

The semantic geometry of action: convergence across foundation models, human behaviour and the brain

Huiguang He

huiguang.he@ia.ac.cn

Institute of Automation, Chinese Academy of Sciences <https://orcid.org/0000-0002-0684-1711>

Changde Du

Institute of Automation, Chinese Academy of Sciences <https://orcid.org/0000-0002-0084-433X>

Yi Sun

Institute of Automation, Chinese Academy of Sciences <https://orcid.org/0009-0005-3848-122X>

Yizhuo Lu

Institute of Automation, Chinese Academy of Sciences <https://orcid.org/0009-0002-1196-6372>

Shuang Qiu

Institute of Automation Chinese Academy of Science

Kun Zhao

Institute of Automation, Chinese Academy of Sciences

Yaozhi Wen

Institute of Automation, Chinese Academy of Sciences

Chen Hu

Zhongguancun Academy, Beijing, China

Article

Keywords:

Posted Date: September 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10930847/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

The semantic geometry of action: convergence across foundation models, human behaviour and the brain

Changde Du^{1,3,4,†,*}, Yi Sun^{1,3,4,†}, Yizhuo Lu^{1,2,†}, Shuang Qiu^{1,3}, Kun Zhao^{1,3}, Yaozhi Wen^{1,2,4}, Chen Hu⁴, and Huiguang He^{1,2,3,4,*}

¹NeuBCI Lab, State Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing, China

²School of Future Technology, University of Chinese Academy of Sciences, Beijing, China

³School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

⁴Zhongguancun Academy, Beijing, China

*corresponding author: Changde Du (changde.du@ia.ac.cn), Huiguang He (huiguang.he@ia.ac.cn)

†these authors contributed equally to this work

ABSTRACT

Interpreting dynamic human actions is central to biological intelligence and essential for embodied artificial intelligence. Although foundation models achieve robust accuracy in action recognition, it remains unclear whether they organize naturalistic actions within a representational geometry aligned with human perception. Here, we evaluated the semantic organization of action representations across foundation models, human behaviour, and the human cortex. By analyzing over two million triplet similarity judgements, we extracted stable, low-dimensional semantic spaces from both text-only large language models and multimodal models. These model-derived spaces predicted human behavioral judgements of action similarity more accurately than baseline spatiotemporal video encoders. To evaluate the functional relevance of these extracted spaces, we demonstrated that modulating individual dimensions within a generative pipeline produced predictable, corresponding semantic shifts in synthesized video content. Furthermore, when transferred without retraining to an independent, large-scale video-fMRI dataset, these model-derived action dimensions successfully predicted stimulus-evoked cortical responses across higher-order visual, temporal, and sensorimotor networks. Together, these findings demonstrate a convergent semantic organization for naturalistic actions across artificial foundation models, human cognitive judgements, and cortical representations.

Introduction

Interpreting the dynamic actions of others is a core component of human cognition and social interaction^{1–3}. This process requires the extraction of abstract, behaviorally relevant information—such as goals, intentions, and environmental contexts—rather than relying solely on the processing of surface-level spatiotemporal kinematics^{4–6}. While conventional video models achieve high accuracy in explicit action recognition tasks⁷, task performance does not necessarily indicate that their internal representations mirror the semantic organization underlying human action judgements^{8,9}.

Foundation models provide an alternative computational framework to investigate representational alignment^{10,11}. Large language models (LLMs) encode broad relational knowledge from text, and multimodal large language models (MLLMs) integrate this knowledge with visual input. Evidence that these models capture aspects of human representational structure for static objects and concepts supports their use as computational models of cognition^{10–14}. However, it remains undetermined whether this alignment extends to the dynamic and context-dependent domain of naturalistic actions. Testing this alignment requires comparing models and humans at the level of representational geometry, and ultimately linking these behavioral architectures to neural responses¹⁰.

Recent methodological advances enable this comparison by recovering interpretable, low-dimensional semantic spaces from similarity judgments, allowing a direct mapping between model preferences and human cognitive or cortical data¹⁵. While previous work has mapped the semantic dimensions underlying human action perception^{16,17}, it remains unresolved which dimensions organize foundation-model representations of naturalistic actions. Furthermore, it is unclear whether observed human–model similarities reflect a shared semantic structure or arise from low-level stimulus statistics^{8,18,19}. We therefore investigated the dimensional organization of foundation-model judgments for naturalistic actions, testing their alignment with human behavioral and cortical data, and objectively assessing their functional utility.

To address these questions, we implemented a machine-behavioral and neurocomputational framework to evaluate action geometry across distinct stimulus sets. We collected zero-shot similarity judgments from text-only and multimodal foundation models across 355 naturalistic action categories, and compared the resulting representational spaces with an independent human

behavioral benchmark¹⁶. We subsequently evaluated the functional relevance of the extracted dimensions through feature attribution, spatial masking, and dimension-guided generative modulation. Finally, we transferred the fixed, model-derived action spaces to an independent neuroimaging dataset (HAD-fMRI²⁰, encompassing 21,600 videos) to predict cortical responses using voxel-wise encoding and searchlight representational similarity analysis (RSA).

Our analyses revealed that both text-only and multimodal foundation models organize actions within a low-dimensional semantic space that aligns with human behavioral judgments, significantly outperforming conventional spatiotemporal video encoders. Modulating individual dimensions of these extracted spaces produced corresponding semantic changes in synthesized video content, confirming their operational relevance. Furthermore, the model-derived dimensions predicted stimulus-evoked cortical responses across higher-order visual, temporal, and sensorimotor networks. Notably, we observed modality-dependent differences where text-derived semantic geometries matched or, in specific higher-order cortical regions, exceeded the predictive performance of multimodal spaces. Together, these findings demonstrate a convergent organizational structure for naturalistic actions across artificial foundation models, human behavior, and the human cortex.

Results

We developed a data-driven framework to evaluate and compare the representational geometry of naturalistic human actions across foundation models and humans. Using the triplet odd-one-out task—a standard paradigm for estimating cognitive representational geometry^{15,21,22}—we collected similarity judgments across 355 naturalistic action categories (Fig. 1a; Supplementary Table 1). To systematically assess the influence of input modality and model scale, we evaluated text-only LLMs (DeepSeek-R1-671B²³ and Qwen2.5-7B²⁴) and MLLMs (Qwen2.5-VL-7B and Qwen2.5-VL-72B²⁵). Multimodal models evaluated raw video clips, whereas text-only models evaluated corresponding textual descriptions refined from speech-recognition transcripts²⁶ (Fig. 1b, c; Supplementary Fig. 1). This procedure yielded over 1.14 million and 1.17 million triplet responses from the multimodal and text-only models, respectively, which were evaluated against an independent human behavioral benchmark comprising 779,467 triplet judgments¹⁶. We applied the sparse positive similarity embedding (SPoSE) algorithm¹⁵ to these behavioral choices (see Methods) to extract low-dimensional representational spaces for each system (Fig. 1d). Finally, we tested whether these model-derived action spaces capture biologically plausible structures by transferring them to an independent neuroimaging dataset (HAD-fMRI²⁰) for voxel-wise encoding and searchlight representational similarity analysis (RSA) (Fig. 1e).

Low-dimensional spaces capture human and model action judgements

To address the inherent stochasticity of the representation learning process, the SPoSE algorithm was executed across 65 distinct random initializations for each system. The derived dimensions exhibited consistent loading profiles across runs (Extended Data Fig. 1), indicating that the extracted spaces reflect stable geometric structures in the judgment distributions rather than algorithmic noise. We evaluated predictive performance on a held-out set of actions to determine the optimal dimensionality of the latent spaces. As shown in Fig. 2a, predictive performance (measured via representational similarity scores) began to plateau at approximately 30 dimensions across the human benchmark, the two LLMs, and the two MLLMs. Following these convergence dynamics, we retained the top 30 principal dimensions (Supplementary Table 2) to construct the final representational embeddings for all five systems.

To quantify the fidelity of these learned representations, we evaluated their accuracy on an independent held-out test set of the odd-one-out task (Fig. 2b). All five embeddings predicted held-out triplet choices well above the 33.3% chance level, confirming that the 30-dimensional spaces captured systematic semantic structure. The human embedding achieved an accuracy of 53.16% ($\pm 0.25\%$). Among the artificial systems, prediction accuracy ranged from 57.70% ($\pm 0.37\%$) for Qwen2.5-7B to 72.28% ($\pm 0.32\%$) for Qwen2.5-VL-72B, while DeepSeek-R1-671B and Qwen2.5-VL-7B achieved accuracies of 67.23% ($\pm 0.35\%$) and 67.62% ($\pm 0.32\%$), respectively. Estimated noise ceilings for fitting individual-trial behavior (see Methods) indicated that the low-dimensional embeddings achieve up to 92.74%, 94.5%, 93.6%, 92.2%, and 95.8% of the optimal predictive accuracy for Human, Qwen2.5-7B, DeepSeek-R1-671B, Qwen2.5-VL-7B, and Qwen2.5-VL-72B, respectively. Because each embedding was evaluated against held-out choices from its own source system, these differences quantify within-system representational consistency rather than alignment with human perception.

We subsequently evaluated whether the SPoSE embeddings could reconstruct empirical pairwise similarity structures. Using a balanced subset of 41 representative actions, we compared model-predicted Representational Similarity Matrices (RSMs) with empirical behavioural RSMs (Fig. 2c; see Methods). The correspondence between predicted and measured similarities was substantiated by strong Pearson correlations across systems. The human embedding achieved $r = 0.85$ (95% CI = [0.83, 0.87], $P < 0.001$). All four artificial models achieved higher internal reconstruction correlations ($r > 0.90$, $P < 0.001$), consistent with the lower intrinsic variance and deterministic response characteristics of artificial computational architectures compared to pooled human participants.

