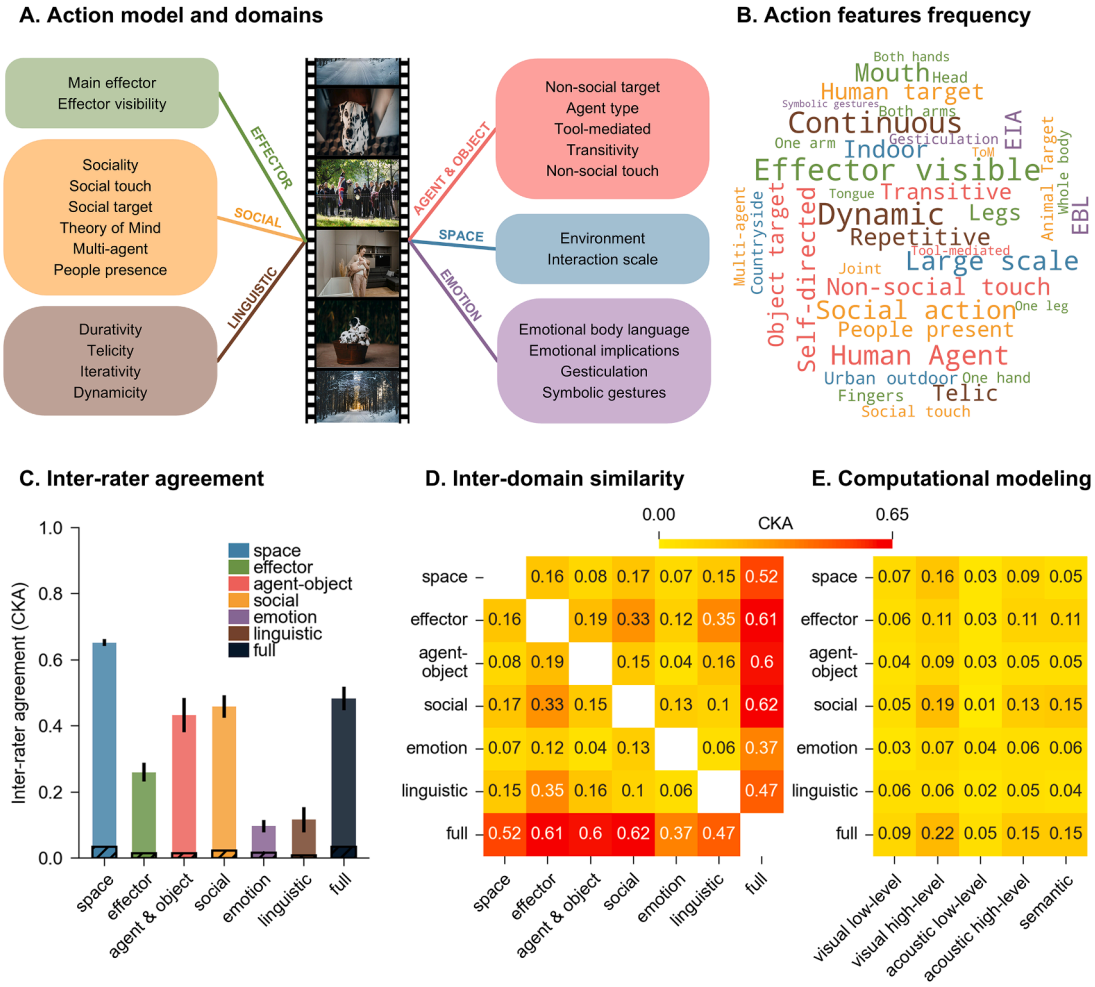


Fig. 1. Taxonomic action model. (A) Schematic representation of the six action domains (*Space, Effector, Agent & Object, Social, Emotion, and Linguistic*) and corresponding features used to annotate actions of the movie stimulus. (B) Frequency of action features occurrence in the stimulus. Features are color-coded to denote the domain they pertain to, with a larger font size indicating a greater frequency. *EBL*: Emotional Body Language. *EIA*: Emotional Implication of Action. *ToM*: Theory of Mind. (C) Mean Inter-rater agreement across domains and for the full model. Error bars indicate standard deviation. Dashed bars represent the 99th percentile of the null distribution obtained via permutation testing ($p < 0.01$). (D) Pairwise similarity between domains and the full model, indicating the amount of shared variability. (E) Similarity between model domains and computational models, measuring the extent to which the action model captures low- and high-level perceptual and semantic information in the stimulus. Visual low-level: motion-energy model; visual high-level: ReLU6 layer of the VGG-19 network; acoustic low-level: spectral descriptors; acoustic high-level: ReLU5.1 layer of the VGGish network; semantic: GPT-4 word embeddings.

Table 1
Characterization of the features and domains of the taxonomic action model.

Domain	Feature	Annotation	References	Annotation Example
Space	*Environment*: the type of environment where the action occurs.	Indoor Urban Outdoor Countryside Outdoor	Dima et al. (2022)	"Dog sniffs the ground": Countryside Outdoor "Girls jogging": Urban Outdoor "Man hits friend with a stick": Indoor
	Interaction Scale: the magnitude of movement in the space required to complete the action.	Minimal / Extensive	Tarhan and Konkle (2020)	"Dog sniffs the ground": Minimal "Girls jogging": Extensive "Man hits friend with a stick": Extensive
Effector	*Main Effector*: the primary body part involved in performing the action.	One hand Both hands One arm Both arms One leg or paw All legs or paws Mouth Head or nose Tongue Whole body Fingers	Tarhan and Konkle (2020)	"Dog sniffs the ground": Head or nose "Girls jogging": Whole body "Man hits friend with a stick": Both arms
	Effector Visibility: whether the effector is visible or hidden from the observer's view.	Visible / Not Visible	Tarhan and Konkle (2020)	"Dog sniffs the ground": Not visible "Girls jogging": Visible "Man hits friend with a stick": Visible
Agent & Object	*Agent Type*: whether the agent performing the action is human or non-human.	Human / Non-Human	Haxby et al. (2020)	"Dog sniffs the ground": Non-Human "Girls jogging": Human "Man hits friend with a stick": Human
	Non-social Action Target: whether the action is directed to an object or to the self.	Object Self	Tarhan and Konkle (2020)	"Dog sniffs the ground": Object "Girls jogging": Self "Man hits friend with a stick": Social
	Tool-Mediated: whether the action requires the use of tools.	Yes / No	Gallivan et al. (2013)	"Dog sniffs the ground": No "Girls jogging": No "Man hits friend with a stick": Yes
	Transitivity: whether the action involves interaction with inanimate objects.	Yes / No	Wurm et al. (2017)	"Dog sniffs the ground": Yes "Girls jogging": No "Man hits friend with a stick": Yes
	Non-Social Touch: whether the agent makes physical contact with an object or self during the action.	Yes / No	Masson and Isik (2021)	"Dog sniffs the ground": Yes "Girls jogging": No "Man hits friend with a stick": Yes
Social	*Sociality*: whether the action involves social interaction with another individual (human or animal).	Social / Non-Social	Wurm et al. (2017)	"Dog sniffs the ground": Non-social "Girls jogging": Yes "Man hits friend with a stick": Yes
	Social Action Target: whether the action is directed to another individual (human or animal).	Human Animal	Tarhan and Konkle (2020)	"Dog sniffs the ground": non-social target "Girls jogging": Non-social target "Man hits friend with a stick": Human
	Multi-Agent: whether the action is performed by multiple agents in a coordinated or concurrent manner.	Single-agent Multiple agents concurrent Joint actions	Sebanz et al. (2006); Sinigaglia and Butterfill (2020)	"Dog sniffs the ground": single-agent "Girls jogging": Multiple agents concurrent "Man hits friend with a stick": Single-agent
	People Present: whether there are >2 people present in the scene.	Yes / No	Dima et al. (2023)	"Dog sniffs the ground": No "Girls jogging": Yes "Man hits friend with a stick": No
	Theory of Mind: whether the action involves inferring another individual's mental state.	Yes / No	Masson and Isik (2021)	"Dog sniffs the ground": No "Girls jogging": No "Man hits friend with a stick": Yes
	Social Touch: whether the agent makes physical contact with an individual (human or animal) during the action.	Yes / No	Masson and Isik (2021)	"Dog sniffs the ground": No "Girls jogging": Yes "Man hits friend with a stick": Yes
Emotion	*Emotional Body Language*: whether the action expresses emotions through body language.	Yes / No	Barliya et al. (2013); Tipper et al. (2015)	"Dog sniffs the ground": No "Girls jogging": No "Man hits friend with a stick": Yes
	Emotional Implications: whether the action has emotional significance due to its narrative or context.	Yes / No	Goldberg et al. (2014)	"Dog sniffs the ground": No "Girls jogging": No "Man hits friend with a stick": Yes

(*continued on next page*)

Table 1 (*continued*)

Domain	Feature	Annotation	References	Annotation Example
Linguistic	*Gesticulation*: whether the actor gesticulates during the action.	Yes / No	Schippers et al. (2010)	"Dog sniffs the ground": No "Girls jogging": No "Man hits friend with a stick": No
	Symbolic Gesture: whether the action executes culturally specific gestures with symbolic meaning (e.g., handshakes, salutes).	Yes / No	Schippers et al. (2010)	"Dog sniffs the ground": No "Girls jogging": No "Man hits friend with a stick": No
	Iterativity: whether the action involves repetitive cycles or a single occurrence.	Repetitive / Single	Xu and Huang (2013)	"Dog sniffs the ground": Repetitive "Girls jogging": Repetitive "Man hits friend with a stick": Single
	Dynamicity: the degree of motion involved in the action.	Dynamic / Static	Zarcone and Lenci (2010)	"Dog sniffs the ground": Static "Girls jogging": Dynamic "Man hits friend with a stick": Dynamic
	Durativity: whether the action is continuous (e.g., speaking) or instantaneous and bounded (e.g., striking).	Continuous / Bounded	Zarcone and Lenci (2010)	"Dog sniffs the ground": Bounded "Girls jogging": Continuous "Man hits friend with a stick": Bounded
	Telicity: whether the action has a clear and defined endpoint.	Telic / Atelic	Zarcone and Lenci (2010)	"Dog sniffs the ground": Atelic "Girls jogging": Atelic "Man hits friend with a stick": Telic

2. Methods

In this study, we investigated the neural encoding of action features across sensory modalities by applying a comprehensive taxonomic model to fMRI data collected during naturalistic stimulation. We analyzed part of a naturalistic fMRI dataset from our previous study (Setti et al., 2023). Thirty participants were divided into three groups and exposed to the audiovisual, visual-only, or auditory-only versions of the same movie. Action events in the stimuli were identified and annotated by three independent raters, according to an ad-hoc action model consisting of six conceptual domains: Space, Effector, Agent & Object, Social, Emotion, and Linguistic. The final model was validated through inter-rater agreement and served as the final input for the multivariate encoding fMRI analysis, offering a detailed and structured representation of action-related information from the movie stimulus as perceived consistently across multiple raters. First, a full-model Canonical Correlation Analysis (CCA) assessed how well the entire action model predicted brain activity, quantifying the proportion of variance explained within cortical regions of interest. Second, a variance partitioning approach measured the unique contribution of each domain. This framework allowed us to disentangle domain-specific and shared effects in brain representations of action features across experimental conditions.

2.1. Participants

Thirty healthy and typically developing participants were assigned to one of three experimental conditions consisting of three versions of naturalistic stimulation: audiovisual (AV) ($N = 10$, 35 ± 13 years, 8 females); visual-only (V) ($N = 10$, 37 ± 15 years, 5 females); or auditory-only (A) ($N = 10$, 39 ± 17 years, 7 females). All subjects were right-handed and native Italian speakers. Participants had no history of neurological or psychiatric conditions, normal hearing, normal or corrected vision, and were drug-free.