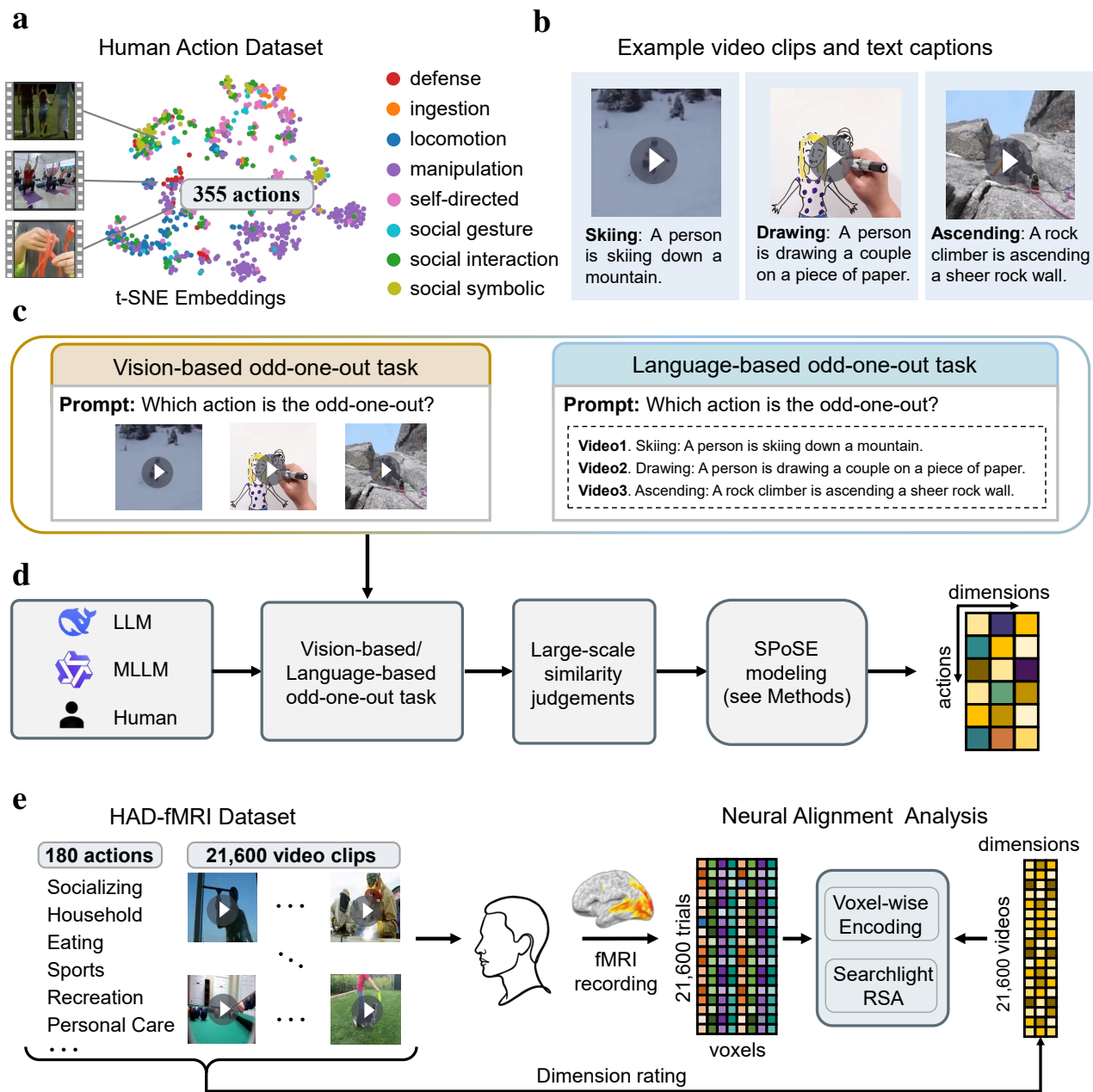


Figure 1. Schematic of the experimental and analytical workflow. **a**, Overview of the human action dataset comprising 355 naturalistic action categories. The t-SNE visualization illustrates the semantic distribution of these actions across eight predefined macro-categories. **b**, Representative stimuli, illustrating the parallel utilization of raw video clips and matched textual captions. **c**, Prompt designs for the vision-based and language-based triplet odd-one-out paradigms. **d**, Computational pipeline for deriving the latent action spaces. Large-scale similarity judgements collected from human participants, LLMs, and MLLMs are processed via the SPoSE algorithm to extract low-dimensional action embeddings. **e**, Neural alignment analysis framework. The model-derived SPoSE dimensions are transferred to an independent, large-scale neuroimaging dataset (HAD-fMRI, encompassing 21,600 video clips). Voxel-wise encoding and searchlight RSA evaluate the geometric correspondence between the dimension-rating matrices and human cortical activity.

Finally, we analyzed representational alignment across different systems by computing a cross-model RSM correlation matrix (Fig. 2d). Three critical patterns emerged. First, representational correspondence was generally higher between models sharing an input modality than across distinct modalities. Second, we observed a systematic asymmetry in human-machine

alignment: human representations predicted model representations more accurately than vice versa. This indicates that the human perceptual space incorporates a broader distribution of semantic features, whereas individual artificial models capture constrained subspaces of human-relevant variance. Third, among the evaluated models, the text-only DeepSeek-R1-671B demonstrated the highest bidirectional correspondence with the human benchmark (predicting human: $r = 0.493$; predicted by human: $r = 0.515$). This outcome provides quantitative evidence that semantic and relational dynamics learned exclusively through linguistic scaffolding can closely approximate the conceptual geometry of human action perception, even in the absence of visually grounded inputs.

Semantic topology and modality-dependent differences in action representations

We examined the macro-scale semantic organization of the extracted action spaces by projecting the 30-dimensional embeddings into a two-dimensional manifold using t-SNE (see Methods). Actions were color-coded according to eight predefined macro-categories¹⁷ (Fig. 3a). The artificial model manifolds map 355 distinct single-exemplar categories, whereas the human manifold visualizes 768 exemplars nested within 256 categories, reflecting their respective experimental designs. This structure tests whether semantic organization persists across broadly sampled categories in the models and repeated within-category exemplars in humans. Despite the lack of explicit categorical supervision in the odd-one-out task, both LLM and MLLM embeddings displayed robust category-related organization. Actions formed visually separable groupings corresponding to specific functional activity types (e.g., *sports*, *socializing*, *housework*). Hierarchical radial dendrograms provided complementary evidence that actions with shared functional or semantic attributes naturally group together across the evaluated systems (Extended Data Fig. 2).

To quantify this categorical structure, we evaluated nearest-centroid classification performance for the eight macro-categories (Fig. 3b; see Methods). The rank order of classification accuracies diverged from that of the held-out triplet prediction (cf. Fig. 2b). DeepSeek-R1-671B achieved the highest top-1 classification accuracy at 72.96% (chance: 19.33%; 95% CI = [15.49%, 23.38%]), followed by Qwen2.5-7B (64.51%). The multimodal models yielded lower classification accuracies (Qwen2.5-VL-72B: 61.69%; Qwen2.5-VL-7B: 56.34%). Under this evaluation, text-only models produced action spaces with greater discrete category separability. This aligns with the cognitive principle that text-based representations abstract away from episodic details to emphasize category-level conceptual semantics, whereas visually grounded representations preserve graded perceptual and kinematic variations within those categories. The human benchmark achieved an intermediate accuracy (59.90%), consistent with a hybrid representational system that incorporates both categorical semantics and perceptual variance.

We next analyzed the intersections of the derived SPoSE dimensions to assess representational overlap (Fig. 3c, d; see Methods). In both three-system comparisons, 18 to 19 dimensions were shared across all three spaces, representing approximately 41% to 46% of the union of dimensions in each comparison. This recurrent intersection indicates a shared core of semantic features underlying action perception (Supplementary Fig. 3). Beyond this shared core, overlap configurations varied by modality and scale. Within the Qwen2.5-VL multimodal family (Fig. 3d), the 72B model shared six dimensions exclusively with the human benchmark, compared with three for the 7B model, suggesting that scale enhances human-aligned abstraction. In the text-only comparison (Fig. 3c), Qwen2.5-7B and DeepSeek-R1 each shared three dimensions exclusively with the human benchmark and retained distinct model-specific dimensions. These intersection patterns demonstrate that human-model correspondence comprises a core set of shared dimensions alongside measurable modality- and architecture-specific components.

Network topology and representational alignment with human behaviour

To examine the topological relationships among individual SPoSE dimensions, we represented each action space as an undirected graph. Nodes correspond to SPoSE dimensions, and edges connect pairs with Pearson correlations exceeding $|r| > 0.19$. Louvain community detection (see Methods) revealed both conserved macro-structures and system-specific semantic organizations (Fig. 4a). Across both biological and artificial systems, communities associated with *Gross Motor Activities*, *Routine Tasks*, and *Socializing* consistently emerged as core structural components. The recurrence of these specific modules indicates that broad physical kinematics, daily routines, and social interactions form a universal semantic backbone for action representation.

Beyond this shared backbone, the network topologies diverged. The human-derived network exhibited a highly modular structure, uniquely utilizing *Hand-related Tasks* as a topological bridge between gross motor and social communities, alongside a distinct *Body / Facial Movement* cluster. The text-only models reorganized action semantics by introducing a *Caregiving* community. DeepSeek-R1 produced a denser, more integrated graph than Qwen2.5-7B, successfully recovering the *Body / Facial Movement* module observed in the human baseline. The multimodal networks displayed distinct topologies: Qwen2.5-VL-7B formed a structure featuring a unique *Symbolic / Ritualistic Activities* module, whereas the Qwen2.5-VL-72B model collapsed into a highly interconnected, centralized network without distinct peripheral branches. Chord diagrams provided complementary visualizations of these system-specific topologies (Extended Data Fig. 3a).

To quantify the representational alignment between the diverse computational systems and human behavior, we calculated the Pearson correlation between the vectorized upper triangles of each system's RSM and the human RSM (Fig. 4b). Human split-half reliability ($r_{hh} = 0.71$) was included as an empirical upper bound. Foundation models (both LLMs and MLLMs)

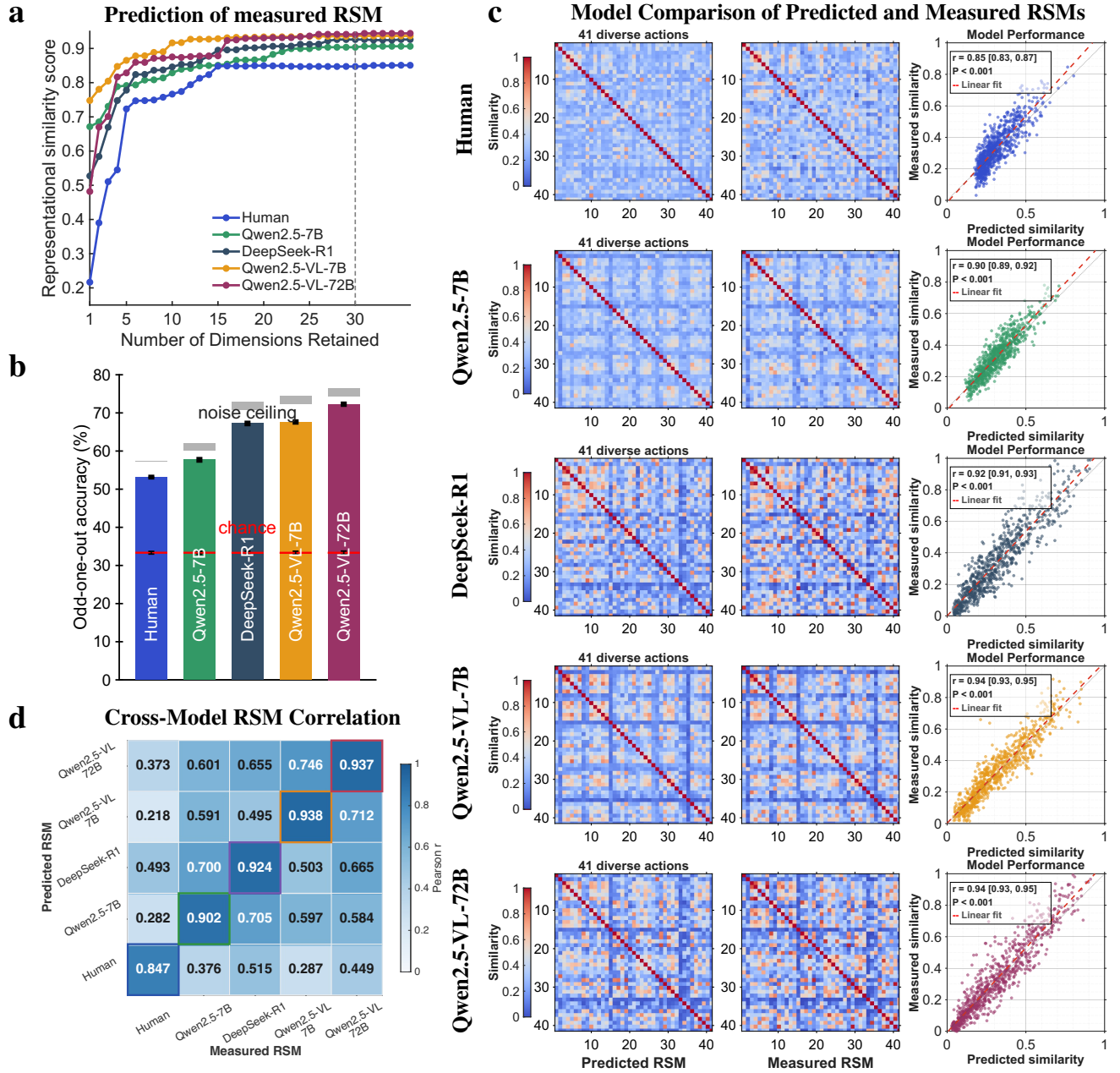


Figure 2. Dimensionality, predictive validation, and cross-system alignment of the extracted action spaces. a, Dimensionality saturation curves. Predictive performance on held-out data plateaus at approximately 30 dimensions across the human benchmark, text-only LLMs, and MLLMs. **b,** Prediction accuracy on an independent held-out odd-one-out test set. The dashed line delineates chance-level performance (33.3%), and the gray shaded band indicates the estimated noise ceiling. Accuracy reflects how well each 30-dimensional embedding predicts novel choices from its own source system, and the error bar reflects 95% CIs ($n=1,000$ bootstraps). **c,** Reconstruction of empirical similarity structures for a balanced subset of 41 representative actions. For each system, panels display (from left to right): the SPoSE-predicted RSM, the empirically measured behavioural RSM, and the corresponding scatter plot. Pearson correlation coefficients (r) indicate within-system reconstruction performance (Human: $r = 0.85$; LLMs: $r = 0.90$, $r = 0.92$; MLLMs: $r = 0.94$, $r = 0.94$). **d,** Cross-system RSM correlation matrix. The diagonal displays within-system reconstruction fidelity. Off-diagonal elements indicate representational alignment across systems, showing stronger intra-modality than cross-modality correlations. Among the evaluated artificial models, the text-only DeepSeek-R1 shows the highest bidirectional correlation with the human benchmark.

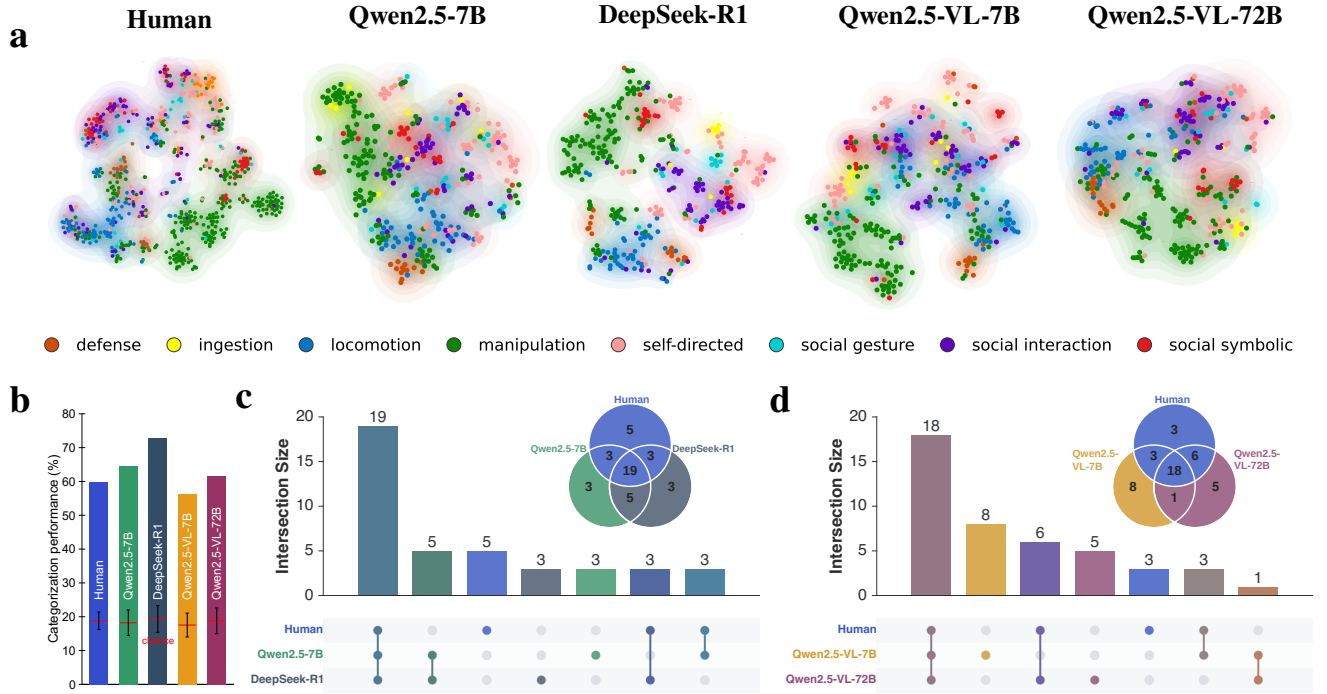


Figure 3. Semantic topology and modality-dependent differences in action representations. **a**, t-SNE projections of the action latent spaces. Points are color-coded by eight high-level categories. To reflect the underlying experimental designs, the human plot comprises 768 exemplars distributed across 256 categories, whereas the AI plots map 355 single-exemplar categories. Both text-only and multimodal artificial models demonstrate category-related organization without explicit categorical supervision during the odd-one-out task. **b**, Nearest-centroid classification accuracy for eight predefined macro-categories. The dashed red line denotes the chance level estimated by permutation testing, and the black intervals indicate the corresponding 95% CIs. Text-only LLMs achieved higher accuracies than visually grounded MLLMs, indicating greater separability of the predefined category labels in the text-derived representational spaces. **c-d**, UpSet and Venn diagrams illustrating the intersection of SPoSE dimensions across systems. A subset of matched dimensions is shared across both the LLM family (**c**) and the MLLM family (**d**). Within the Qwen2.5-VL family (**d**), the 72B model shows greater exclusive overlap with the human benchmark than the 7B model. In contrast, Qwen2.5-7B and DeepSeek-R1 (**c**) exhibit a shared subset of dimensions matched to the human benchmark alongside distinct model-specific dimensions.

demonstrated robust alignment with human judgments. The strongest model–human correlations were observed for gpt-oss-120B ($r = 0.53$), DeepSeek-V4-284B ($r = 0.53$), MiniMax-V2.7 ($r = 0.51$), and Qwen2.5-VL-72B ($r = 0.51$). Notably, these foundation models vastly outperformed conventional video encoders, which yielded substantially lower alignments (e.g., PerceptionEncoder, $r = 0.22$; VideoChat, $r = 0.22$; VideoMAE, $r = 0.06$). This stark performance gap highlights that standard spatiotemporal feature extraction is insufficient to recover human-like action similarity, underscoring the necessity of abstract semantic integration. Variance partitioning analysis (Extended Data Fig. 3b,c; see Methods) further confirmed that model–human alignment is primarily driven by high-level action semantics (*category* and *activity* contexts) rather than surface-level target similarities.

Feature attribution and semantic modulation of generated actions

To objectively evaluate the functional relevance of the extracted SPoSE dimensions (Supplementary Figs. 4–7), we developed a two-stage framework comprising feature attribution in auxiliary projection models and dimension-guided generative modulation.

First, we examined how dimension activations predicted by the auxiliary models relate to raw sensory inputs (Fig. 5a; Extended Data Fig. 4). Using Grad-CAM²⁷ for visual inputs and Grad-ECLIP²⁸ for textual inputs (see Methods), we generated attribution maps for the predicted SPoSE dimensions. These maps successfully highlighted behaviorally relevant features rather than background regions. For example, in a *singing* scene evaluated using the DeepSeek-R1 space, the text-based projection assigned high attribution to the tokens “singing” and “mike” for the *blowing / mouth-related* dimension, while the visual projection highlighted the performer’s facial region. To quantify sensitivity, we applied saliency-guided spatial masking. Selectively occluding high-attention visual regions—such as the hand and pot in a *removing* action—drastically reduced the

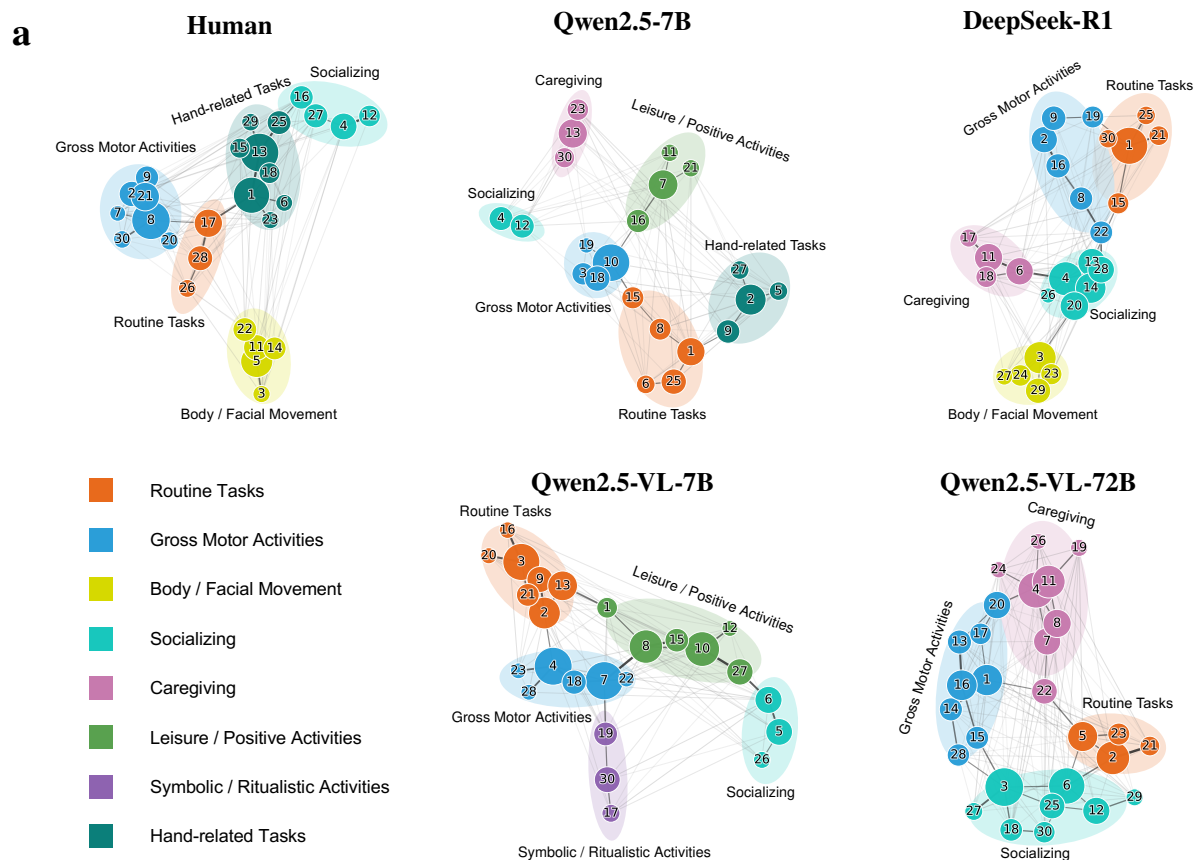
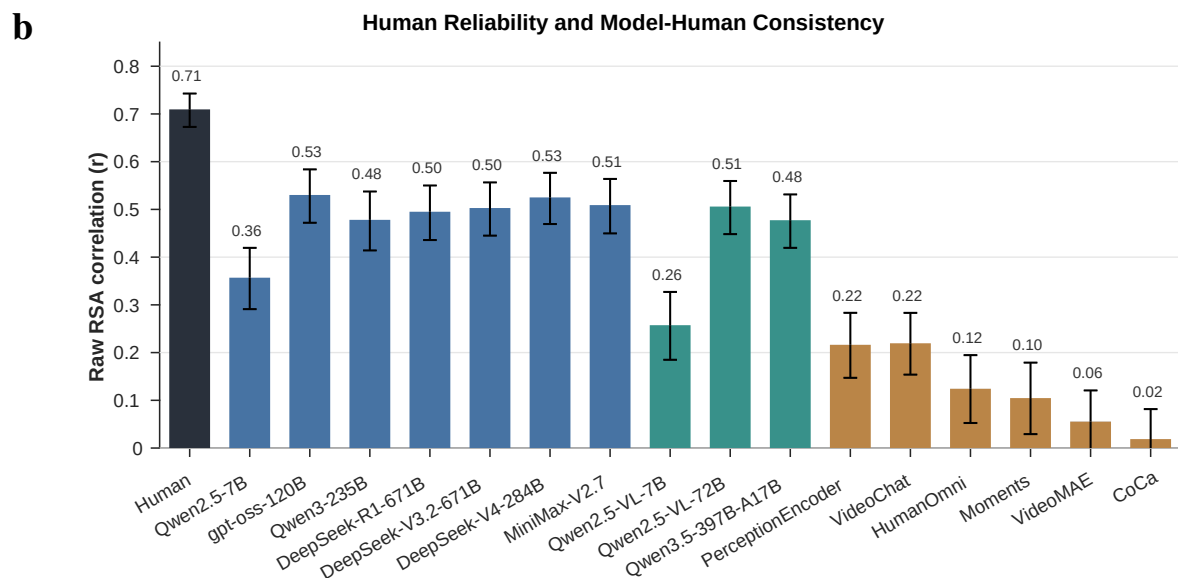


Figure 4. Network topology and representational alignment with human behaviour. **a**, Community structure of the extracted action spaces visualized as undirected graphs. Nodes represent individual SPoSE dimensions, with node size scaled by PageRank centrality. Edges indicate inter-dimensional Pearson correlations exceeding $|r| > 0.19$. **b**, Representational alignment across computational architectures. Bars display the Pearson correlations between each model’s RSM and the human RSM. Left black bar indicates human split-half reliability, serving as an empirical noise ceiling. Error bars indicate 95% CIs derived from 1,000 bootstrap resamples of the 820 unique category pairs. No correction for attenuation was applied.