Each participant was instructed about the nature of the research and gave written informed consent. The study was approved by the Ethical Committee of the University of Turin (protocol number 195874/2019) and conforms to the Declaration of Helsinki.

2.2. Stimuli and experimental conditions

Naturalistic stimulation consisted of the presentation of a shortened and edited version (~54 min) of the movie "101 Dalmatians" (Herek, 1996), split into six runs (~9 min each).

The stimulation was delivered in three different versions: (1) audiovisual (AV), including both visual and auditory features of the movie stimulus; (2) visual (V) features only with no auditory stimulation; (3) auditory (A) features only with no visual stimulation.

In the A and AV versions, an Italian voice-over serving as a narrator's verbal description of the movie was superimposed. The audio description included all the aspects of the visual scenery that can not be conveyed through dialogue, music, or environmental sounds. Likewise, in the V and AV versions, Italian subtitles transcribing the whole soundscape (dialogues, narrator's voice-over, environmental sounds) were added. For further details on the editing procedure, we refer the reader to the original paper (Setti et al., 2023).

In all conditions, participants were instructed to simply enjoy the movie, and in the A condition, participants were instructed to keep their eyes closed.

2.3. Taxonomic action model

2.3.1. Features selection and domains creation

To create a detailed and systematic representation of action-related events, the taxonomic model was constructed through an iterative process that integrated multiple conceptual frameworks ranging from embodied cognition to ecological psychology and action semantics. We first identified candidate features based on their relevance to action perception and representation in the brain, as supported by prior studies (e.g., Tarhan and Konkle, 2020; Haxby et al., 2020; Wurm et al., 2017; see Table 1 and Fig. 1A). Each feature was operationalized to capture specific aspects of actions.

The selected features were grouped into six domains to minimize overlap and ensure orthogonality: *Space, Effector, Agent & Object, Social, Emotion*, and *Linguistic*. This organization reflects both theoretical coherence and empirical evidence suggesting distinct neural representations for these feature sets. These domains were derived from a comprehensive review of existing literature on action representation, combining theoretical insights and practical considerations to ensure

their relevance to naturalistic stimuli. Each domain includes features that define specific dimensions of actions, allowing their characterization at multiple levels of granularity.

The *Space* domain refers to the type of environment where an action occurs and the extent of movement required to complete it. It captures the environment's nature - e.g., indoor, urban outdoor, or countryside outdoor (Dima et al., 2022) - and quantifies the interaction scale, defined by the movement scope of the actor within the scene (Tarhan and Konkle, 2020). Notably, the interaction scale was included in the *Space* domain as the spatial extent of an action often covaries with the scene it is set in. The *Effector* domain characterizes the main body part involved in performing the action and whether this effector is visible or hidden from the observer's view during the action (Tarhan and Konkle, 2020). The *Agent & Object* domain addresses the attributes of the action's agent and its non-social target. It distinguishes actions based on whether the agent is human or non-human (Haxby et al., 2020) and whether the action is directed toward an object or the self (Tarhan and Konkle, 2020). This domain also identifies whether the action involves interactions or physical contact with inanimate objects (Masson and Isik, 2021) and whether it requires the use of tools (Gallivan et al., 2013). The *Social* domain includes a set of features related mainly to human or animal interactions (Wurm et al., 2017). In particular, this domain specifies whether the action has a social target (Tarhan and Konkle, 2020), i.e., directed toward another individual (human or animal), or involves interaction or physical contact with another agent (Masson and Isik, 2021). Furthermore, it considers the number of individuals visible in the scene (Dima et al., 2023), whether the action is performed by multiple agents in a coordinated or concurrent fashion (Sebanz et al., 2006; Sinigaglia and Butterfill, 2020), and, finally, whether the action requires inferring another's mental state (Masson and Isik, 2021). The *Emotion* domain describes the emotional dimensions of actions, including the presence of emotional implications of the action (Goldberg et al., 2014), the use of gesticulation and symbolic gestures (Schippers et al., 2010), and whether the action expresses portrayed emotions through body language, conveyed for instance by movement velocity, acceleration, pitch in the voice or urge (Barliya et al., 2013; Tipper et al., 2015). Lastly, the *Linguistic* domain characterizes actions through features derived from verbal descriptors, focusing on semantic properties that contribute to their representation. Specifically, these features were borrowed from experimental linguistics, specifically from Vendler's classification of event types (Vendler, 1967), and capture attributes of actions as conveyed by the meaning of the action verb. The linguistic domain includes durativity (the temporal extent of an action), telicity (whether an action has a clear and intrinsically defined endpoint), dynamicity (the degree of motion involved in an action - Zarcone and Lenci, 2010), and iterativity (whether the action consists of repetitive cycles - Xu and Huang, 2013). Altogether, these domains provide a structured taxonomic model, grounded in the literature, along with a comprehensive set of descriptors for classifying action-related events.

2.3.2. Stimulus feature space

Based on our taxonomic model of actions, we created a stimulus feature space to capture detailed action-related aspects present in the movie stimulus. For this purpose, three raters (1F, average age 27) were recruited and instructed to watch the movie in the AV condition, identify each action in the naturalistic stimulus, and describe it in terms of the taxonomic model. Each rater began with segmenting the stimulus into discrete events. This involved independently detecting individual actions and marking their specific onset and offset times with a temporal resolution of 0.1 s. Raters were instructed to annotate as many actions as they could detect, including both primary actions - which occur predominantly in the foreground of a scene - as well as secondary actions, which may (co)-occur in the background. To preserve the ecological validity of the naturalistic stimulation, no specific instructions were given regarding the level of granularity for defining actions. Rather, the raters were free to apply their criteria, ensuring the annotations reflected their subjective interpretations of the action boundaries. For each identified action, the raters then evaluated each of the 23 features of the model by manually annotating the corresponding value. As a result, we obtained a matrix of actions by features for each rater.

2.3.3. Model preprocessing

These matrices were then preprocessed to create a single, time-resolved feature space. First, we converted categorical annotations of individual features into a binary format by applying one-hot encoding. Specifically, each multi-level feature was decomposed into n binary features, where n is the number of possible annotations for that feature. Of the 23 original conceptual features, 18 were originally binary, while the remaining 5 features were multi-level, with 11, 3, 2, 2, and 2 levels. After applying one-hot encoding, these five features contributed a total of 20 binary features, yielding a final set of 38 predictors.

We then transformed the datasets from an event-based to a time-resolved format. Specifically, based on each action onset and offset, we mapped the respective features to a continuous time sequence with 0.1 s resolution, obtaining a time by features matrix for each rater.

The resulting time-based feature matrices from individual raters were aggregated to create a group-level model. To ensure consistency among raters, a value of 1 was assigned to a feature at each time point only if at least two out of three raters agreed on its presence. The group-level model was then down-sampled to match the temporal resolution of the fMRI data (2 s). To account for events shorter than the fMRI sampling interval, each feature was assigned a value of 1 at each time point if it was present in any sample within the 2-second bin. For the multivariate encoding procedure, each column of the model was convolved with a canonical hemodynamic response function (gamma function, duration $= 12$ s, $p = 8.6$, $q = 0.547$) to account for the delayed hemodynamic response of the fMRI signal.

2.3.4. Inter-rater agreement

To verify the consistency of feature annotations, we quantified inter-rater agreement across individual domains and the full model. We employed Centered Kernel Alignment (CKA) (Kornblith et al., 2019) with a debiasing step (Murphy et al., 2024) as a similarity index. CKA measures the shared variance between two multi-dimensional feature spaces, providing a robust method for quantifying similarity between datasets of different size. Debiasing was applied to account for potential confounds arising from significant differences in matrix ranks (i.e., domains' dimensionality ranged from 4 to 12, with the full model comprising 38 descriptors), ensuring a more reliable assessment of similarity.

For each rater pair, we computed CKA scores for individual domains and the full model using their respective downsampled (i.e., 2-seconds temporal window) binary matrices. To establish statistical significance, a permutation test was implemented: a null distribution ($n = 1000$) was obtained by randomly shuffling chunks of 15 timepoints of one rater's matrix prior to CKA estimation, to account for temporal autocorrelation in the data. The observed similarity scores, averaged across rater pairs, were then compared against the 99^{th} percentile of the null distribution to identify significant agreements ($p < 0.01$). This analysis ensured that the group model accurately reflected consistent action feature annotations across raters.

2.3.5. Inter-domain similarity

To evaluate the relationship between domains, we assessed how features from different domains uniquely or redundantly captured aspects of the described actions. This was accomplished by calculating pairwise similarity through CKA between all domains of the down-sampled binary group-level model.

2.3.6. Computational modeling

As a final step, we sought to determine the extent to which perceptual or semantic information from the stimulus was captured by the

features of the action model. To achieve this, we leveraged a series of computational models to extract diverse representations of the movie stimulus, encompassing: 1) low-level visual descriptors (i.e., motion energy derived from the spatiotemporal integration of Gabor-like filters); 2) high-level visual descriptors (i.e., VGG-19 convolutional deep neural network architecture; Simonyan and Zisserman, 2014); 3) low-level auditory descriptors (spectral and envelope-based properties capturing frequency modulations); 4) high-level auditory descriptors (VGGish deep neural network architecture; Hershey et al., 2017); 5) a high-level semantic representation derived from GPT-4 embeddings (Achiam et al., 2023) extracted from the movie's subtitles.

In detail, low-level visual descriptors were obtained for each two-second segment of the movie videoclip using a comprehensive set of 4715 motion energy descriptors derived from space-time Gabor filters. These filters included Gabor wavelets varying in spatial frequencies, orientations, and integrated across three distinct temporal frequencies: 0 Hz (static energy), 2 Hz, and 4 Hz (Nishimoto et al. 2011). Regarding high-level visual descriptors, we utilized the VGG-19 (Simonyan and Zisserman, 2014) convolutional deep neural network architecture to extract a comprehensive set of 4096 visual properties from the central frame of each two-second segment. Specifically, we used the output from ReLU6, the final layer in the stack of convolutional layers, which captures high-level visual information essential for object recognition and image classification. To model low-level auditory characteristics, we extracted for each two-second segment 449 spectral descriptors in the 0–15k Hz frequency range, following the methodology outlined by de Heer and colleagues (2017). Instead, for high-level auditory descriptors, we employed the VGGish (Hershey et al., 2017) model, a convolutional neural network based on the VGGNet architecture and adapted for audio classification. Specifically, we extracted the output of the ReLU5.1 layer, obtaining 4096 descriptors for each two-second segment. These features captured more abstract properties of the movie's audio track, including contextual information about the soundscape, such as the presence of background noise, music, or speech. Finally, regarding semantic descriptors, we extracted contextual word embeddings from each subtitle sentence using the pre-trained GPT-4 model (Achiam et al., 2023) (i.e., text-embedding-3-small) via the OpenAI API (https://openai.com/). This process generated a 1536-dimensional vector for each two-second segment.

We then compared the representations of individual domains as well as the full model to these computational representations using CKA as a similarity measure. This allowed us to evaluate whether specific domains aligned more closely with perceptual or semantic features, offering insights into the nature of the information encoded by each domain.

2.4. fMRI data acquisition and preprocessing

Structural and functional data were acquired in the same session with a Philips 3T Ingenia scanner equipped with a 32-channel head coil. For anatomical images, a magnetization-prepared rapid gradient echo sequence use employed (TR = 7 ms; TE = 3.2 ms; FA = 9°; FOV = 224 mm, acquisition matrix =224 × 224; slice thickness = 1 mm; voxel size 1 × 1 × 1 mm; 156 sagittal slices). Functional images were acquired using gradient recall echo planar imaging (GRE-EPI; TR = 2000 ms; TE = 30 ms; FA = 75°; FOV = 240 mm; acquisition matrix (in-plane resolution) = 80 × 80; acquisition slice thickness = 3 mm; acquisition voxel size =3 × 3 × 3 mm; reconstruction voxel size = 3 × 3 × 3 mm; 38 sequential axial ascending slices; total volumes 1614 for the six runs of the movie).