164 predicted target-dimension activation for Qwen2.5-VL-7B from 0.686 to 0.098 (Supplementary Table 3). This reduction
165 demonstrates that the auxiliary prediction mechanism is functionally coupled to the localized semantic features identified by the
166 attribution maps.

167 Second, we tested whether individual SPoSE dimensions could serve as functional semantic coordinates for video synthesis.
168 We trained a cross-space Action Projector mapping the 30-dimensional SPoSE representation into a CLIP semantic space, which
169 subsequently guided a Stable Diffusion model via DDIM inversion (Extended Data Fig. 5; see Methods). Modulating specific
170 SPoSE dimensions produced predictable and corresponding semantic shifts in the generated video content (Fig. 5b). For
171 instance, starting with an indoor *assembling* action (Fig. 5b, left), increasing the *nature-related* dimension smoothly introduced
172 outdoor environmental cues. From this intermediate state, further increasing the *extreme sports-related* dimension shifted the
173 setting to a snowy mountain environment, whereas alternatively increasing the *hand-related* dimension produced a close-up of
174 hands manipulating soil. Sequential modulations on a *bumping* scene (Fig. 5b, right) similarly demonstrated complementary
175 shifts in contextual and social features (Extended Data Figs. 6, 7, 8).

176 These interventions objectively demonstrate that the extracted action dimensions are not merely statistical artifacts of the
177 dimensionality reduction process, but operate as functionally meaningful semantic coordinates within the generative pipeline.

178 **Cross-dataset transfer of model-derived action geometry predicts distributed cortical responses**

179 To determine whether the extracted semantic geometry captures neurobiologically grounded structures beyond its source stimulus
180 distribution, we evaluated the fixed 30-dimensional SPoSE spaces on an independent neuroimaging dataset (HAD-fMRI). This
181 dataset contains cortical BOLD responses from 30 participants viewing 21,600 unseen clips spanning 180 action categories.
182 Crucially, no HAD judgments or fMRI data were used to optimize the embeddings or their transfer mappings. Instead, we
183 estimated a ridge regression mapping from frozen Perception Encoder features to the SPoSE spaces using solely the 355 S-MiT
184 training videos, and applied this mapping zero-shot to the HAD clips (see Methods). We then employed voxel-wise encoding
185 and searchlight RSA to test whether this generalized geometry predicts stimulus-evoked cortical responses (Fig. 6, Extended
186 Data Fig. 9, and Supplementary Figs. 11–13).

187 Whole-brain encoding analysis revealed a consistent spatial distribution of predictive accuracy, with notable magnitude
188 variations across the evaluated foundation models (Fig. 6a; Supplementary Figs. 8–10). Rather than being confined to
189 early visual areas, encoding performance peaked across a distributed network associated with high-level action perception.
190 ROI-averaged evaluations (Fig. 6b) confirmed significant predictive performance across lateral temporal and occipital regions
191 (e.g., MT, MST, FST, LO1, LO2), the intraparietal sulcus (IPS0-IPS2), the temporo-parieto-occipital junction (TPOJ3), and
192 face-selective regions (FFC). Significant alignment also extended into the primary somatosensory (S1, S2) and motor/premotor
193 (M1, PM) cortices, consistent with the vicarious engagement of bodily representations during action observation²⁹. Notably,
194 pairwise statistical comparisons demonstrated that the text-derived DeepSeek-R1 embedding significantly outperformed the
195 visually grounded Qwen2.5-VL-72B embedding across nearly all of these higher-order visual, parietal, and sensorimotor ROIs.

196 To characterize the functional selectivity of these regions, we extracted regional tuning profiles from the voxel-wise fractional
197 ridge regression weights (Fig. 6c and Extended Data Fig. 10; see Methods). FST exhibited strong positive weights for the
198 *exercise-related* and *hand-related* dimensions, aligning with topographical gradients for action effectors and graspability in
199 the lateral occipitotemporal cortex¹⁹. Concurrently, regions such as MT and TPOJ3 were heavily weighted toward *cheering*
200 / *crowd-related*, *playing*, and *sports-related* dimensions, reflecting a sensitivity to multi-agent and socially contextualized
201 action features. The FFC was distinctly tuned to the *smiling* dimension, consistent with its specialized role in facial processing.
202 Collectively, these zero-shot cross-dataset predictions demonstrate that the semantic coordinate systems recovered from
203 foundation models maintain robust biological validity. The compelling performance of text-derived spaces further substantiates
204 the profound influence of conceptual and linguistic scaffolding on the cortical organization of action representation.

205 **Discussion**

206 Whether artificial systems represent natural actions in a human-aligned manner has remained an unresolved question. Recognition
207 accuracy alone is insufficient to establish representational alignment, as comparable performance can arise from divergent
208 computational strategies^{8,19}. Human action perception integrates perceptual kinematics with abstract features such as goals,
209 object relations, and social contexts^{6,30,31}. By evaluating LLMs and MLLMs within a shared machine-behavioral framework,
210 we identified a low-dimensional semantic geometry that organizes naturalistic actions. This structure predicts human behavioral
211 judgments better than conventional visual encoders, supports targeted semantic modulation, and successfully predicts distributed
212 cortical responses. Together, these findings indicate that a human-aligned semantic organization emerges spontaneously in
213 foundation models, providing a shared computational space across artificial systems, human behavior, and the brain^{11,32,33}.

214 The correspondence between model and human representations was established using complementary stimulus sampling
215 regimes. The model evaluation utilized one representative clip per category across 355 categories, capturing broad categorical
216 structure. In contrast, the human benchmark sampled 256 categories with three exemplars each, capturing within-category



Figure 5. Feature attribution and dimension-guided modulation of generated actions. **a**, Attribution and spatial masking for DeepSeek-R1 and Qwen2.5-VL-7B across four representative action scenes (*singing*, *bandaging*, *rubbing*, and *removing*). Rows display original frames and text descriptions, followed by modality-specific attribution maps (Grad-CAM for visual regions, Grad-ECLIP for textual tokens) highlighting features associated with the targeted SPoSE dimension. Hatched overlays indicate targeted spatial masking of high-attention regions, leading to corresponding reductions in the predicted dimension activations (e.g., 0.686 → 0.098 for cooking/removing). **b**, Dimension-guided video generation utilizing the Qwen2.5-VL-7B space. Starting from source videos (*assembling*, left; *bumping*, right), specific dimension values in the SPoSE embedding were modified prior to conditional video generation. Editing Step 1 illustrates the visual output following a single dimension modification (e.g., increasing *nature-related*). Editing Step 2 illustrates parallel branching outputs after subsequent, additive dimension modifications (e.g., branching into *hand-related* versus *extreme sports-related*). These interventions demonstrate that manipulating the extracted dimensions yields corresponding semantic shifts in the synthesized scene context and action content.

variance¹⁶. The alignment observed over the shared categories indicates that the recovered semantic organization persists across independently sampled video exemplars. This cross-dataset generalization suggests that the foundation models capture abstract category-level action structure rather than relying solely on item-specific visual statistics. Furthermore, the recovered dimensions systematically relate to action targets, broad categories, and everyday activities. This pattern aligns with evidence that human action representations are organized by behavioral goals, interaction scale, and semantic contexts rather than

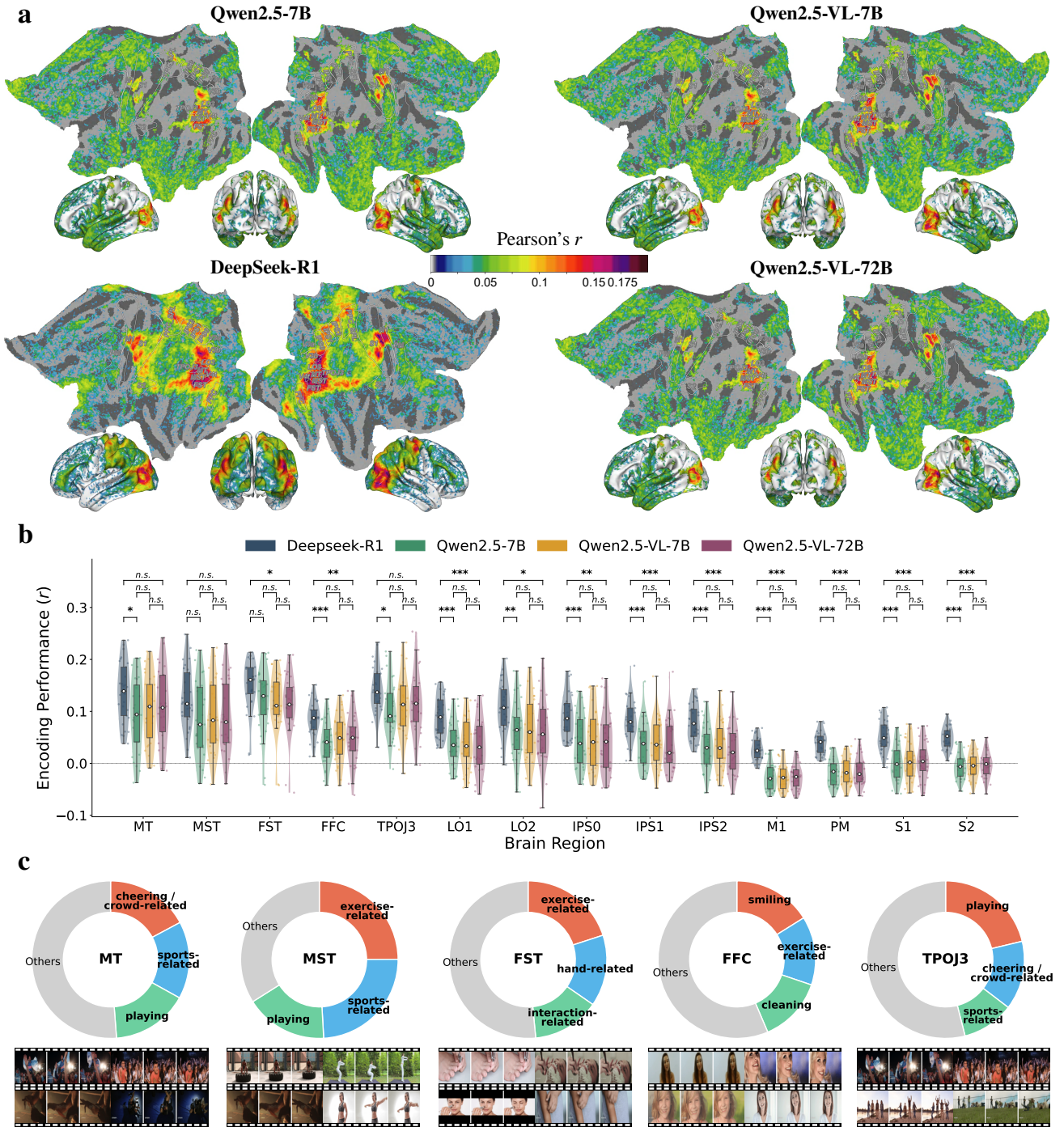


Figure 6. Cross-dataset prediction of distributed cortical responses using model-derived action geometries. **a**, Flattened cortical maps displaying group-level voxel-wise encoding accuracy (Pearson's r , averaged across $n = 30$ subjects). Colored voxels denote statistically significant predictions ($P < 0.05$, FDR corrected, two-tailed t-test). **b**, ROI-averaged encoding performance. Violin plots show the distribution of individual subject accuracies, overlaid with boxplots (median, interquartile range [IQR], and $1.5 \times$ IQR whiskers). White-filled circles indicate medians. Brackets denote pairwise statistical comparisons (two-tailed Welch's t-test, Benjamini-Hochberg FDR corrected; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; n.s., not significant). **c**, Regional neurosemantic tuning profiles. Donut plots illustrate the action dimensions with the largest positive regression weights within five specific ROIs (data shown for Qwen2.5-VL-7B), demonstrating region-specific tuning. Below, the top four representative HAD videos exhibiting the highest model-predicted regional similarity are shown.

kinematics alone^{6,16}.