Acquired fMRI data were preprocessed with the AFNI 17.1.12 software package (Cox, 1996), following the standard steps. First, scanner-related noise was corrected through spike removal (3dDespike). Then, all volumes underwent run-wise temporal realignment (3dTshift) and head motion correction using as base the first run (3dvolreg). Spatial smoothing was then performed with a Gaussian kernel (3dBlurToFWHM, 6 mm, full width at half maximum), followed by run-wise percentage normalization. Next, in order to smooth time series and remove unwanted trends and outliers, the normalized runs underwent detrending through Savitzky-Golay filtering (sgolayfilt, polynomial order = 3, frame length = 200 timepoints) in MATLAB R2019b (MathWorks Inc., Natick, MA, USA) and were concatenated into a single time series.

Afterward, signals related to head motion parameters and framewise displacement were regressed-out through multiple regression analysis (3dDeconvolve). Lastly, the anatomical volumes were aligned to the reference functional image and non-linearly registered (3dQWarp) to the MNI-152 standard space (final voxel size 3 mm iso; Fonov et al., 2009).

2.5. fMRI data analysis

To investigate the relationship between action features and brain activity, we employed a variance partitioning framework based on multivariate analysis, using a region of interest (ROI)-based approach.

2.5.1. Parcellation and PCA

First, fMRI data were parcellated into 200 distinct cortical ROIs using Schaefer's atlas (Schaefer et al., 2018). The parcellation, including grey matter cortical voxels only, was applied in the original anatomical space of each subject.

To measure the association between the action model and brain activity, we employed a multivariate encoding procedure based on CCA. Since CCA extracts canonical components that maximize the linear association between two matrices, the algorithm limits the dimensionality of the final components to the lower rank of the input matrices. Thus, to accurately reconstruct the action model and account for variability in the number of voxels across ROIs (ranging from 21 to 492; see Supplementary Fig. S1), we set the dimensionality of the ROIs to match the rank of the action model. Specifically, for each ROI, principal component analysis (PCA) was performed voxel-wise, retaining a number of components equal to the 38 predictors in the action model. To address cases where an ROI contained fewer than 38 voxels, additional voxels were added by replicating the average time series across existing voxels within that ROI. This approach ensured consistent dimensionality across ROIs and subjects while preserving the minimum number of components required for subsequent analyses, without excluding brain regions or altering the original informational content of the ROI signals. Importantly, while the amount of variance explained by the resulting 38 components varied across ROIs (see Supplementary Fig. S1), it consistently exceeded 76 % for all subjects and conditions.

2.5.2. Full-model canonical correlation analysis

Then, to quantify how well the predictors in the action model explained brain activity within each ROI, we employed CCA. CCA is a multivariate statistical technique designed to measure linear relationships between multidimensional variables (Hotelling, 1992; Bilenko and Gallant, 2016). Specifically, CCA finds pairs of canonical components - linear projections of each dataset - that maximize the correlation between the two datasets in a shared canonical space.

In our analysis, CCA was iteratively applied to each ROI to relate the action model (X, time points × 38 features) to the brain components (Y, time points × 38 components). This process yielded two sets of canonical components U and V, along with the canonical coefficients A and B, for X and Y, respectively. These were used to compute the proportion of variance in Y that could be predicted by X, through the following steps. First, we used U to predict the canonical projections of V (V_{pred}) by fitting a set of coefficients (W), obtained through least squares regression mapping $U \rightarrow V$ (Eq. (1)).

$$V_{pred} = U \cdot W \tag{1}$$

Next, we reconstructed predicted brain activity (Y_{pred}) by mapping back the predicted canonical components (V_{pred}) to the original brain

activity space, using the canonical coefficients B derived during the CCA process (Eq. (2)).

$$Y_{pred} = V_{pred} \cdot B \quad (2)$$

Finally, we calculated the coefficient of determination R^2, which quantifies the proportion of variance in the observed Y that was explained by the reconstructed Y (Y_{pred}) (Eqs. (3),4,5).

$$SS_{residual} = \sum (Y - Y_{pred})^2 \quad (3)$$

$$SS_{total} = \sum (Y - \overline{Y})^2 \quad (4)$$

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}} \quad (5)$$

Here, R^2 represents the proportion of variance in the brain activity of a given ROI that can be explained by the predictors of the action model.

2.5.3. Statistical significance and permutation tests

To assess the significance of R^2, we generated null distributions by permuting the temporal order of the fMRI data. For each ROI, the time points of the brain components (in chunks of contiguous 30 s, to account for temporal autocorrelation of the signal) were shuffled randomly across 1000 iterations while keeping the structure of the action model intact. For each permutation, the CCA procedure was repeated, generating a null distribution of R^2 values for each ROI. Each of these R^2 values was then compared against the null distribution to determine their *p*-value. Once this process was done for all subjects, *p*-values were combined across subjects through Fisher's sum (Fisher, 1970) (Eq. (6)), yielding a null distribution of combined statistics for each ROI.

$$\chi^2_{2k} = -2\sum_{i=1}^{k} ln(p_i) \quad (6)$$

The group-level *p*-value for each ROI was obtained by comparing the observed combined statistics with the relative null distribution, using Pareto tail approximation (Winkler et al., 2016) and t-max correction for multiple comparisons ($p_{corr} < 0.05$; Westfall and Young, 1993). Specifically, we obtained the t-max distribution by taking the maximum value across all ROIs, for each permutation. Then, we modelled the right tail (90^{th} percentile) of the t-max distribution by fitting a generalized Pareto distribution (Winkler et al., 2016; Pickands, 1975), from which we derived the *p*-values.

First, this process was done for all subjects of the AV condition. Then the same pipeline was repeated in the A and V conditions, this time including only ROIs that obtained a significant R^2 ($p_{corr} < 0.05$) in the AV condition. As a result, we obtained a comprehensive map of the task-related variance in brain activity for each experimental condition, highlighting cortical regions where brain activity was significantly predicted by the full action model.

2.5.4. Variance partitioning

To disentangle the contributions of individual domains in the action model, a variance partitioning approach was employed. This method quantified the unique explanatory power of each domain by measuring the drop in R^2 when its predictors were shuffled.

For each domain, the corresponding predictors were shuffled across timepoints while all other domains were left intact. The partially shuffled action model was then used in the CCA to estimate R^2, reflecting the model's explanatory power after disrupting the specific domain of interest. To obtain a stable estimate of the impact of shuffling, we repeated the process 50 times for each domain and took the average R^2 from the 50 shuffled models. Lastly, the shuffled R^2 was subtracted from the full-model R^2, obtaining a domain-specific R^2. These calculations were performed for each domain, ROI, subject, and condition, generating domain-specific maps of unique variance contributions to brain activity.

We then computed the average across subjects to obtain a group R^2 and, for each domain, the regression coefficient between conditions and the Spearman correlation coefficients with associated *p*-values were calculated pairwise.

2.5.5. Cortical representation of domains

To evaluate the cortical representation of each action domain, we employed two complementary procedures. The first aimed to maximize the functional specificity of each brain region for a given domain, while the second assessed the consistency of domain-related representations by measuring the overlap of the highest-associated brain regions across sensory modalities.

The first method relied on a non-parametric rank-based approach and tested whether the representation of specific domains within individual ROIs was significantly greater than others, based on the distribution of R^2 values across ROIs.

For each subject and condition, R^2 values were ranked within each domain across all ROIs, identifying regions where the domain exhibited the highest and lowest explanatory power. Converting R^2 values into ranks allowed us to assess the relative, rather than absolute, representation of each domain across ROIs. By focusing on rank-based comparisons, this approach mitigated the influence of global differences in R^2 magnitudes between domains, emphasizing the relative contribution of each domain within a given ROI.

To compare the rank-based representation of domains within each ROI, we applied the Wilcoxon signed-rank test, a non-parametric method for paired data. Pairwise one-tailed tests were conducted for all domain combinations within each ROI and condition. This test determined whether the ranks of one domain consistently exceeded those of another, reflecting greater relative specificity in that region. Then, *p*-values from the three conditions were aggregated across modalities using Fisher's method (Fisher, 1970). To control for multiple comparisons across ROIs and domain pairs, the aggregated *p*-values were corrected using False Discovery Rate (FDR) control ($\alpha = 0.05$). A significant result for a domain pair within an ROI indicates that one domain consistently ranks higher than the other across subjects. This implies that the ROI exhibits a relatively stronger representation of the more dominant domain, independent of absolute R^2 values. To identify the dominant domain(s) within each ROI we adopted a "maximum-takes-all" procedure. Each domain was compared based on the number of significant pairwise tests it won within the ROI. The domain (s) with the highest number of wins were flagged as dominant. This analysis produced spatial maps to reveal putative domain-specific representations, highlighting ROIs where action-related features were most prominently encoded relative to other brain regions and other features.

The second approach relied on the integration of information across sensory modalities while examining each action domain independently. Specifically, we identified brain regions that consistently ranked in the top 20 % of R^2 values for each action domain and each sensory modality. Only areas identified in all three modalities were retained, obtaining for each domain a spatial map of the top responding ROI across the audiovisual, visual-only, and auditory-only conditions.

Only the first rank-based approach was retained in the main manuscript, while the results of the second procedure are reported in Supplementary Fig. S6.

3. Results

3.1. Model validation

Our action model was constructed by selecting a comprehensive set of features from the literature on action representation and asking three raters to identify action events and then manually annotate each feature.

Stimulus segmentation to identify action occurrences was performed independently by each of the three raters, who identified 1635, 1304, and 1025 events, respectively (average event count = 1321). Event duration ranged from 0.1 to 85.7 s and averaged 3.5 s. After

downsampling to the fMRI temporal resolution and aggregating individual raters' models, the number of timepoints in which at least one action event was present was 1562 out of 1614 (Supplementary Fig. S2, Supplementary Table S3).

The 38 action features comprising the final group model had variable frequency counts (Fig. 1B), ranging from 11 timepoints (0.7 % for Symbolic gestures) to 1530 timepoints (97.95 % for Dynamicity; see Supplementary Table S1). Frequencies of *Space* features highlighted that the actions in the dataset predominantly occurred in indoor environments and at an extended spatial scale. The *Effector* was almost always visible and most actions were performed using the mouth (e.g., speaking), followed by legs/paws, and fingers. In the *Agent & Object* domain, the most frequent feature was Human Agent, and roughly half of the time, actions were targeted toward the agents themselves and involved contact with an inanimate object. Transitive and Intransitive actions were mostly balanced, while Tool-mediated actions were underrepresented. More than half of the tagged events were considered social, that is, they involved some sort of interaction between agents; of these, most were directed towards humans rather than animals, and only one-fifth were concurrently or jointly performed by multiple agents. An emotional component was frequently present in the stimulus, either conveyed through the actor's body language or implicit in the action itself. Concerning the *Linguistic* domain, the actions in the dataset were predominantly dynamic and continuous rather than static and instantaneous.

To assess the validity of the tagging procedure, the inter-rater agreement was first evaluated by measuring the similarity between individual raters' action models: considering that these were multidimensional binary matrices differing in size, Centered Kernel Alignment (CKA) was chosen as an appropriate metric. CKA is an index of multivariate similarity and ranges from 0 to 1, where 0 means the two matrices are completely dissociable and 1 denotes maximal similarity. The average CKA value across all parings of raters was 0.48 ± 0.04 for the full model, indicating an adequate level of consistency among raters (Fig. 1C). The same analysis was also performed for each domain separately (*Space*: 0.65 ± 0.01; *Effector*: 0.26 ± 0.03; *Agent & Object*: 0.43 ± 0.05; *Social*: 0.46 ± 0.03; *Emotion*: 0.1 ± 0.02; *Linguistic*: 0.12 ± 0.04). We observed the lowest similarities in the *Emotion* domain, which requires the rater to extrapolate and interpret implicit information from the scene, and the *Linguistic* domain, which captures attributes conveyed by the meaning of the action verb. Nonetheless, all CKA scores were statistically significant when compared against the null distribution ($p < 0.01$).