Beyond their overall alignment with human judgments, the recovered spaces exhibited modality-dependent characteristics. Text-only LLMs demonstrated higher nearest-centroid category separability than visually grounded MLLMs. This difference is consistent with language models compressing representations around semantic regularities, whereas multimodal models retain graded perceptual variation related to kinematics, body form, and scene context¹³. Model capacity provided a secondary source of variation, with larger models within the same family (e.g., Qwen2.5-VL-72B) sharing more dimensions with the human benchmark than their smaller counterparts. However, size-related effects cannot be isolated from differences in training data and post-training procedures, which are known to substantially influence representational alignment and brain predictivity^{8,34,35}.

To objectively evaluate the functional relevance of this recovered geometry, we employed dimension-guided generative modulation. Modulating individual dimensions produced human-interpretable shifts in the corresponding visual content of synthesized videos. Rather than serving merely as descriptive statistical artifacts of the dimensionality reduction process, these dimensions functioned as actionable semantic coordinates capable of guiding visual outputs within the constructed pipeline. This operational control supports the interpretation of these axes as functionally meaningful representations of action content. Importantly, these interventions demonstrate operational modulation within the specific auxiliary projection framework utilized; they are not intended to establish that these dimensions exist as fully independent or compositionally separable causal variables native to the original foundation models' hidden states.

This behavioral and functional correspondence was paralleled by the prediction of stimulus-evoked cortical responses. Model-derived action dimensions predicted fMRI responses across motion-sensitive visual, lateral temporal, temporoparietal, and sensorimotor cortices. This distribution aligns with theoretical models of complementary perceptual and conceptual action codes across the dorsal and ventral pathways^{36,37}, as well as the engagement of whole-body action planning networks^{29,38}. Notably, these neural predictions were generated by transferring the fixed geometry learned from the original 355 categories to an independent dataset of 21,600 unseen videos using a frozen perception encoder. Because the semantic spaces were not re-estimated on the neuroimaging stimuli, their predictive performance effectively rules out the possibility of overfitting to specific stimulus sets. This zero-shot cross-dataset generalization indicates that the recovered coordinates capture a robust action structure that generalizes across distinct visual distributions and experimental contexts³⁹.

Several limitations constrain the interpretation of these findings and identify directions for future research. First, the human and model spaces were estimated under unmatched sampling regimes. While this rigorously tests category-level generalization, establishing item-level correspondence under identical perceptual input requires matched-stimulus designs. Second, the heterogeneity in architecture, training diets, and modalities across the evaluated foundation models prevents the isolation of specific mechanistic drivers of representational alignment. Future work should employ controlled comparisons with matched architectures and encoder-only baselines to disentangle these factors³⁴. Third, the generative editing and masking analyses relied on auxiliary modules, demonstrating control over the constructed pipeline rather than direct causal intervention on the foundation models' internal states.

Finally, because action perception unfolds dynamically over time³⁷, future studies should align frame-resolved model trajectories with temporally precise neural measurements, such as MEG or EEG, as the BOLD responses used here cannot resolve these rapid dynamics²⁰. Despite these limitations, the present framework demonstrates how representations extracted from artificial foundation models can serve as computational probes for understanding human action cognition. By identifying a convergent semantic geometry across behavior, functional modulation, and cortical activity, this work provides a structured basis for investigating and aligning artificial action representations with human conceptual frameworks.

Methods

Experimental stimuli and behavioural paradigms

Experimental stimuli and content descriptions for AI models

To probe action representations in LLMs and MLLMs, we obtained our primary video stimuli from the Spoken Moments in Time (S-MiT) dataset²⁶. Of the original 366 categories, we retained 355 categories depicting human actions and excluded 11 categories representing non-human events. Retaining 355 of the 366 original categories preserved a wide sample of naturalistic human-action concepts while equalising the number of selected clips across categories. The one-clip-per-category design was chosen to recover a broad category-level geometry under a fixed per-category stimulus budget. It was not intended to estimate within-category variability, which was instead represented more richly in the independent human benchmark.

To evaluate text-only LLMs, equivalent textual descriptions were required. We utilized the Google Automatic Speech Recognition (ASR) engine outputs provided in the S-MiT dataset, which transcribe the audio descriptions accompanying each video. The transcripts were independently edited by three annotators to correct transcription errors and retain only visually observable action content. Editors corrected lexical and grammatical errors, removed inferred intentions and non-visual contextual information, and did not introduce content unsupported by the corresponding video. The resulting captions served as textual stimuli for the odd-one-out behavioural tasks.

Human behavioural benchmark and stimuli

All human behavioural choices and corresponding visual stimuli analysed in this study were obtained from the large-scale triplet odd-one-out experiment conducted by Bockes et al.¹⁶. Bockes et al. curated 256 human-action categories from the Moments in Time (MiT) dataset⁴⁰ using a hierarchical clustering procedure based on ResNet-50 representations. Three representative one-second video exemplars were selected for each category, yielding 768 visual stimuli.

The raw behavioural dataset collected by Bockes et al. contained 1,252,704 triplet judgements, organised into 39,147 sets. A set was defined as a block of 32 trials completed by one Amazon Mechanical Turk worker. Bockes et al. excluded workers showing excessively fast responses ($\geq 12.5\%$ of trials below 900 ms and $\geq 50\%$ below 1,200 ms in at least three 32-trial sets) or deterministic position biases ($\geq 50\%$ of responses to the same position in at least six sets). Remaining trials with response times below 900 ms were subsequently removed. Following the exclusion criteria reported by Bockes et al., the final dataset contained 779,467 valid triplet judgements, which we used without further participant-level filtering.

To estimate between-participant choice consistency, Bockes et al. randomly sampled 1,000 triplets, each of which was judged 40 times across the participant pool. Thus, the design comprised 40 judgements of each triplet, rather than 40 independent repetitions of the entire experiment. For a repeated triplet t , let $n_{t,r}$ denote the number of retained judgements in which stimulus r was selected as the odd one out. Triplet-level consistency was defined as the proportion of judgements corresponding to the modal response:

$$C_t = \frac{\max_{r \in \{1,2,3\}} n_{t,r}}{\sum_{r=1}^3 n_{t,r}}.$$

Overall consistency was calculated as the unweighted mean across the 1,000 repeated triplets:

$$C = \frac{1}{1000} \sum_{t=1}^{1000} C_t.$$

Bockes et al. reported an overall human choice consistency of 57.32%, compared with a three-alternative chance level of 33.3%.

Cross-system intersection and inferential scope. The model stimulus set and the published human behavioural benchmark differed in sampling emphasis. The model evaluation retained 355 human-action categories from S-MiT, with one representative clip per category, thereby prioritising categorical breadth. The human benchmark comprised 256 categories with three exemplars per category, thereby providing greater within-category depth. Direct comparisons were restricted to categories shared by both sets. Of the 256 categories in the human dataset, 255 were also present in the model stimulus set, and all cross-RSM and human-model alignment analyses were conducted over these 255 categories. Because model and human judgements were not collected on identical videos and triplets, these analyses were interpreted at the category and representational-geometry levels. They test whether semantic organisation converges across independently sampled stimuli, but do not establish item-level agreement under matched perceptual input.

Triplet odd-one-out paradigm and sampling strategy for AI models

The 7,393,585 possible triplets from 355 action categories were computationally prohibitive to evaluate exhaustively. We therefore employed a pair-based sparse-sampling strategy. First, all possible unique category pairs were enumerated, yielding $\binom{355}{2} = 62,835$ dyads. For each pair, third categories were randomly collected from the remaining 353 categories. Candidate triplets were sorted, and duplicates were removed. This procedure yielded 547,335 unique training triplets, covering 7.40% of the complete triplet space. Each category appeared in 4,625.37 triplets on average and each pair appeared 26.13 times on average.

A test set of 50,000 triplets and a repeated-response set of 2,000 triplets were sampled without replacement from the remaining space. The training, test, and repeated-response sets were mutually exclusive. Each repeated triplet was queried 20 times per model. Empirical within-model consistency was defined as the mean proportion of responses matching the modal choice across the 2,000 triplets. Furthermore, to ensure a balanced evaluation across broad taxonomic classes, we selected a representative subset of 41 distinct actions (uniformly distributed across classes such as *ingestion*, *locomotion*, *manipulation*, *social*, etc.) to comprehensively validate categorical structure alignment.

Odd-one-out task via text-only LLMs

We queried Qwen2.5-7B-Instruct²⁴ (model identifier: Qwen/Qwen2.5-7B-Instruct) and DeepSeek-R1-0528 (671B)²³ (model identifier: deepseek-r1:671b-0528) through the OpenAI-compatible APIs provided by SiliconFlow (endpoint: <https://api.siliconflow.cn/v1>) and the China Science and Technology Cloud (CSTCloud; endpoint: <https://uni-api.cstcloud.cn/v1/>), respectively. Each prompt contained three action labels and their corresponding textual video descriptions, followed by the odd-one-out instructions described below. To mitigate autoregressive positional bias, we iteratively refined our prompt to explicitly instruct the model to perform internal sequence shuffling (Supplementary Fig. 2). Generation used provider defaults: temperature and top_p were 1.0 and 1.0 for both Qwen2.5-7B and DeepSeek-R1. For Qwen2.5-7B, only

the final `content` stream was stored; for DeepSeek-R1, `reasoning_content` and `content` were recorded separately, with the latter parsed to obtain the selected item. The exact prompt provided to the LLMs, injecting the textual captions, was as follows:

*Here are three video descriptions each belongs to a specific category. The first video description corresponds to **category_1: caption_1**; the second video description corresponds to **category_2: caption_2**; the third video description corresponds to **category_3: caption_3**.*

Now we need to perform a role-playing task. Assume you are an action judgement expert with the ability to classify actions performed by humans. Firstly, you should analyse the content and figure out the hidden dimensions individually. Then you should perform a second-level comparison of the pairs to identify the best option with strongly supported evidence you discover. Ensure the comparisons are not influenced by the order of priority in which the descriptions are presented. To mitigate positional bias, follow these steps: Mentally shuffle the options into different sequences (e.g., reverse order, random permutation). Analyse each action's merits/demerits in each shuffled configuration. Identify if any option consistently emerges as superior across different orderings. If inconsistencies arise, conduct content-based comparisons using criteria: factual accuracy, logical coherence, contextual relevance, completeness of information. Document your reasoning process transparently before finalizing the selection. Then, tell me which human action that is noticeably different from the other descriptions. Explain the reason for this difference. Finally, give me the answer. (The answer format is "Video description ID is noticeably different from the other two.")

After response validation, the final datasets contained 536,457 training and 49,976 test responses for Qwen2.5-7B, and 542,359 training and 49,749 test responses for DeepSeek-R1-671B.

Odd-one-out task via multimodal LLMs

For the visually grounded MLLMs, we evaluated two Qwen2.5-VL checkpoints²⁵: a locally hosted Qwen2.5-VL-7B model and a provider-hosted Qwen2.5-VL-72B model queried through the CSTCloud API described above (model identifier: `qwen2.5-vl:72b`). The prompting strategy followed the LLM pipeline with adaptations for multimodal input. Each user-only prompt presented three action clips, requested a description of each action, and instructed the model to identify and justify the odd one out after mentally permuting the presentation order. The final answer was required in a fixed format. For both the 7B and 72B models, eight frames were sampled uniformly across each video. Temperature and `top_p` were both set to 1.0, and only the final response content was retained. The exact prompt was as follows:

*Now we need to perform a role-playing task. Here are three videos which depict actions performed by humans. **Video Input_1, Video Input_2, Video Input_3***

Assume you are an action judgement expert with the ability to classify actions performed by humans. Firstly, you should describe the content of these clips individually. Then you should carefully evaluate each content independently of its order and distinguish which action is noticeably different from the other videos. Explain the reason for this difference. To mitigate positional bias, follow these steps: Mentally shuffle the options into different sequences (e.g., reverse order, random permutation). Analyse each action's merits/demerits in each shuffled configuration. Identify if any option consistently emerges as superior across different orderings. If inconsistencies arise, conduct content-based comparisons using criteria: factual accuracy, logical coherence, contextual relevance, completeness of information. Document your reasoning process transparently before finalizing the selection. Finally, give me the answer. (The answer format is "Video[number] is noticeably different from the other two".)

After response validation, the final datasets contained 521,912 training and 48,849 test responses for Qwen2.5-VL-7B, and 547,335 training and 50,000 test responses for Qwen2.5-VL-72B.

Additional models for behavioural-consistency analysis

To broaden the architecture comparison in Fig. 4b, we additionally collected triplet odd-one-out judgements from five text-only foundation models: `gpt-oss-120b`⁴¹, `Qwen3-235B-A22B`⁴², `DeepSeek-V3.2 (671B)`⁴³, `DeepSeek-V4-Flash (284B)`⁴⁴, and `MiniMax-M2.7`⁴⁵. All provider-hosted additional models were queried through the CSTCloud API described above, using the following model identifiers: `gpt-oss-120b`, `qwen3:235b`, `deepseek-v3.2`, `deepseek-v4-flash`, and `minimax-m27`. Each trial contained three action-category labels and their corresponding manually corrected video descriptions. The models received the same role-playing, pairwise-comparison, mental-shuffling, and fixed-answer-format instructions used for the primary text-only LLMs. When returned by the API, `reasoning_content` and `final_content` were accumulated and stored separately, and only the final content was parsed to obtain the selected video-description identifier. If no identifier could be parsed from a response, the same trial was queried again for up to ten attempts.

We also evaluated the multimodal Qwen3.5-397B-A17B model⁴⁶ using the CSTCloud API identifier qwen3.5. Each clip was represented by eight frames sampled at uniformly spaced decoded-frame indices from the first to the last frame, yielding 24 images per triplet. Frames were JPEG-encoded at their decoded spatial resolution and supplied as ordered image inputs labelled *Video 1*, *Video 2*, and *Video 3*. The prompt required the model to describe each clip, compare the actions independently of their presentation order, mentally permute the alternatives, and return the outlier using the fixed format specified for the other MLLMs. When available, the reasoning and final-answer streams were stored separately. Any `<think>...</think>` block embedded in the final content was removed before response parsing. Unparseable responses were retried up to ten times.

Baseline feature extraction

We extracted features from a set of baseline vision models with different architectures and training objectives. These models included approaches based on spatiotemporal convolution, masked autoencoding, and contrastive vision-language pre-training. Model outputs were converted into fixed-length video-level feature vectors using aggregation procedures appropriate to their output structures.

For Transformer-based video models, including VideoMAE⁴⁷ and the visual encoders of VideoChat-Flash-Qwen2.5-2B⁴⁸ and HumanOmni⁴⁹, we extracted the final hidden states. These representations were mean-pooled across time, followed by a spatially weighted summation across patch tokens. For the ResNet3D-50 model pre-trained on Moments in Time⁴⁰, features were obtained from the pre-logits global average pooling layer. For CoCa⁵⁰, frame-wise spatial embeddings were aggregated across time to obtain video-level representations. For the Perception Encoder⁵¹ (PE-Core-G-14-448), we used its L_2 -normalized video-level embeddings directly in subsequent analyses.

Derivation and characterization of latent action spaces

Sparse Positive Similarity Embedding (SPoSE)

To reconstruct the latent similarity space from the discrete triplet judgements, we used the Sparse Positive Similarity Embedding (SPoSE) algorithm¹⁵. Each action n was represented by an embedding vector $\mathbf{x}_n \in \mathbb{R}^d$. Each observed triplet was ordered as (i, j, k) , where i and j constituted the selected most-similar pair and k was the odd one out. The probability of selecting (i, j) was defined as

$$\Pr((i, j) \mid (i, j, k)) = \frac{\exp(\mathbf{x}_i^\top \mathbf{x}_j / \tau)}{\exp(\mathbf{x}_i^\top \mathbf{x}_j / \tau) + \exp(\mathbf{x}_i^\top \mathbf{x}_k / \tau) + \exp(\mathbf{x}_j^\top \mathbf{x}_k / \tau)}, \quad (1)$$

where the softmax temperature τ was set to 1. Let $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]^\top$ denote the embedding matrix. The model was fitted by minimizing the mean negative log-likelihood together with sparsity and non-negativity penalties:

$$\hat{\mathbf{X}} = \arg \min_{\mathbf{X}} \left[-\frac{1}{|\mathcal{B}|} \sum_{(i, j, k) \in \mathcal{B}} \log \Pr((i, j) \mid (i, j, k)) + \frac{\lambda}{N} \|\mathbf{X}\|_1 + \gamma \|\mathbf{X}\|_+ \right], \quad (2)$$

where $[z]_+ = \max(z, 0)$, λ controls sparsity, and $\gamma = 0.01$ penalizes negative embedding weights. We used the publicly available PyTorch implementation at <https://github.com/VicCo-Group/SPoSE> and optimized the model using Adam with a learning rate of 0.001. The regularization coefficient λ was set to 0.008 for the human data, 0.012 for DeepSeek-R1-671B, 0.025 for Qwen2.5-7B, 0.0072 for Qwen2.5-VL-7B, and 0.0073 for Qwen2.5-VL-72B.

Reproducibility and selection of embedding dimensions

To assess embedding stability, we performed 65 independent SPoSE runs for each system using different random initializations. For each dimension in a given run, we identified the dimension in every other run with the highest Pearson correlation. These correlations were Fisher z -transformed, averaged across runs, and inverse-transformed to obtain a reproducibility score for each dimension. Dimensions were subsequently ranked by their total embedding weight across actions. To remove within-system redundancy, when two dimensions had correlated loading profiles ($r \geq 0.3$), the lower-weight dimension was excluded. This manually selected threshold minimized conceptual overlap while preserving the broadest coverage of action concepts. The procedure was applied independently to each system and retained 30 dimensions per system.

Semantic labeling of action dimensions

For each retained dimension, we selected the nine actions with the highest embedding weights as representative exemplars. Nine independent participants examined these exemplars separately and provided short free-text descriptions of their shared semantic content. Their responses were aggregated by term frequency to generate a word cloud for each dimension. The annotation study was approved by the Ethics Committee of CASIA (IA21-2511-020201), and all participants provided informed consent.

For the human embedding, labels from Bockes et al.¹⁶ were retained when they were consistent with the current exemplars and participant descriptions. The present analysis additionally identified three dimensions not reported in the previous analysis, which were labeled *hand*, *body care*, and *violence* based on recurring participant descriptions.

The same human annotation procedure was applied independently to each AI embedding. Qwen3-235B-A22B was also prompted to propose candidate labels from the corresponding top-nine actions. Final labels were assigned by integrating these candidates with the human-generated word clouds and were retained only when they were consistent with both the recurring human descriptions and the representative actions. The LLM therefore served only as an auxiliary labeling tool.

Cross-system matching and intersection of SPoSE dimensions

Dimension intersections were evaluated over the 255 action categories shared by the human and artificial stimulus sets. After removing the three human exemplars belonging to the non-overlapping category, each 255-row model embedding was expanded by repeating every category row three times, yielding matrices aligned to the 765 human exemplars. For each pair of systems, we calculated Pearson correlations between all dimension-loading profiles. Dimensions in the second system were ordered according to the dimension in the anchor system with which they showed the maximum correlation. An adjusted matching-score matrix was then obtained by subtracting the off-diagonal within-anchor dimension-correlation matrix, with zero padding or truncation when the two systems contained different numbers of dimensions. Dimensions with adjusted scores of at least 0.2 were matched using a sequential one-to-one procedure. Human dimensions were matched first and independently to each of the two model spaces; remaining unmatched dimensions were then matched between the two models. Each dimension could be assigned only once. The resulting matched groups were classified into the seven possible three-set membership patterns and summarized using UpSet and Venn diagrams. The procedure was applied separately to the human-Qwen2.5-7B-DeepSeek-R1 comparison and the human-Qwen2.5-VL-7B-Qwen2.5-VL-72B comparison.

Human-model representational similarity and geometric alignment

We quantified human-model correspondence using the balanced subset of 41 action categories described above. In the published human benchmark, each category was represented by three exemplars, yielding 123 target stimulus objects. We retained human odd-one-out trials containing at least two of these exemplars. After collapsing exemplar identities to the corresponding 41 category labels, the same triplets were used for subsequent AI evaluation. This matched-triplet design ensured that human and artificial systems were evaluated on an identical set of triplet comparisons. No new human data were collected. Trials whose third item fell outside the target subset were retained because they still provided evidence about the target pair.

Odd-one-out judgements were converted into 41×41 RSMs. The two non-odd items in each trial were treated as the similar pair. For system s , pairwise similarity was the proportion of valid judgements containing actions i and j in which they formed the similar pair:

$$\hat{p}_{ij}^{(s)} = \frac{\sum_{k \notin \{i,j\}} \sum_{q=1}^{R_{ijk}^{(s)}} \mathbb{I}[y_{ijkq}^{(s)} = (i,j)]}{\sum_{k \notin \{i,j\}} R_{ijk}^{(s)}}.$$

Here, $R_{ijk}^{(s)}$ is the number of valid judgements for triplet (i,j,k) , and $y_{ijkq}^{(s)}$ is the similar pair in judgement q . The matrices were symmetrized, with diagonal entries set to 1. This estimator was applied to human judgements and to the behavioural responses of the LLMs and MLLMs.

For feature-based comparison models, we calculated Euclidean distances between the extracted video-level feature vectors. Within each triplet, the item with the largest summed distance to the other two was designated the odd one out. The remaining pair was entered into the same RSM estimator, placing behavioural and feature-based systems in a common decision space.

RSMs were constructed for each artificial system and for the human judgements. For each system, model-human correspondence was quantified as the Pearson correlation r_{mh} between the vectorized upper triangles of the model and human RSMs, comprising the $\binom{41}{2} = 820$ unique category pairs. To estimate human internal consistency, the ordered human judgement list was divided at its midpoint and an RSM was constructed separately from each half. Human split-half reliability was then calculated as the Pearson correlation r_{hh} between the two vectorized human split-half RSMs and was included as an empirical upper bound. No correction for attenuation was applied. Confidence intervals (CIs) were estimated from 1,000 bootstrap resamples of the 820 category pairs. For each estimate, the 95% CI was calculated in Fisher- z space as $\text{atanh}(r) \pm 1.96\text{SD}[\text{atanh}(r_{\text{boot}})]$ and then transformed back to the correlation scale.

Multi-level semantic annotation of human actions

To systematically capture the semantic organisation of human actions in naturalistic contexts, we implemented a hierarchical classification framework informed by prior behavioural and neurocognitive research. Specifically, we categorized actions into

three distinct levels of granularity: *target*, *category*, and *activity*.

At the broadest level of abstraction, the *target* of an action indicates the primary entity toward which the behaviour is directed. Following established models of action processing^{17,30,52}, this level differentiates among actions directed toward an object, another person, or the self. This attribute is essential for distinguishing social interactions from object-directed manipulations.

At the intermediate level, the action *category* groups behaviours based on fundamental motor repertoires and behaviourally relevant classes, such as locomotion, ingestion, or manipulation^{5,17,53}. This level captures the shared functional objectives of different movements, independent of the specific environment.

At the most specific level, the *activity* defines the everyday context and routine in which the action occurs. These definitions were derived from the American Time Use Survey (ATUS), which is administered by the Bureau of Labor Statistics of the United States Department of Labor⁵⁴. By utilizing the structure of the ATUS, we ensured that the selected activities—such as preparing food, playing games, or performing housework—accurately represent the broad ecological distribution of human behaviours in daily life^{17,31}.

Evaluation of categorical structure and t-SNE visualization

To quantitatively assess the categorical structure of the learned embeddings across humans and artificial models, we performed a classification analysis based on the aforementioned medium-grained action categories (including defense, ingestion, locomotion, manipulation, self-directed behaviours, social gestures, social interactions, and social symbolic activities). We utilized a nearest-centroid classifier within a leave-one-out cross-validation (LOOCV) framework across all action stimuli. During each fold of the cross-validation, the centroid of a specific category was computed as the mean embedding vector of all stimuli belonging to that category, with the explicit exclusion of the held-out stimulus. The held-out stimulus was subsequently assigned to the category associated with the nearest centroid, as quantified by Euclidean distance. The overall performance was evaluated using mean classification accuracy. To establish an empirical baseline, the chance level of the classification was determined through 1,000 permutation tests in which the category labels of the stimuli were randomly shuffled.