To ensure consistency in the final aggregated model, only annotations agreed upon by at least two out of three raters were retained.

Since multiple features from different domains may co-occur at the same timepoint, we assessed collinearity between domains by measuring the similarity structure between each domain and with the full model (Fig. 1D). CKA scores between domains varied from a minimum of 0.04 to a maximum of 0.35, with the *Effector* and *Emotion* domains exhibiting the highest and lowest similarity, respectively. CKA values between domains and the full model were 0.62 for the *Social* domain; 0.61 for *Effector*; 0.60 for *Agent & Object*; 0.53 for *Space*; 0.37 for *Emotion,* and 0.47 for the *Linguistic* domain. Hence, the adopted domains capture sufficiently unique representational content and unbiasedly contribute to the full model.

As our action model was built based on hypothesis-driven taxonomic features derived from the literature, some degree of collinearity with visual, acoustic, and verbal properties is expected. To measure the impact of these descriptors on both the full model and individual domains, we computed CKA between different computational models and the action model matrices (Fig. 1E). As for the full model, stronger correlations were observed with computational models capturing higher-level perceptual descriptors, both visual and auditory (0.22 and 0.15, respectively), and semantic descriptors (0.15), rather than low-level perceptual characteristics (0.09 for the visual and 0.05 for the auditory model). This effect was also present when considering individual domains, though it was more evident for *Space, Effector*, and *Social*, compared to the other domains.

3.2. fMRI results in the audiovisual modality

Next, the ability of our action model to predict brain activity was assessed. Since the aim was to measure the association between two sets of multivariate data, i.e., the multidimensional action model and each multi-voxel brain region (i.e., ROI), we used Canonical Correlation Analysis (CCA, Fig. 2A). From CCA, we computed the R^2 as a metric of brain-model association, which expresses the proportion of variance in ROI patterns explained by the action model (see Methods section *Full-model Canonical Correlation Analysis*). Importantly, to quantify the unique contribution of each domain, we also performed variance partitioning, thus obtaining a domain-specific R^2 for each ROI.

In the AV modality, we identified an extended network of occipital, temporal, posterior parietal, prefrontal, and right pre-central areas (127 ROIs in total out of 200 cortical parcels defined in Schaefer et al., 2018), in which the full action model could significantly explain neural activation (average $R^2 = 0.061 \pm 0.017$, $p < 0.05$, t-max correction for multiple comparisons). Posterior temporal and lateral occipital areas showed the highest R^2 values (Fig. 2B; peak $R^2 = 0.129$ in right posterior superior temporal sulcus -STS- at $x = 56$, $y = -51$, $z = 14$, LPI), consistent with previous literature denoting these brain regions as critical for action recognition (Rizzolatti and Craighero, 2004; Orban et al., 2021; Wurm and Caramazza, 2022; Karakose-Akbiyik et al., 2023).

As for the contribution of individual domains (Fig. 2C), the Effector domain had on average the highest explanatory power (0.0086 ± 0.0031, 14 % of full model R^2), followed by the Social domain (0.0059 ± 0.002, 9.6 % of full model R^2), Agent & Object (0.0041 ± 0.0017, 6.7 % of full model R^2), Linguistic (0.0034 ± 0.0019, 5.5 % of full model R^2), Space (0.0034 ± 0.0014, 5.5 % of full model R^2), and Emotion (0.0028 ± 0.0011, 4.6 % of full model R^2). Therefore, 54.1 % of the brain variance explained by the full model was not attributable to the unique contributions of single domains, but rather to shared variance across combinations of multiple domains.

3.3. Impact of sensory modality

To test the degree of invariance of identified areas to sensory modality, we replicated the same analysis in the V and A modalities, quantifying the amount of task-related variance in brain activity when subjects were presented with visual-only or auditory-only versions of the stimulus.

In the V modality, brain activity was significantly predicted by the full model in 80 ROIs out of 127 ($p < 0.05$, t-max correction, average $R^2 = 0.06 \pm 0.015$, highest $R^2 = 0.109$ in right LOTC, at $x = 47$, $y = -75$, $z = -1$, LPI), while in the A modality, only 12 ROIs survived statistical thresholding ($p < 0.05$, t-max correction, average $R2 = 0.065 \pm 0.005$; peaking in middle right STS at $x = 61$, $y = -30$, $z = -4$, LPI, $R^2 = 0.073$), reflecting a partial loss of information when one modality is neglected. The average full model R^2 across all tested ROIs ($N = 127$) for the V and A modalities was 0.055 ± 0.014 and 0.05 ± 0.009, respectively (Fig. 3B). The decrease in explanatory power as compared to the multisensory stimulation was mainly observed in frontal and primary visual areas for the V modality. The statistically significant ROIs in the A stimulation all pertained to the superior and middle temporal cortices (Fig. 3A).

Variance partitioning for the V and A modalities (Fig. 3C) revealed a similar trend in individual domain contributions as for the AV stimulation. For all three conditions, the Effector domain explained the highest amount of variance in brain activity (V: 0.0085 ± 0.0029, 14.2 % of full model R^2; A: 0.0113 ± 0.0016, 17.4 % of full model R^2), followed by the Social domain (V: 0.0051 ± 0.0017, 8.5 % of full model R^2; A: 0.0077 ± 0.0014, 11.9 % of full model R^2). Likewise, the Emotion

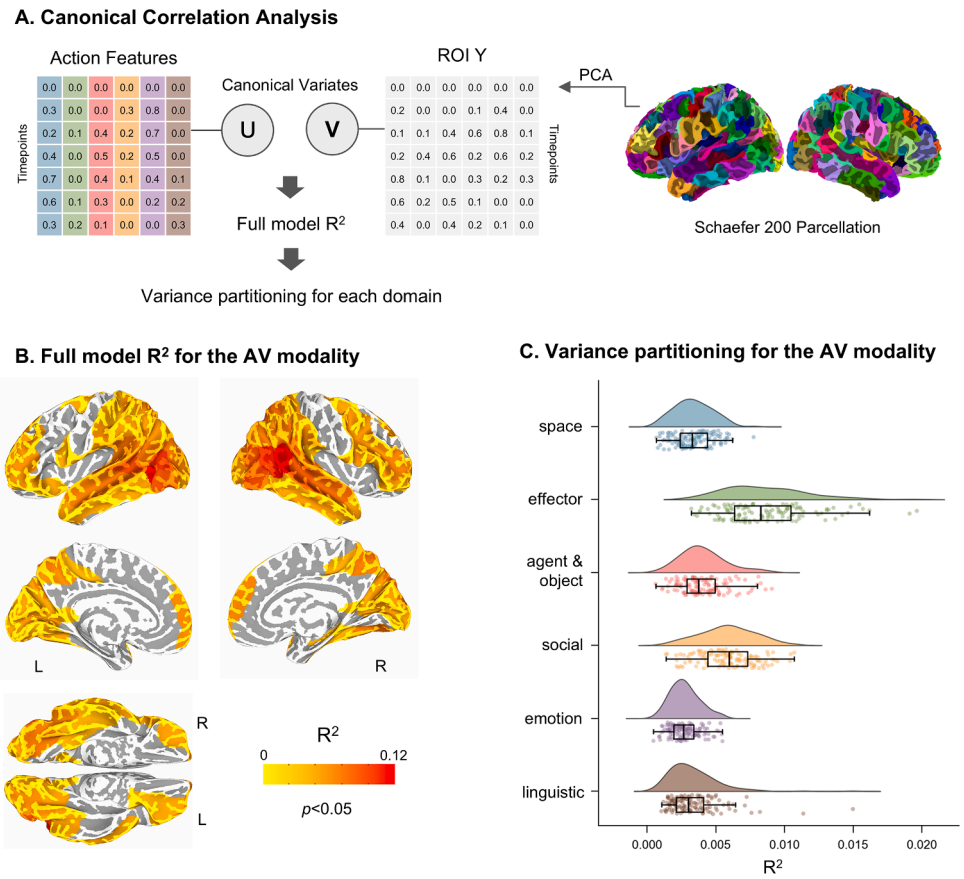


Fig. 2. Brain encoding of action features in the audiovisual modality. (A) Analysis pipeline for assessing the relationship between action features and brain activity. fMRI data were parcellated into 200 cortical regions of interest (ROIs) using Schaefer's atlas. Principal Component Analysis (PCA) was applied to voxels time series within each ROI and Canonical Correlation Analysis (CCA) was then used to map the action model onto the fMRI data, identifying canonical components that maximize the correlation between the two datasets. The proportion of variance in brain activity explained by the action model was quantified using R^2, which reflects model fit within each reduced ROI. The same process was repeated in a variance partitioning framework to isolate contributions of individual domains. (B) Brain regions showing significant ($p_{corr} < 0.05$, t-max correction) R^2 values for the full action model in the audiovisual (AV) modality. *L*: left, *R*: right. (C) Unique contributions of individual domains. Individual points reflect domain-specific R^2 values for each ROI.

domain had the least explanatory power for all three modalities (V: 0.002 ± 0.0009, 3.3 % of full model R^2; A: 0.0022 ± 0.0003, 3.4 % of full model R^2).

In the V modality, the domain-specific R^2 was 0.005 ± 0.0017 (8.3 % of full model R^2) for Agent & Object, $R^2 = 0.0028 \pm 0.0011$ (4.7 % of full model R^2) for Space, and $R^2 = 0.0027 \pm 0.0016$ (4.5 % of full model R^2) for the Linguistic domain. In the A modality, the remaining domains R^2 were: 0.0055 ± 0.0009 (8.5 % of full model R^2) for Space, 0.0035 ± 0.0007 (5.4 % of full model R^2) for Agent & Object, and 0.003 ± 0.0005 (4.6 % of full model R^2) for the Linguistic domain. Similarly to the AV modality, a substantial portion of brain variance explained by the full model (56.4 % for V and 48.8 % for A) was shared across combinations of multiple domains.

The invariance of action information organization across sensory modalities was further explored by directly comparing domain-specific R^2 distributions between modalities. For each domain, Spearman's rho was computed between ROIs R^2 of each pair of sensory modalities. The correlation was significant for all domains and modalities combinations (Table 2), indicating that information was similarly organized across the cortex for all three sensory inputs (Fig. 3D). The AV and V modalities had the most similar spatial distribution, with all domains having a correlation coefficient higher than 0.7. For the A modality, the most dissimilar domains were Emotion, whose features were predominantly expressed through visual attributes, and Linguistic.

Overall, the distribution of explained variance across cortical regions and modalities indicates that the multidimensional framework of action representation is stable across sensory inputs.

3.4. Cortical representation of domains

We observed that ROIs showing a high fit to the full model have high R^2 values for all domains, leading to overlapping spatial maps across domains. Therefore, to maximize ROIs functional specificity for a given domain, we employed a non-parametric rank-based approach: rather than comparing the absolute magnitude of domain-specific R^2, we focused on the relative contribution of each domain across regions (see Methods section *Cortical representation of domains*, Supplementary Fig. S3–5 and Supplementary Table S2). Data from all conditions were aggregated to evaluate tuning across sensory modalities. An additional analysis was also performed, which assessed the consistency of domain-related representations across sensory modalities; its results are reported in Supplementary Fig. S6. Only the results from the first analysis are described here, as we considered it a more effective approach for maximizing differences between domains.

From the non-parametric rank-based analysis, we obtained spatial maps of domain-sensitive representations, highlighting ROIs predominantly tuned to a domain versus the others, independently of the sensory modality (Fig. 4). Part of these regions were shared across multiple action domains, even though variance partitioning estimated unique domain contributions. This pattern of results implies a model of the AON characterized by distributed and overlapping representations across different features, rather than one composed of strictly domain-specific

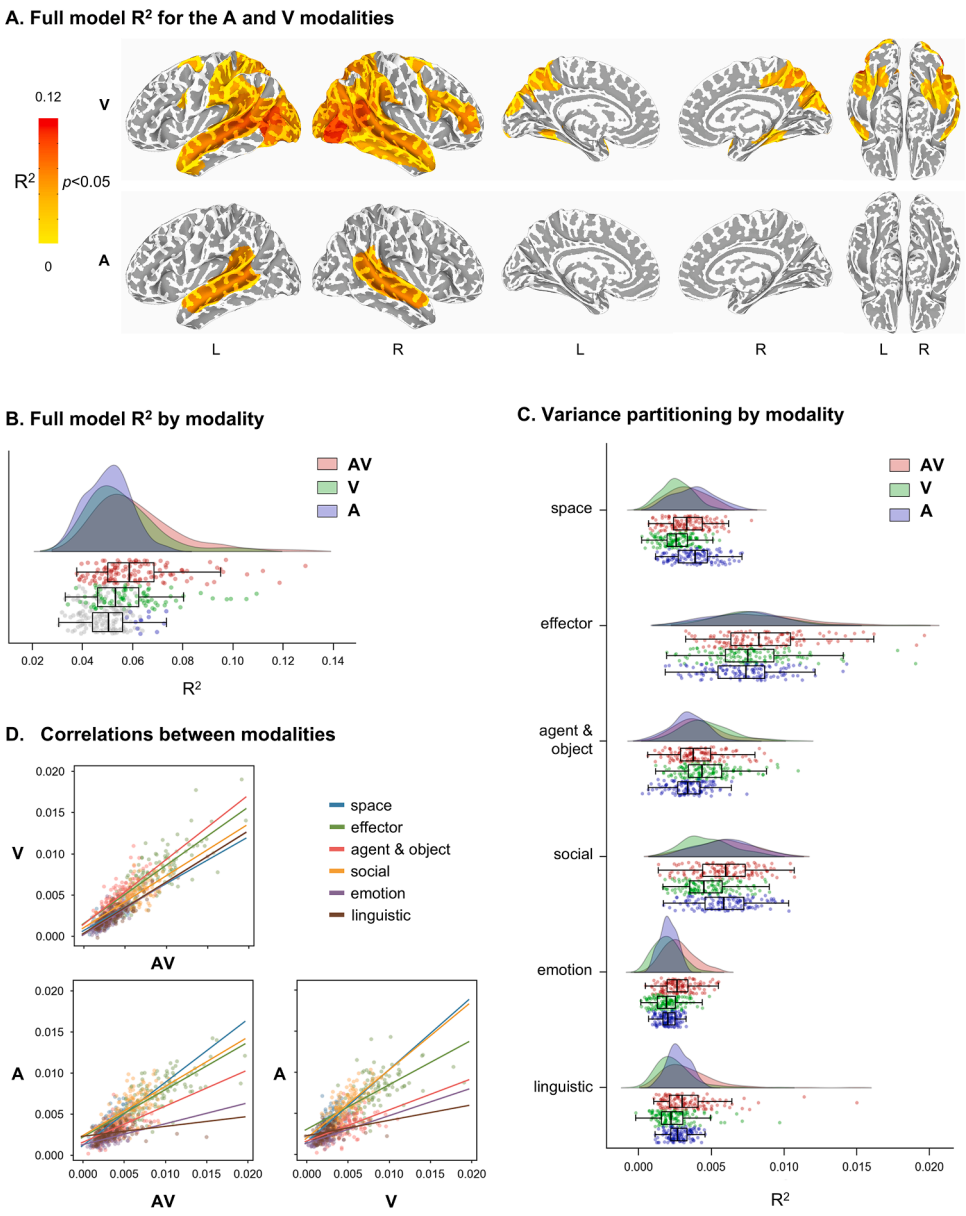


Fig. 3. Impact of sensory modality on action representation in the brain. (A) Brain regions showing significant ($p_{corr} < 0.05$, t-max correction) R^2 values for the full action model in the visual (V, top) and auditory (A, bottom) modalities. *L*: left, *R*: right. (B) Distributions of full model R^2 values across all tested ROIs for the three conditions (AV, V, A). Each point represents an ROI full model R^2 value; ROIs not surviving statistical thresholding in each modality are colored in grey. (C) Unique contributions of individual domains across sensory modalities. (D) Domains correlations between different modalities. Each point represents an individual ROI domain-specific R^2. Solid lines represent the least square fit between modalities pairings for each domain.

Table 2
Spearman correlations between modalities. ** $p < 0.001$.

	Space	*Effector*	*Agent & Object*	*Social*	*Emotion*	*Linguistic*
AV - V	0.76**	0.77**	0.78**	0.74**	0.77**	0.82**
AV - A	0.78**	0.75**	0.64**	0.63**	0.5**	0.47**
V - A	0.69**	0.67**	0.6**	0.73**	0.51**	0.49**

modules.

In the following paragraph, we report and briefly discuss the results for each domain, which expand on the current knowledge of feature specificity in AON.

3.4.1. *Space*

This domain describes the context in which the action takes place and its spatial extent. Research on scene representation has highlighted the role of the retrospenial cortex (RSC) and the parahippocampal cortex (PHC) in processing indoor vs outdoor and natural vs urban environments (Henderson et al., 2007, 2011; Stobbe et al., 2024). In the context of action representation, the setting of an action is often conflated with other perceptual features (Masson and Isik, 2021; Dima et al., 2022), contributing to brain activity in early visual areas and fusiform gyrus (Masson and Isik, 2021). A previous study investigating the representation of actions spatial scale found regions tuned to fine-scale actions in the intraparietal sulcus (IPS) extending to the transverse occipital sulcus (TOS), LOTC, and PHC; areas tuned to intermediate/near space and large interaction scale were identified in posterior portions of middle (MTG) and inferior temporal cortices (ITG), medial parietal cortex and RSC (Tarhan and Konkle, 2020).

Consistently, we showed tuning to the Space domain in inferior and middle temporal cortices. Other areas were identified in bilateral orbitofrontal cortex (OFC), pars orbitalis of the inferior frontal gyrus

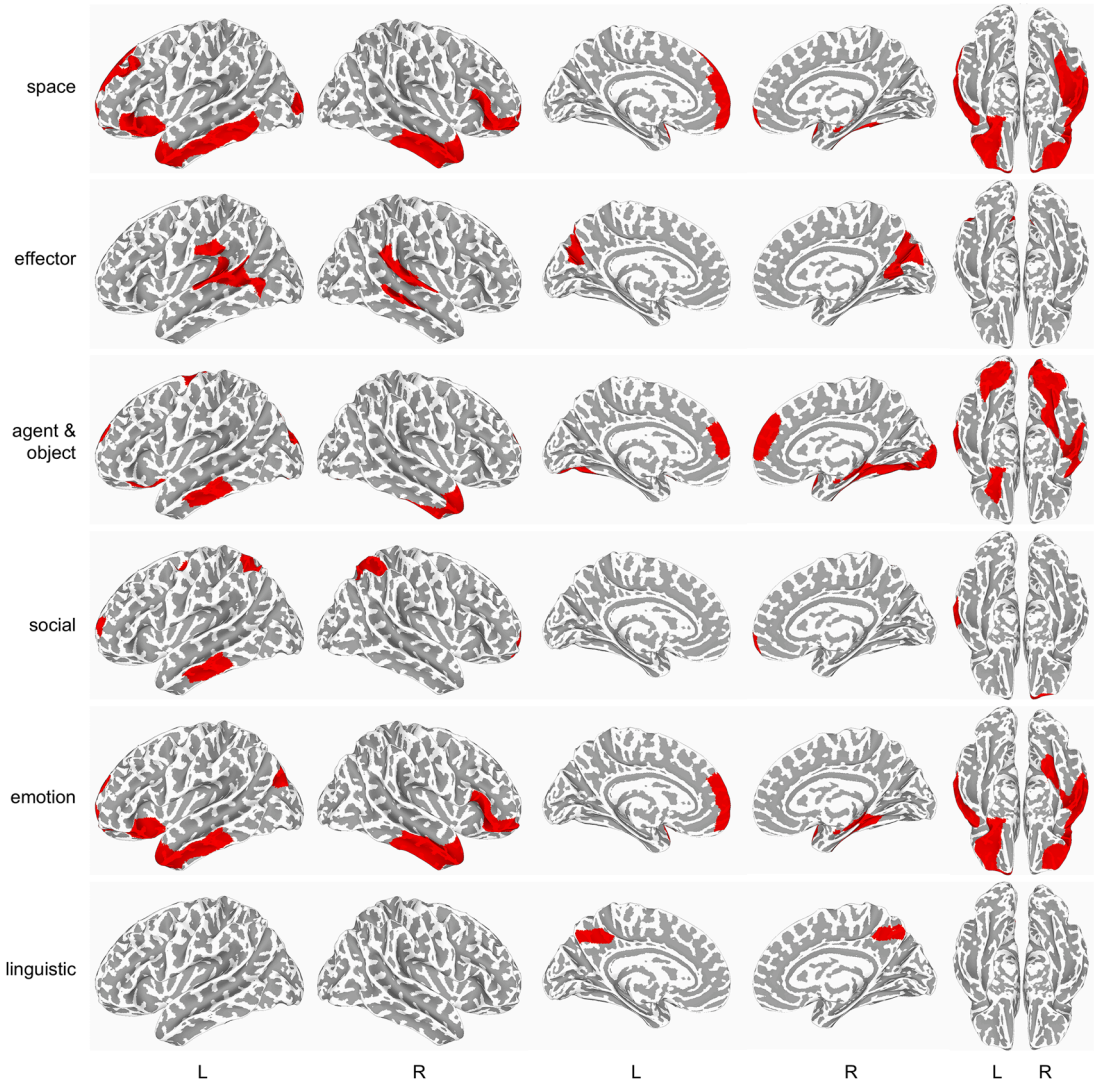


Fig. 4. Spatial domain specificity in cortical action representations. Cortical regions predominantly tuned to each action domain. ROIs were identified using a rank-based approach and aggregating results across modalities. *L*: left, *R*: right.

(IFGorb), left pars triangularis of the inferior frontal gyrus (IFGtri), left anterior medial prefrontal cortex (mPFC), and left dorsomedial prefrontal cortex (dmPFC).

3.4.2. *Effector*

Tarhan & Konkle (2020) underlined the importance of the effector in action representation and found that effector-related features modulate responses in occipital, temporal, and parietal cortices, with dorsal areas preferentially encoding effector visibility rather than identity.

Here, effector-specific ROIs were identified in the left parietal operculum (OP), which is considered part of the extended human Mirror Neuron System (Bonini, 2017), and parieto-occipital sulcus (POS). We also found tuning in areas associated with auditory and language processing (i.e., STS and MTG); this response was likely driven by the mouth effector, engaged during speaking.

3.4.3. *Agent & object*

This domain predominantly identifies actions involving or directed towards non-social targets, such as transitive actions. Observation of object manipulation has been shown to activate parietal regions (Gallivan and Culham, 2015; Urgen and Orban, 2021). Moreover, ventral LOTC represents actions based on transitivity (Wurm et al., 2017; Tucciarelli et al., 2019).

Here, we found that Agent & Object features were preferentially encoded in ventral occipito-temporal areas, part of the ventral pathway for object recognition (Goodale and Milner, 1992). Other identified areas included: bilateral anterior mPFC, right temporal pole (TP), left dorsal premotor area (PMd), left inferior temporal sulcus (ITS), and right primary visual cortex (V1).