Complementary to the quantitative classification, we employed t-distributed Stochastic Neighbor Embedding (t-SNE; dual perplexity: 5 and 30; 1,000 iterations) initialized with multidimensional scaling (MDS) to qualitatively visualize the topological organisation of the semantic spaces. This non-linear dimensionality reduction technique projects the high-dimensional feature vectors into a two-dimensional plane, thereby preserving the local similarities between action representations. Within the resulting visualizations, individual data points were color-coded according to the predefined action categories to visualize the distribution of predefined action labels in two dimensions.

Action graph analysis

To characterize relationships among the latent semantic dimensions, we represented the SPoSE embedding space as an undirected, weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Each node in $\mathcal{V}$ represented one SPoSE dimension. For each pair of dimensions, we calculated the Pearson correlation between their activation-weight vectors across the 355 stimulus actions. An edge was included when the absolute correlation exceeded a predefined threshold of $|r| > 0.19$, with $|r|$ used as the edge weight. This threshold produced a sparse graph for subsequent network analyses.

Community structure was estimated by optimizing modularity (Q) using the Louvain algorithm⁵⁵. The resulting communities comprised dimensions with relatively dense within-community connections in the thresholded graph. Nodes without supra-threshold connections were removed. Finally, PageRank centrality⁵⁶ was calculated for each remaining node. Dimensions with the largest PageRank values were described as high-centrality dimensions in the thresholded graph.

Variance partitioning

To partition the contributions of different semantic descriptions to model similarity structure, we constructed three categorical predictor RDMs representing the *target*, *category*, and *activity* levels. For each model, the outcome RDM was derived separately from its observed similarity judgements.

We performed leave-one-action-out cross-validation across the 355 stimulus actions. In each fold, seven regression models containing all non-empty subsets of the three predictor RDMs were fitted using the remaining actions. Predictive performance was evaluated on the corresponding held-out entries of the outcome RDM using squared Spearman’s rho-A (ρ_A^2). The scores were averaged across folds and combined using inclusion–exclusion to estimate the unique contribution of each predictor and the shared variance components among predictors.

Feature attribution, spatial perturbation, and dimension-guided generative modulation

Attribution analysis via Grad-ECLIP and Grad-CAM

To examine the linguistic and visual features associated with SPoSE-dimension activations predicted by auxiliary projection models, we performed modality-specific attribution analyses. For the text-only LLMs, we adapted Grad-ECLIP²⁸, a gradient-based attribution framework originally designed for image-text matching. We redirected the optimization target from cross-modal

similarity to the scalar activation of a specific SPoSE dimension. Specifically, the native latent representations of the text encoder, denoted as $\mathbf{f}_T$, were projected into the SPoSE space:

$$\mathbf{f}_{\text{SPoSE}} = \mathbf{f}_T \mathbf{W}_{\text{proj}} + \mathbf{b}_{\text{proj}}. \quad (3)$$

Here, $\mathbf{W}_{\text{proj}}$ and $\mathbf{b}_{\text{proj}}$ are the learned parameters of the linear projection head. The gradient target for dimension d was defined as the scalar activation $S_d = \mathbf{f}_{\text{SPoSE}}[d]$. Gradients of S_d were backpropagated to the output of the final attention layer, $\mathbf{o}_{\text{eos}}$. A dimension-specific importance score, $I_{i,d}$, was computed for each linguistic token i by combining the gradients with its value vector $\mathbf{v}_i$ and attention weight λ_i :

$$I_{i,d} = \text{ReLU} \left(\sum_k \frac{\partial S_d}{\partial \mathbf{o}_{\text{eos}}[k]} \cdot \lambda_i \cdot \mathbf{v}_i[k] \right), \quad (4)$$

where k indexes the feature channels. These scores were rendered as textual heatmaps to visualize token contributions to the projected SPoSE activations.

For the visually-grounded MLLMs, we established a homomorphic mapping between visual features and the behaviorally derived SPoSE space. A linear probe was trained via Mean Squared Error (MSE) minimization to project temporally averaged video features extracted by a pre-trained CLIP visual encoder⁵⁷ into the SPoSE space. Utilizing this differential projection, we employed Grad-CAM²⁷ to compute the gradients of the target dimension's activation with respect to the feature maps of the final convolutional layer of the CLIP encoder. The resultant heatmaps were spatiotemporally upsampled and superimposed on the original video frames, highlighting the visual regions.

Frame-wise spatiotemporal masking and perturbation analysis

To test whether the predicted action-dimension scores depended on regions identified by Grad-CAM, we applied frame-wise spatial perturbations. For a video comprising T frames, each Grad-CAM activation map $A_t \in \mathbb{R}^{H \times W}$ was normalized to $[0, 1]$. The frame-specific salient region was defined as

$$\Omega_t = \{p = (x, y) \mid A_t(p) > \tau\}, \quad \tau = 0.4. \quad (5)$$

The spatial coordinates in Ω_t were parameterized by a two-dimensional Gaussian with centroid $\boldsymbol{\mu}_t = (\mu_{x,t}, \mu_{y,t})$ and coordinate-wise standard deviations $\boldsymbol{\sigma}_t = (\sigma_{x,t}, \sigma_{y,t})$. The radius of the circular mask was defined as

$$r_t = \lfloor \lambda \max(\sigma_{x,t}, \sigma_{y,t}) \rfloor, \quad \lambda = 0.75. \quad (6)$$

For each frame, the targeted mask was centered on the Grad-CAM activation centroid:

$$M_t^{\text{target}}(p) = \begin{cases} 0, & \|p - \boldsymbol{\mu}_t\|_2 \leq r_t, \\ 1, & \text{otherwise.} \end{cases} \quad (7)$$

A shifted-control mask of the same radius was centered at a location $\boldsymbol{\mu}'_t$ sampled from the valid image domain, subject to $\|\boldsymbol{\mu}'_t - \boldsymbol{\mu}_t\|_2 \geq 2r_t$. Thus, both the targeted and control masks varied across frames while differing only in their spatial relationship to the Grad-CAM-localized region.

The masked frames, $I_t^c = M_t^c \odot I_t$, were passed through the frozen Perception Encoder, and action-dimension scores were estimated using the previously fitted linear regression probe. For each masking condition $c \in \{\text{target}, \text{control}\}$, the change in the target-dimension score was quantified as

$$\Delta s_d^c = s_d(V) - s_d(V^c). \quad (8)$$

Score changes were then compared between the targeted and shifted-control conditions.

Generative modulation of SPoSE dimensions

To objectively evaluate the functional relevance of the extracted SPoSE dimensions, we constructed a dimension-guided generative modulation pipeline. The framework mapped modulated SPoSE representations into a pre-trained semantic space, which subsequently conditioned video generation.

The core component was an Action Projector comprising a five-layer multilayer perceptron (MLP) with LayerNorm. It mapped a 30-dimensional SPoSE embedding, f_{SPoSE} , into the 768-dimensional joint latent space of CLIP:

$$f_{\text{proj}} = \text{MLP}_{\text{LN}}(f_{\text{SPoSE}}). \quad (9)$$

The projector output was aligned with the corresponding visual and textual embeddings of the source video using a tri-modal contrastive objective:

$$\mathcal{L}_{\text{CLIP}} = \mathcal{L}_{\text{BiInfoNCE}}(f_{\text{proj}}, f_{\text{vision}}) + \mathcal{L}_{\text{BiInfoNCE}}(f_{\text{proj}}, f_{\text{language}}). \quad (10)$$

552 The bidirectional InfoNCE loss was defined as

$$\mathcal{L}_{\text{BiInfoNCE}} = -\frac{1}{B} \sum_{i=1}^B \left[\log \frac{\exp(s(\hat{z}_i, z_i)/\tau)}{\sum_{j=1}^B \exp(s(\hat{z}_i, z_j)/\tau)} + \log \frac{\exp(s(\hat{z}_i, z_i)/\tau)}{\sum_{k=1}^B \exp(s(\hat{z}_k, z_i)/\tau)} \right], \quad (11)$$

553 where $s(\cdot, \cdot)$ denotes cosine similarity, B is the batch size, and τ is a learned temperature parameter.

554 Two auxiliary mean squared error losses were included. The first directly aligned the projected representation with the
555 language embedding:

$$\mathcal{L}_{\text{proj1}} = \|f_{\text{proj}} - f_{\text{language}}\|_2^2. \quad (12)$$

556 The second aligned an up-projected representation with the flattened token-level text features:

$$\mathcal{L}_{\text{proj2}} = \|f_{\text{proj}} \mathbf{W}_{\text{up}} + \mathbf{b}_{\text{up}} - f_{\text{text-token}}\|_2^2, \quad \mathbf{W}_{\text{up}} \in \mathbb{R}^{768 \times (77 \times 768)}. \quad (13)$$

557 The composite training objective was

$$\mathcal{L} = \mathcal{L}_{\text{proj1}} + \lambda_1 \mathcal{L}_{\text{CLIP}} + \lambda_2 \mathcal{L}_{\text{proj2}}. \quad (14)$$

558 The projector was trained for 600 epochs, including 50 warm-up epochs, using Adam with a learning rate of 2×10^{-4} and a
559 batch size of 128. The loss weights were set to $\lambda_1 = 0.5$ and $\lambda_2 = 1.0$.

560 During inference, a target dimension in the original SPoSE embedding was manually modified to produce a modulated
561 embedding, $\hat{f}_{\text{SPoSE}}$. The Action Projector transformed this embedding into target semantic tokens. In parallel, the source video
562 was encoded using a pre-trained VQVAE⁵⁸. DDIM inversion⁵⁹, with 50 inference steps, was then used to obtain an initial noise
563 tensor derived from the source video.

564 A pre-trained Stable Diffusion v1.5 model⁶⁰ was conditioned on the projected semantic tokens through cross-attention⁶¹.
565 The model denoised the initial tensor frame by frame in the latent space. Finally, the modulated latent frames were decoded by
566 the VQVAE to produce the output video.

567 Neural alignment framework

568 HAD-fMRI dataset acquisition and preprocessing

569 To link the computational representations of action to human neural substrates, we utilized the independent Human Action
570 Dataset (HAD)²⁰. This functional magnetic resonance imaging (fMRI) dataset encompasses neural responses from 30 healthy
571 participants viewing a total of 21,600 video clips, spanning 180 action categories. Whole-brain fMRI data were acquired on a
572 Siemens MAGNETOM Prisma 3 Tesla scanner utilizing a multi-band, gradient-echo accelerated echo-planar imaging (EPI)
573 sequence (2 mm isotropic spatial resolution, TR=2.0 seconds). The raw MRI data were preprocessed using the standardized
574 *fMRIPrep* pipeline and projected from the native anatomical space onto the standard fsLR cortical surface space.

575 Cross-dataset transfer of fixed SPoSE spaces (Dimension rating)

576 The neural analysis was designed as a zero-shot cross-dataset transfer test, rather than a re-estimation of semantic geometry on
577 the neuroimaging stimuli. The four 30-dimensional model-derived SPoSE spaces were learned from judgements on 355 S-MiT
578 videos and then held fixed. No triplet judgements on the 21,600 HAD clips, and no fMRI measurements, were used to update
579 these spaces. We instead learned a stimulus-level mapping that estimated coordinates in each fixed space from perceptual
580 features.

581 For each S-MiT video, we extracted frozen Perception Encoder features from the video as predictors. A multi-output
582 ridge regression with an intercept mapped these features to the 30 SPoSE coordinates. For each coordinate, the regularisation
583 parameter α was selected independently from 100 logarithmically spaced values between 10^{-3} and 10^3 by 5-fold cross-validation.
584 No further feature standardisation was applied, and no coordinate was excluded on the basis of prediction accuracy. After
585 hyperparameter selection, the mapping was refitted on all 355 S-MiT videos and applied strictly without adaptation to all 21,600
586 HAD clips.