3.4.4. *Social*

Studies investigating the social dimension of actions have typically implicated STS (Isik et al., 2017; Masson et al., 2018; Tarhan and Konkle, 2020), the temporoparietal junction (TPJ; Arioli and Canessa, 2019; Arioli et al., 2021; Masson and Isik, 2021), and mPFC (Wagner et al., 2016; Masson and Isik, 2021). Furthermore, some evidence suggests that sociality is encoded in the dorsal parts of LOTC (Wurm et al., 2017; Han et al., 2024), while the superior parietal lobule (SPL) has been shown to respond to observed socio-affective touch (Masson et al., 2018).

Observation of joint actions recruits areas in the temporal poles, STS/MTG, precuneus, and TPJ (Leube et al., 2012; Eskenazi et al., 2015); these activations overlap with areas comprising the Theory of Mind network (Schurz et al., 2014; Arioli and Canessa, 2019; Arioli et al., 2021), underlying mentalizing processes necessary for inferring others' state of mind.

Adding to the existing literature, we found tuning to the Social

domain in left ITS, frontal pole (FP), left frontal eye field (FEF), and SPL and IPS. IPS and FEF are part of the Dorsal Attention Network, which has been shown to encode embodied aspects of social cognition (Lahnakoski et al., 2012). Moreover, a recent meta-analysis (Zhao et al., 2024) suggested that SPL may be part of an action observation pathway dedicated to the processing of non-social actions, therefore capturing modulation of social features.

3.4.5. Emotion

Previous research on naturalistic viewing indicates that affective features, including valence and arousal, partly predict responses in STS, TPJ, anterior temporal lobe (ATL), and mPFC (Masson and Isik; 2021). Moreover, lateral OFC has been shown to map emotional content regardless of the sensory modality (Lettieri et al., 2024). Emotional information conveyed specifically through body language activates ATL (Tipper et al., 2015) and other areas specialized in emotional processing, such as amygdala, OFC, anterior cingulate cortex, and anterior insula (de Gelder, 2006; Sokolov et al., 2020).

Other features included in the domain were Gesticulation and Symbolic gestures, which have been shown to evoke responses in frontal (e.g., IFG) and temporal (e.g., MTG, STG extending to supramarginal gyrus) cortices (Andric et al., 2013; Möttönen et al., 2016; Papeo et al., 2019).

Consistent with these findings, our results revealed areas exhibiting specificity to the *Emotion* domain in OFC, left anterior mPFC, bilateral ATL, and right IFG*orb* and IFG*tr*.

3.4.6. Linguistic

The literature on the neural correlates of action verb processing has highlighted that Telicity modulates activity in the left posterior MTG (Romagno et al., 2012) and bilateral precuneus (Malaia and Newman, 2014), while Iterativity engages IPS (Lai et al., 2023), and Dynamicity MTG/STS (Peelen et al., 2012).

In the present study, only the precuneus showed specific tuning to the Linguistic domain.

4. Discussion

This study assessed a comprehensive taxonomic model of action representation, designed to capture six key domains of action descriptors (i.e., Space, Effector, Agent & Object, Social, Emotion, and Linguistic) in naturalistic contexts. Using fMRI data collected across different sensory modalities (i.e., audiovisual, visual-only, and auditory-only presentations of the same movie), we demonstrated that this action feature-based model effectively predicts a significant portion of brain activity. Furthermore, the results showed that domain representations were consistently maintained across sensory modalities, both in their cortical spatial distribution and in their relative contribution to explaining brain activity.

4.1. Naturalistic approach to model design and stimulation

The taxonomic model of naturalistic action representation was developed through a multi-step process. First, we identified and categorized critical action features into six conceptual domains, drawing on existing literature on action representation. Next, multiple raters were engaged to extract and annotate action-related information from the movie stimulus. This step comprised two main stages: event segmentation, during which raters were asked to detect discrete actions from the continuous movie stream and define their temporal locations; and feature tagging, i.e., raters described each identified action according to the features outlined in the taxonomic model.

For the segmentation step, raters were instructed to freely identify the temporal boundaries of distinct actions. Event segmentation research suggests that actions follow a partonomic organization: they can be defined at multiple levels of granularity, ranging from fine-grain atomic movements to broader, coarse-grain events based on actor intentions and goals (Shipley and Zacks, 2008; Gu et al., 2018). We deliberately avoided imposing strict constraints or providing explicit instructions about the grain size of actions. This approach aligns with findings indicating that partitioning a continuous stream of events into discrete units is a subjective and non-trivial process for human observers. In fact, event segmentation emerges from the interplay of perceptual cues and inferential processes involving prior knowledge, observer goals, and task instructions (Shipley and Zacks, 2008). Notably, instructions can modulate segmentation density by guiding observers toward coarser or finer-grain segmentation criteria (Zacks et al., 2001). By refraining from providing specific instructions, raters were allowed to adhere to their own criteria for defining action, fostering the emergence of a more ecologically valid and naturalistic segmentation unit. At the same time, the absence of a common criteria inevitably led to greater variability across raters, as reflected in differences in the number of identified events. Nevertheless, overall inter-rater agreement on the annotated features was found to be robust, and, by including only annotations that were common to at least two out of three raters, we constructed an average group model that effectively captures a structured representation of action domains. Importantly, the domains were designed to be orthogonal, and analyses of inter-domain correlations confirmed their minimal overlap and independent contributions to the final model.

Notably, most of the features incorporated in our model have previously been explored within the literature, often with neural correlates identified for each of them (Grafton and Hamilton, 2007; Van Elk et al., 2014; Kemmerer, 2021). However, traditional paradigms usually focus on isolated actions, employing highly-controlled single clips with low ecological validity. These approaches often overlook the complex interplay and contextual dependencies between multiple actions. Therefore, to enhance ecological validity, we evaluated our taxonomic model using a naturalistic stimulation paradigm which enables a more comprehensive understanding of how the brain interprets and processes actions within the dynamic context of everyday life, thus ultimately improving the reliability and generalizability of research findings (Hasson et al., 2004).

4.2. Model goodness of fit and invariance to sensory input

Testing whether our model could predict fMRI data during AV stimulation, we found that most of the cortex encodes action features. Statistically significant ROIs spanned prefrontal, temporal, parietal, and occipital cortices, substantially overlapping with the extended AON (Han et al., 2024). As previously noted, prior literature mostly focused on isolated action features and highly controlled experimental stimuli. Here, the naturalistic approach adopted in both action model creation and stimulus design uncovered a much more extended network, encompassing areas involved in mentalizing and social cognition as well, namely, the temporo-parietal junction, medial prefrontal cortex, precuneus, and anterior temporal cortex (Schurz et al., 2014; Arioli et al., 2021). In a naturalistic setting, actions are complex stimuli, often embedded within a social context, relaying the affective state of the agent, and requiring the observer to infer and understand agents' goals and intentions (Stern, 2010; Tarhan et al., 2021). Consequently, observing naturalistic actions entails the recruitment of a wide portion of the cortex (Haxby et al., 2020), encompassing both the action observation and mentalizing networks, which jointly support the representation of action meaning and agents' mental states (Arioli and Canessa, 2019).

On average, the full model explained around 6 % of the variance in neural activation, reaching a maximum of 13 % in posterior temporal and lateral occipital areas. While this effect size is consistent with previous literature (Huth et al., 2012; Cichy et al., 2021; Lahner et al., 2024), a considerable proportion of variance in neural activity remained unexplained. In the present study, we employed a high-level taxonomic

action model which showed higher collinearity with high-level perceptual and semantic rather than low-level properties. Yet, the AON might be preferentially tuned to lower-level perceptual information. In such a case, our results would be driven by the marginal collinearities between our model and perceptual lower-level features. To address this possibility, we measured the impact of lower-level perceptual features by re-testing the full-model in the AV modality after orthogonalization (see Supplementary Figure S7 for detailed methods). The results of this control analysis showed that, after accounting for computational low-level properties, the model still retained a robust predictive pattern, peaking in the right posterior MTG (10 % of variance explained by the taxonomic features), confirming that this model did not rely on perceptual features. The existing literature appears to support this possibility: studies comparing or estimating unique contributions of perceptual vs. higher-level models demonstrated an advantage of the latter in explaining brain activations (Urgen et al., 2019; Masson and Isik, 2021; Dima et al., 2022). These findings are paralleled by behavioral investigations suggesting that action categorization and similarity judgments are driven by higher-level dimensions such as sociality and action goals rather than visual characteristics and kinematics (Tarhan et al., 2021; Kabulska and Lingnau, 2023; Dima et al., 2024).

The action model's ability to explain brain activity independently of stimulation modality was also assessed. Results in the video-only modality were similar to the multimodal stimulation, as the model significantly predicted brain activity in most temporal, occipital, and parietal areas. On the other hand, auditory-only stimulation resulted in a loss of explanatory power across the cortex, with only a few areas in the temporal cortex surviving statistical thresholding for multiple comparisons. It should be noted that, for all three modalities, we tested the same action model, which was developed by tagging events in the audio-video version of the stimulus. While raters were instructed to tag every action independently of the sensory modality of stimulus presentation, the action descriptors inevitably encompass multisensory information. Indeed, the effect size in the multisensory modality was greater than in the visual-only and auditory-only modalities. The increased responsivity of the AON to multimodal audio-video stimulation as compared to unimodal conditions has already been proved (Kaplan and Iacoboni, 2007; McGarry et al., 2012; Bischoff et al., 2014), with evidence pointing toward an advantage of the visual over the auditory modality (Copelli et al., 2022). Equally, while action processing as a whole results to be modality-independent (Ricciardi et al., 2009; 2013), whether the representation of individual action features is due to concurrent additive modality-dependent responses or to a truly shared, modality-independent coding is still uncertain (Alaerts et al., 2009; Rezk et al., 2020).

4.3. Domains representation across modalities

Previous investigations into action representation have usually adopted univariate approaches, testing action properties in isolation (Kemmerer, 2021). Therefore, the potential interactions and the complex manner in which such properties may jointly contribute to brain activity have been overlooked. To address this limitation, we employed a variance partitioning approach, which enabled us to disentangle the unique contribution of each action domain, while accounting for collinearity between features. This approach yielded equivalent results across sensory modalities, both in terms of the magnitude of domain-specific R^2 and their cortical distribution. Specifically, we observed a stability of domain contributions across brain regions and sensory modalities.

In this regard, the Effector domain explained the highest amount of variance in all modalities, highlighting the importance of this feature in action perception (Beurze et al., 2007; Abdollahi et al., 2013; Tarhan and Konkle, 2020). The Social domain was the second most contributing domain, consistent with previous accounts proposing the socio-affective dimension as an organizing principle for action representation (Wurm et al., 2017; Tarhan and Konkle, 2020; Dima et al., 2022; Kabulska and Lingnau, 2023; Zhao et al., 2024; Han et al., 2024). On the other hand, we found that the Emotion domain had the least impact on the full model across all modalities, which may be due to the higher level of complexity of our Emotion features with respect to previous studies (e.g., emotional information relayed by the narrative rather than scene valence; Masson and Isik, 2021; Kabulska and Lingnau, 2023). Altogether, these findings suggest that the representational organization of action features remains consistent despite differences in sensory input.

The absolute magnitude of domain-specific R^2 revealed an imbalance toward the Effector domain, which was consistent across ROIs. Therefore, to highlight potential preferences in feature tuning, we employed a rank-based approach assessing domains' relative specificity within ROIs. We found weak evidence of domain specificity, with multiple regions preferentially recruited by more than one domain. Nonetheless, as discussed above in the Results section, the spatial distribution of domain preference is consistent with previous literature. Indeed, the lack of specificity of the AON is corroborated by lesion studies, which provide sparse evidence of associations between these specific action features and brain regions (Kalénine et al., 2010, 2013; Bonivento et al., 2014; Urgesi et al., 2014).

In the audio-video, visual-only, and auditory-only conditions, single domain contributions added up to 54 %, 57 %, and 49 % of the full model R^2, respectively. Thus, around half of the neural activity explained by the action model can be attributed to variance shared across domains rather than to their unique contributions. Again, the lack of domain specificity in our results is consistent with this notion. Because of computational constraints, it was not feasible to estimate common sources of variance between all possible domain combinations. Interestingly, our action features were grouped into conceptually meaningful domains so that collinearities between domains were minimized. Despite model orthogonality, we found that shared information across domains accounted for the majority of explained brain activity, suggesting that our taxonomic categorization may not reflect how the brain organizes action-related information. Rather than relying on theoretically motivated stimulus descriptors, recent works (Zheng et al., 2019; Hebart et al., 2020) have employed data-driven approaches to stimulus feature extraction in object coding, leveraging human similarity judgments of stimuli. Such approaches may provide a better approximation of the dimensions governing mental representations (Hebart et al., 2020) and may reveal critical features that previous literature might have missed (Kabulska and Lingnau, 2023), thus increasing the ecological validity of the investigated dimensions. Implementations of such approaches in the framework of action representation may identify action features tailored to describe naturalistic stimuli and explain their neural representations.

5. Conclusions

In conclusion, this study has extended the understanding of action representation in the human brain across varying sensory modalities by taking advantage of a naturalistic stimulation and a comprehensive taxonomic model of action features. The findings reveal that the AON is capable of robustly encoding action-related information across sensory modalities, highlighting the modality-general nature of action processing. Furthermore, by means of variance partitioning, we demonstrated that specific action domains contribute distinctly to neural representation, with some domains, such as Effector and Social, showing stronger influence than others. Importantly, the work underscores the value of a naturalistic approach to neuroscience research, providing a more ecologically valid insight into how actions are processed in real-world settings. This study not only enhances our understanding of neural mechanisms underlying action perception but also opens avenues for further research into how actions are integrated across different sensory inputs to form a cohesive understanding of others' behaviors.

Ethics statement

The study was approved by the Ethical Committee of the University of Turin (protocol number 195874/2019) and conforms to the Declaration of Helsinki.

Data availability

Functional MRI data are available on https://osf.io/j8x6h/. Only preprocessed functional data were shared. Code and action model are available here: https://github.com/LauraMarras/Action101.

Funding

This work was supported by the PRIN grants (20223K8B3X and P20228PHN2) by the Italian Ministry of University and Research to E.R. and by the "Tuscany Health Ecosystem—THE" Project, Spoke 8, granted by Next Generation EU—National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, NRRP)—Mission 4 Component 2 Investment 1.4—Ministry of University and Research (MUR) Call N. 3277, Project Code ECS_00000017 to E.R. and G.H.

CRediT authorship contribution statement

Laura Marras: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation. **Lorenzo Teresi:** Methodology, Investigation, Conceptualization. **Francesca Simonelli:** Writing – original draft, Visualization, Software, Methodology, Investigation. **Francesca Setti:** Writing – original draft, Investigation. **Alessandro Ingenito:** Investigation. **Giacomo Handjaras:** Writing – original draft, Visualization, Supervision, Methodology, Conceptualization. **Emiliano Ricciardi:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare no conflicts of interest.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2025.121439.

References

Abdollahi, R.O., Jastorff, J., Orban, G.A., 2013. Common and segregated processing of observed actions in human SPL. Cereb. Cortex 23 (11), 2734–2753.
Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., McGrew, B., 2023. Gpt-4 technical report. arXiv preprint. *arXiv:2303.08774*.
Alaerts, K., Swinnen, S.P., Wenderoth, N., 2009. Interaction of sound and sight during action perception: evidence for shared modality-dependent action representations. Neuropsychologia 47 (12), 2593–2599.
Andric, M., Solodkin, A., Buccino, G., Goldin-Meadow, S., Rizzolatti, G., Small, S.L., 2013. Brain function overlaps when people observe emblems, speech, and grasping. Neuropsychologia 51 (8), 1619–1629.
Arioli, M., Canessa, N., 2019. Neural processing of social interaction: coordinate-based meta-analytic evidence from human neuroimaging studies. Hum. Brain Mapp. 40 (13), 3712–3737.
Arioli, M., Cattaneo, Z., Ricciardi, E., Canessa, N., 2021. Overlapping and specific neural correlates for empathizing, affective mentalizing, and cognitive mentalizing: a coordinate-based meta-analytic study. Hum. Brain Mapp. 42 (14), 4777–4804.
Barliya, A., Omlor, L., Giese, M.A., Berthoz, A., Flash, T., 2013. Expression of emotion in the kinematics of locomotion. Exp. Brain Res. 225, 159–176.
Beauchamp, M.S., 2005. See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex. Curr. Opin. Neurobiol. 15 (2), 145–153.
Beurze, S.M., De Lange, F.P., Toni, I., Medendorp, W.P., 2007. Integration of target and effector information in the human brain during reach planning. J. Neurophysiol. 97 (1), 188–199.
Bilenko, N.Y., Gallant, J.L., 2016. Pyrcca: regularized kernel canonical correlation analysis in python and its applications to neuroimaging. Front. Neuroinform. 10, 49.
Bischoff, M., Zentgraf, K., Pilgramm, S., Stark, R., Krüger, B., Munzert, J., 2014. Anticipating action effects recruits audiovisual movement representations in the ventral premotor cortex. Brain Cogn. 92, 39–47.
Bonini, L., 2017. The extended mirror neuron network: anatomy, origin, and functions. Neurosci. 23 (1), 56–67.
Bonivento, C., Rothstein, P., Humphreys, G., Chechlacz, M., 2014. Neural correlates of transitive and intransitive action imitation: an investigation using voxel-based morphometry. NeuroImage 6, 488–497.
Chambon, V., Sidarus, N., Haggard, P., 2014. From action intentions to action effects: how does the sense of agency come about? Front. Hum. Neurosci. 8, 320.
Cichy, R.M., Dwivedi, K., Lahner, B., Lascelles, A., Iamshchinina, P., Graumann, M., Oliva, A., 2021. The algonauts project 2021 challenge: how the human brain makes sense of a world in motion. arXiv preprint. *arXiv:2104.13714*.
Copelli, F., Rovetti, J., Ammirante, P., Russo, F.A., 2022. Human mirror neuron system responsivity to unimodal and multimodal presentations of action. Exp. Brain Res. 240 (2), 537–548.
Cox, R.W., 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29 (3), 162–173.
De Gelder, B., 2006. Towards the neurobiology of emotional body language. Nat. Rev. Neurosci. 7 (3), 242–249.
De Gelder, B., Van den Stock, J., 2011. The bodily expressive action stimulus test (BEAST). Construction and validation of a stimulus basis for measuring perception of whole body expression of emotions. Front. Psychol. 2, 181.
de Heer, W.A., Huth, A.G., Griffiths, T.L., Gallant, J.L., Theunissen, F.E., 2017. The hierarchical cortical organization of human speech processing. J. Neurosci. 37 (27), 6539–6557.
Decety, J., Grèzes, J., 1999. Neural mechanisms subserving the perception of human actions. Trends Cogn. Sci. (Regul. Ed.) 3 (5), 172–178.
Dima, D.C., Tomita, T.M., Honey, C.J., Isik, L., 2022. Social-affective features drive human representations of observed actions. Elife 11, e75027.
Dima, D.C., Hebart, M.N., Isik, L., 2023. A data-driven investigation of human action representations. Sci. Rep. 13 (1), 5171.
Dima, D.C., Janarthanan, S., Culham, J.C., Mohsenzadeh, Y., 2024. Shared representations of human actions across vision and language. Neuropsychologia 202, 108962.
Eskenazi, T., Rueschemeyer, S.A., de Lange, F.P., Knoblich, G., Sebanz, N., 2015. Neural correlates of observing joint actions with shared intentions. Cortex 70, 90–100.
Fisher, R.A., 1970. Statistical methods for research workers. Breakthroughs in statistics: Methodology and Distribution. Springer New York, New York, NY, pp. 66–70.
Fonov, V.S., Evans, A.C., McKinstry, R.C., Almli, C.R., Collins, D.L., 2009. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. Neuroimage 47, S102.
Gallivan, J.P., McLean, D.A., Valyear, K.F., Culham, J.C., 2013. Decoding the neural mechanisms of human tool use. Elife 2, e00425.
Gallese, V., Fadiga, L., Fogassi, L., Rizzolatti, G., 1996. Action recognition in the premotor cortex. Brain 119 (2), 593–609.
Gallivan, J.P., Culham, J.C., 2015. Neural coding within human brain areas involved in actions. Curr. Opin. Neurobiol. 33, 141–149.
Giese, M.A., Rizzolatti, G., 2015. Neural and computational mechanisms of action processing: interaction between visual and motor representations. Neuron 88 (1), 167–180.
Goldberg, H., Preminger, S., Malach, R., 2014. The emotion–action link? Naturalistic emotional stimuli preferentially activate the human dorsal visual stream. Neuroimage 84, 254–264.
Goodale, M.A., Milner, A.D., 1992. Separate visual pathways for perception and action. Trends Neurosci. 15 (1), 20–25.
Grafton, S.T., Hamilton, A.F.D.C., 2007. Evidence for a distributed hierarchy of action representation in the brain. Hum. Mov. Sci. 26 (4), 590–616.
Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Malik, J., 2018. Ava: a video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6047–6056.
Han, J., Chauhan, V., Philip, R., Taylor, M.K., Jung, H., Halchenko, Y.O., Nastase, S.A., 2024. Behaviorally-relevant features of observed actions dominate cortical representational geometry in natural vision. bioRxiv., 2024–11
Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R., 2004. Intersubject synchronization of cortical activity during natural vision. Science (1979) 303 (5664), 1634–1640.
Haxby, J.V., Gobbini, M.I., Nastase, S.A., 2020. Naturalistic stimuli reveal a dominant role for agentic action in visual representation. Neuroimage 216, 116561.
Haynes, J.D., 2015. A primer on pattern-based approaches to fMRI: principles, pitfalls, and perspectives. Neuron 87 (2), 257–270.
Hebart, M.N., Zheng, C.Y., Pereira, F., Baker, C.I., 2020. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. Nat. Hum. Behav. 4 (11), 1173–1185.
Henderson, J.M., Larson, C.L., Zhu, D.C., 2007. Cortical activation to indoor versus outdoor scenes: an fMRI study. Exp. Brain Res. 179, 75–84.
Henderson, J.M., Zhu, D.C., Larson, C.L., 2011. Functions of parahippocampal place area and retrosplenial cortex in real-world scene analysis: an fMRI study. Vis. Cogn. 19 (7), 910–927.
Herek, S., 1996. 101 Dalmatians. Great Oaks Entertainment & Walt Disney (Director).
Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Wilson, K., 2017. CNN architectures for large-scale audio classification. In: *2017 ieee international conference on acoustics, speech and signal processing* (icassp). IEEE, pp. 131–135.
Hotelling, H., 1992. Relations between two sets of variates. Breakthroughs in statistics: Methodology and Distribution. Springer New York, New York, NY, pp. 162–190.

Huth, A.G., Nishimoto, S., Vu, A.T., Gallant, J.L., 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76 (6), 1210–1224.
Isik, L., Koldewyn, K., Beeler, D., Kanwisher, N., 2017. Perceiving social interactions in the posterior superior temporal sulcus. In: Proceedings of the National Academy of Sciences, 114, pp. E9145–E9152.
James, T.W., VanDerKlok, R.M., Stevenson, R.A., James, K.H., 2011. Multisensory perception of action in posterior temporal and parietal cortices. Neuropsychologia 49 (1), 108–114.
Kabulska, Z., Lingnau, A., 2023. The cognitive structure underlying the organization of observed actions. Behav. Res. Methods 55 (4), 1890–1906.
Kalénine, S., Buxbaum, L.J., Coslett, H.B., 2010. Critical brain regions for action recognition: lesion symptom mapping in left hemisphere stroke. Brain 133 (11), 3269–3280.
Kalénine, S., Shapiro, A.D., Buxbaum, L.J., 2013. Dissociations of action means and outcome processing in left-hemisphere stroke. Neuropsychologia 51 (7), 1224–1233.
Kaplan, J.T., Iacoboni, M., 2007. Multimodal action representation in human left ventral premotor cortex. Cogn. Process. 8 (2), 103–113.
Karakose-Akbiyik, S., Caramazza, A., Wurm, M.F., 2023. A shared neural code for the physics of actions and object events. Nat. Commun. 14 (1), 3316.
Kemmerer, D., 2021. What modulates the mirror neuron system during action observation?: multiple factors involving the action, the actor, the observer, the relationship between actor and observer, and the context. Prog. Neurobiol. 205, 102128.
Kilner, J.M., 2011. More than one pathway to action understanding. Trends Cogn. Sci. (Regul. Ed.) 15 (8), 352–357.
Kirsch, L.P., Cross, E.S., 2015. Additive routes to action learning: layering experience shapes engagement of the action observation network. Cereb. Cortex 25 (12), 4799–4811.
Kornblith, S., Norouzi, M., Lee, H., Hinton, G., 2019. Similarity of neural network representations revisited. In: International conference on machine learning. PMLR, pp. 3519–3529.
Lahnakoski, J.M., Glerean, E., Salmi, J., Jääskeläinen, I.P., Sams, M., Hari, R., Nummenmaa, L., 2012. Naturalistic fMRI mapping reveals superior temporal sulcus as the hub for the distributed brain network for social perception. Front. Hum. Neurosci. 6, 233.
Lahner, B., Dwivedi, K., Iamshchinina, P., Graumann, M., Lascelles, A., Roig, G., Cichy, R., 2024. Modeling short visual events through the BOLD moments video fMRI dataset and metadata. Nat. Commun. 15 (1), 6241.
Lai, Y.Y., Sakai, H., Makuuchi, M., 2023. Neural underpinnings of processing combinatorial unstated meaning and the influence of individual cognitive style. Cereb. Cortex 33 (18), 10013–10027.
Lettieri, G., Handjaras, G., Cappello, E.M., Setti, F., Bottari, D., Bruno, V., Cecchetti, L., 2024. Dissecting abstract, modality-specific and experience-dependent coding of affect in the human brain. Sci. Adv. 10 (10), eadk6840.
Leube, D., Straube, B., Green, A., Blümel, I., Prinz, S., Schlotterbeck, P., Kircher, T., 2012. A possible brain network for representation of cooperative behavior and its implications for the psychopathology of schizophrenia. Neuropsychobiology. 66 (1), 24–32.
Malaia, E., Newman, S., 2014. Neural bases of event knowledge and syntax integration in comprehension of complex sentences. Neurocase 21 (6), 753–766.
Masson, H.L., Van De Plas, S., Daniels, N., de Beeck, H.O., 2018. The multidimensional representational space of observed socio-affective touch experiences. Neuroimage 175, 297–314.
Masson, H.L., Isik, L., 2021. Functional selectivity for social interaction perception in the human superior temporal sulcus during natural viewing. Neuroimage 245, 118741.
McGarry, L.M., Russo, F.A., Schalles, M.D., Pineda, J.A., 2012. Audio-visual facilitation of the mu rhythm. Exp. Brain Res. 218, 527–538.
Möttönen, R., Farmer, H., Watkins, K.E., 2016. Neural basis of understanding communicative actions: changes associated with knowing the actor's intention and the meanings of the actions. Neuropsychologia 81, 230–237.
Murphy, A., Zylberberg, J., Fyshe, A., 2024. Correcting Biased Centered Kernel Alignment Measures in Biological and Artificial Neural Networks. arXiv preprint. *arXiv:2405.01012*.
Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2011. Encoding and decoding in fMRI. Neuroimage 56 (2), 400–410.
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 (19), 1641–1646.
Orban, G.A., Lanzilotto, M., Bonini, L., 2021. From observed action identity to social affordances. Trends Cogn. Sci. (Regul. Ed.) 25 (6), 493–505.
Papeo, L., Agostini, B., Lingnau, A., 2019. The large-scale organization of gestures and words in the middle temporal gyrus. J. Neurosci. 39 (30), 5966–5974.
Peelen, M.V., Romagno, D., Caramazza, A., 2012. Independent representations of verbs and actions in left lateral temporal cortex. J. Cogn. Neurosci. 24 (10), 2096–2107.
Pickands III, J., 1975. Statistical inference using extreme order statistics. Ann. Stat. 119–131.
Rezk, M., Cattoir, S., Battal, C., Occelli, V., Mattioni, S., Collignon, O., 2020. Shared representation of visual and auditory motion directions in the human middle-temporal cortex. Curr. Biol. 30 (12), 2289–2299.
Ricciardi, E., Bonino, D., Sani, L., Vecchi, T., Guazzelli, M., Haxby, J.V., Pietrini, P., 2009. Do we really need vision? How blind people "see" the actions of others. J. *Neurosci.* 29 (31), 9719–9724.
Ricciardi, E., Handjaras, G., Bonino, D., Vecchi, T., Fadiga, L., Pietrini, P., 2013. Beyond motor scheme: a supramodal distributed representation in the action-observation network. PLoS. One 8 (3), e58632.
Rizzolatti, G., Craighero, L., 2004. The mirror-neuron system. Annu. Rev. Neurosci. 27, 169–192.
Romagno, D., Rota, G., Ricciardi, E., Pietrini, P., 2012. Where the brain appreciates the final state of an event: the neural correlates of telicity. Brain Lang. 123 (1), 68–74.
Schaefer, A., Kong, R., Gordon, E.M., Laumann, T.O., Zuo, X.N., Holmes, A.J., Yeo, B.T., 2018. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. Cereb. Cortex 28 (9), 3095–3114.
Schurz, M., Radua, J., Aichhorn, M., Richlan, F., Perner, J., 2014. Fractionating theory of mind: a meta-analysis of functional brain imaging studies. Neurosci. Biobehav. Rev. 42, 9–34.
Sebanz, N., Bekkering, H., Knoblich, G., 2006. Joint action: bodies and minds moving together. Trends Cogn. Sci. (Regul. Ed.) 10 (2), 70–76.
Setti, F., Handjaras, G., Bottari, D., Leo, A., Diano, M., Bruno, V., Ricciardi, E., 2023. A modality-independent proto-organization of human multisensory areas. Nat. Hum. Behav. 7 (3), 397–410.
Simonelli, F., Handjaras, G., Benuzzi, F., Bernardi, G., Leo, A., Duzzi, D., Lui, F., 2024. Sensitivity and specificity of the action observation network to kinematics, target object, and gesture meaning. Hum. Brain Mapp. 45 (11), e26762.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint. *arXiv:1409.1556*.
Sinigaglia, C., Butterfill, S.A., 2020. Motor representation and action experience in joint action. Minimal Cooperat. Shared Agency 181–193.
Schippers, M.B., Roebroeck, A., Renken, R., Nanetti, L., Keysers, C., 2010. Mapping the information flow from one brain to another during gestural communication. In: Proceedings of the National Academy of Sciences, 107, pp. 9388–9393.
Shipley, T.F., Zacks, J.M., 2008. Understanding events: From perception to Action. Oxford University Press.
Sokolov, A.A., Zeidman, P., Erb, M., Pollick, F.E., Fallgatter, A.J., Ryvlin, P., Pavlova, M. A., 2020. Brain circuits signaling the absence of emotion in body language. In: Proceedings of the National Academy of Sciences, 117, pp. 20868–20873.
Stern, D.N., 2010. Forms of vitality: Exploring dynamic Experience in psychology, the arts, psychotherapy, and Development. Oxford University Press, UK.
Stobbe, E., Forlim, C.G., Kühn, S., 2024. Impact of exposure to natural versus urban soundscapes on brain functional connectivity, BOLD entropy and behavior. Environ. Res. 244, 117788.
Tarhan, L., Konkle, T., 2020. Sociality and interaction envelope organize visual action representations. Nat. Commun. 11 (1), 3002.
Tarhan, L., De Freitas, J., Konkle, T., 2021. Behavioral and neural representations en route to intuitive action understanding. Neuropsychologia 163, 108048.
Tipper, C.M., Signorini, G., Grafton, S.T., 2015. Body language in the brain: constructing meaning from expressive movement. Front. Hum. Neurosci. 9, 450.
Tucciarelli, R., Wurm, M., Baccolo, E., Lingnau, A., 2019. The representational space of observed actions. Elife 8, e47686.
Urgen, B.A., Pehlivan, S., Saygin, A.P., 2019. Distinct representations in occipito-temporal, parietal, and premotor cortex during action perception revealed by fMRI and computational modeling. Neuropsychologia 127, 35–47.
Urgen, B.A., Orban, G.A., 2021. The unique role of parietal cortex in action observation: functional organization for communicative and manipulative actions. Neuroimage 237, 118220.
Urgesi, C., Candidi, M., Avenanti, A., 2014. Neuroanatomical substrates of action perception and understanding: an anatomic likelihood estimation meta-analysis of lesion-symptom mapping studies in brain injured patients. Front. Hum. Neurosci. 8, 344.
van Elk, M., van Schie, H., Bekkering, H., 2014. Action semantics: a unifying conceptual framework for the selective use of multimodal and modality-specific object knowledge. Phys. Life Rev. 11 (2), 220–250.
Vendler, Z., 1967. Linguistics in Philosophy. Cornell Univ. Pr., Ithaca.
Wagner, D.D., Kelley, W.M., Haxby, J.V., Heatherton, T.F., 2016. The dorsal medial prefrontal cortex responds preferentially to social interactions during natural viewing. J. Neurosci. 36 (26), 6917–6925.
Westfall, P.H., Young, S.S., 1993. Resampling-based Multiple testing: Examples and Methods For P-Value Adjustment, 279. John Wiley & Sons.
Winkler, A.M., Ridgway, G.R., Douaud, G., Nichols, T.E., Smith, S.M., 2016. Faster permutation inference in brain imaging. Neuroimage 141, 502–516.
Wurm, M.F., Caramazza, A., Lingnau, A., 2017. Action categories in lateral occipitotemporal cortex are organized along sociality and transitivity. Journal of Neuroscience 37 (3), 562–575.
Wurm, M.F., Caramazza, A., 2022. Two 'what' pathways for action and object recognition. Trends Cogn. Sci. (Regul. Ed.) 26 (2), 103–116.
Xu, H., Huang, C.R., 2013. Primitives of events and the semantic representation. In: Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013), pp. 54–61.
Zacks, J.M., Tversky, B., Iyer, G., 2001. Perceiving, remembering, and communicating structure in events. J. Exp. Psychol. 130 (1), 29.
Zarcone, A., Lenci, A., 2010. Priming effects on event types classification: Effects of word and picture stimuli. In: Proceedings of the Annual Meeting of the Cognitive Science Society, 32.
Zhao, M., Li, R., Xiang, S., Liu, N., 2024. Two different mirror neuron pathways for social and non-social actions? A meta-analysis of fMRI studies. Soc. Cogn. Affect. Neurosci. 19 (1), nsae068.
Zheng, C.Y., Pereira, F., Baker, C.I., Hebart, M.N., 2019. Revealing interpretable object representations from human behavior. arXiv preprint. *arXiv:1901.02915*.