587 Voxel-wise encoding analysis

588 To construct comprehensive voxel-wise encoding models, we employed a regularized ridge regression approach⁶² within a
589 nested cross-validation (CV) framework. The dataset was partitioned utilizing an outer 6-fold CV loop. Within each training
590 fold, an inner 5-fold CV was executed using RidgeCV⁶³ to optimize the L_2 regularization penalty (α). The hyperparameter
591 search space comprised 100 candidate values logarithmically spaced between 10^{-3} and 10^3 . Upon identifying the optimal α ,
592 the encoding model was retrained on the aggregate training set and deployed to predict the fMRI time courses of the held-out

test set. Model efficacy was assessed at the single-voxel level by calculating the Pearson correlation coefficient (r) between the predicted and empirically observed fMRI time series. Statistical significance was established via a bootstrapping procedure comprising 2,000 resamples, with P -values corrected for multiple comparisons across voxels using the False Discovery Rate (FDR) procedure⁶⁴.

Feature contribution analysis via fractional ridge regression

To quantitatively disentangle the unique variance explained by specific LLM- and MLLM-derived action dimensions in predicting voxel-wise fMRI activity, we employed Fractional Ridge Regression (FRR). For each voxel, the optimal regularization fraction hyperparameter was determined via an independent 7-fold CV procedure. The final model was fit to the complete dataset, yielding a weight matrix $\beta \in \mathbb{R}^{n_{\text{voxels}} \times n_{\text{dimensions}}}$. Each coefficient quantified the conditional association between a dimension and voxel activity within the regularized multivariate model.

Cortical parcellation and regional tuning profiles

Neuroanatomical localization and Region of Interest (ROI) definitions were based on the HCP-MMP1.0 parcellation⁶⁵, which delineates 180 distinct cortical areas per hemisphere. To characterize functional selectivity, FRR weights (β) were averaged across voxels within each ROI to obtain subject-specific tuning profiles, which were subsequently averaged across subjects. Both positive and negative weights were retained. We calculated the cosine similarity between each group-level regional tuning vector and the transferred SPoSE dimension scores for the 21,600 HAD videos, selecting those with the highest cosine similarities as representative exemplars of each ROI's functional preference.

Searchlight RSA

Complementary to the encoding models, we conducted a whole-brain searchlight RSA⁶⁶ in the native anatomical space of each participant to map the second-order correspondence between neural and model representational geometries. For every voxel, a local neural RSM was computed based on the multi-voxel activation patterns enclosed within a 12 mm diameter (6 mm radius) spherical neighborhood. We quantified representational alignment by computing the Pearson correlation (r) between the lower-triangle vector of this local neural RSM and the corresponding model-derived RSM. The correlation coefficients were assigned to the center voxel of each searchlight sphere, and the final statistical maps were projected onto flattened cortical surface meshes via the Pycortex software suite⁶⁷.

Data availability

The behavioural triplet odd-one-out judgements generated by the LLMs and MLLMs, the learned 30-dimensional SPoSE action embeddings, and the stimulus video annotations utilized in this study are publicly available at https://osf.io/5t23v/overview?view_only=3b0e6fe802604a969c4e31fa88401a30. The raw video stimuli from the Spoken Moments in Time (S-MiT) dataset can be requested from the original MIT repository (<http://moments.csail.mit.edu/spoken.html>). The human neuroimaging data (HAD-fMRI) are publicly available from the original dataset providers at <https://openneuro.org/datasets/ds004488/versions/2.0.1>. The human behavioural benchmark data can be accessed via the repository provided by Bockes et al. (<https://osf.io/b79eu/overview>).

Code availability

The custom scripts and pipelines used for data collection, embedding learning, result analysis, mechanistic attribution and figure plotting are publicly available at <https://github.com/Sunyi20/Human-LLM-Shared-Action>. The original PyTorch implementation of the SPoSE algorithm is available at <https://github.com/ViCCo-Group/SPoSE>.

References

1. Kabulska, Z. & Lingnau, A. The cognitive structure underlying the organization of observed actions. *Behav. Res. Methods* **55**, 1890–1906 (2023).
2. Tucciarelli, R., Wurm, M., Baccolo, E. & Lingnau, A. The representational space of observed actions. *elife* **8**, e47686 (2019).
3. Lingnau, A. & Downing, P. E. The lateral occipitotemporal cortex in action. *Trends cognitive sciences* **19**, 268–277 (2015).
4. Giese, M. A. & Poggio, T. Neural mechanisms for the recognition of biological movements. *Nat. Rev. Neurosci.* **4**, 179–192, [10.1038/nrn1057](https://doi.org/10.1038/nrn1057) (2003).
5. Orban, G., Lanzilotto, M. & Bonini, L. From observed action identity to social affordances. *Trends Cogn. Sci.* **25**, 493–505 (2021).

- 640 6. Tarhan, L. & Konkle, T. Sociality and interaction envelope organize visual action representations. *Nat. Commun.* **11**, 3002
641 (2020).
- 642 7. Pang, B. *et al.* Complex sequential understanding through the awareness of spatial and temporal concepts. *Nat. Mach.*
643 *Intell.* **2**, 245–253 (2020).
- 644 8. Mahner, F. P., Muttenthaler, L., Güçlü, U. & Hebart, M. N. Dimensions underlying the representational alignment of deep
645 neural networks with humans. *Nat. Mach. Intell.* **7**, 848–859 (2025).
- 646 9. Muttenthaler, L. *et al.* Aligning machine and human visual representations across abstraction levels. *Nature* **647**, 349–355
647 (2025).
- 648 10. Du, C. *et al.* Human-like object concept representations emerge naturally in multimodal large language models. *Nat. Mach.*
649 *Intell.* **7**, 860–875 (2025).
- 650 11. Binz, M. *et al.* A foundation model to predict and capture human cognition. *Nature* **644**, 1002–1009 (2025).
- 651 12. Schulze Buschoff, L. M., Akata, E., Bethge, M. & Schulz, E. Visual cognition in multimodal large language models. *Nat.*
652 *Mach. Intell.* **7**, 96–106 (2025).
- 653 13. Xu, Q. *et al.* Large language models without grounding recover non-sensorimotor but not sensorimotor features of human
654 concepts. *Nat. human behaviour* 1–16 (2025).
- 655 14. Binz, M. & Schulz, E. Using cognitive psychology to understand GPT-3. *Proc. Natl. Acad. Sci.* **120**, e2218523120 (2023).
- 656 15. Hebart, M. N., Zheng, C. Y., Pereira, F. & Baker, C. I. Revealing the multidimensional mental representations of natural
657 objects underlying human similarity judgements. *Nat. human behaviour* **4**, 1173–1185 (2020).
- 658 16. Bockes, A., Hebart, M. N. & Lingnau, A. Revealing key dimensions underlying the recognition of dynamic human actions.
659 *Commun. Psychol.* **3**, 149 (2025).
- 660 17. Dima, D. C., Janarthanan, S., Culham, J. C. & Mohsenzadeh, Y. Shared representations of human actions across vision and
661 language. *Neuropsychologia* **202**, 108962 (2024).
- 662 18. Garcia, K., McMahon, E., Conwell, C., Bonner, M. & Isik, L. Modeling dynamic social vision highlights gaps between
663 deep learning and humans. In *The Thirteenth International Conference on Learning Representations* (2024).
- 664 19. Cortinovis, D., Truong, N., Op de Beeck, H. & Bracci, S. Investigating action topography in visual cortex and deep artificial
665 neural networks. *Nat. Commun.* **17**, 1094, [10.1038/s41467-025-67855-6](https://doi.org/10.1038/s41467-025-67855-6) (2026).
- 666 20. Zhou, M. *et al.* A large-scale fMRI dataset for human action recognition. *Sci. Data* **10**, 415 (2023).
- 667 21. Hebart, M. N. *et al.* THINGS-data, a multimodal collection of large-scale datasets for investigating object representations
668 in human brain and behavior. *elife* **12**, e82580 (2023).
- 669 22. Josephs, E. L., Hebart, M. N. & Konkle, T. Dimensions underlying human understanding of the reachable world. *Cognition*
670 **234**, 105368 (2023).
- 671 23. Guo, D. *et al.* DeepSeek-R1 incentivizes reasoning in llms through reinforcement learning. *Nature* **645**, 633–638 (2025).
- 672 24. Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115* (2025). [2412.15115](https://arxiv.org/abs/2412.15115).
- 673 25. Bai, S. *et al.* Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923* (2025). [2502.13923](https://arxiv.org/abs/2502.13923).
- 674 26. Monfort, M. *et al.* Spoken moments: Learning joint audio-visual representations from video descriptions. In *Proceedings*
675 *of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14871–14881 (2021).
- 676 27. Selvaraju, R. R. *et al.* Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings*
677 *of the IEEE international conference on computer vision*, 618–626 (2017).
- 678 28. Zhao, C., Wang, K., Zeng, X., Zhao, R. & Chan, A. B. Gradient-based visual explanation for transformer-based clip. In
679 *International Conference on Machine Learning*, 61072–61091 (PMLR, 2024).
- 680 29. Hedger, N., Naselaris, T., Kay, K. & Knapen, T. Vicarious body maps bridge vision and touch in the human brain. *Nature*
681 **650**, 173–181 (2026).
- 682 30. Wurm, M. F., Caramazza, A. & Lingnau, A. Action categories in lateral occipitotemporal cortex are organized along
683 sociality and transitivity. *J. Neurosci.* **37**, 562–575 (2017).
- 684 31. Dima, D. C., Tomita, T. M., Honey, C. J. & Isik, L. Social-affective features drive human representations of observed
685 actions. *elife* **11**, e75027 (2022).

32. Yamins, D. L. K. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci.* **111**, 8619–8624, [10.1073/pnas.1403112111](https://doi.org/10.1073/pnas.1403112111) (2014).
33. Richards, B. A. *et al.* A deep learning framework for neuroscience. *Nat. Neurosci.* **22**, 1761–1770, [10.1038/s41593-019-0520-2](https://doi.org/10.1038/s41593-019-0520-2) (2019).
34. Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A. & Konkle, T. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nat. Commun.* **15**, 9383, [10.1038/s41467-024-53147-y](https://doi.org/10.1038/s41467-024-53147-y) (2024).
35. Gao, C. *et al.* Increasing alignment of large language models with language processing in the human brain. *Nat. computational science* **5**, 1080–1090 (2025).
36. Wurm, M. F. & Caramazza, A. Distinct roles of temporal and frontoparietal regions in representing actions across vision and language. *Nat. Commun.* **10**, 289, [10.1038/s41467-018-08084-y](https://doi.org/10.1038/s41467-018-08084-y) (2019).
37. Dima, D. C., Culham, J. C. & Mohsenzadeh, Y. Distinct perceptual and conceptual representations of natural actions along the lateral and dorsal visual streams. *Commun. Biol.* **9**, 577 (2026).
38. Gordon, E. M. *et al.* A somato-cognitive action network alternates with effector regions in motor cortex. *Nature* **617**, 351–359 (2023).
39. Huth, A. G., Nishimoto, S., Vu, A. T. & Gallant, J. L. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* **76**, 1210–1224 (2012).
40. Monfort, M. *et al.* Moments in time dataset: One million videos for event understanding. *IEEE Transactions on Pattern Analysis Mach. Intell.* **42**, 502–508, [10.1109/TPAMI.2019.2901464](https://doi.org/10.1109/TPAMI.2019.2901464) (2020).
41. OpenAI. gpt-oss-120b & gpt-oss-20b model card. OpenAI model card (2025).
42. Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* [10.48550/arXiv.2505.09388](https://arxiv.org/abs/2505.09388) (2025). [2505.09388](https://arxiv.org/abs/2505.09388).
43. DeepSeek-AI. DeepSeek-V3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556* [10.48550/arXiv.2512.02556](https://arxiv.org/abs/2512.02556) (2025). [2512.02556](https://arxiv.org/abs/2512.02556).
44. DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348* [10.48550/arXiv.2606.19348](https://arxiv.org/abs/2606.19348) (2026). [2606.19348](https://arxiv.org/abs/2606.19348).
45. MiniMax. The MiniMax-M2 series: Mini activations unleashing max real-world intelligence. *arXiv preprint arXiv:2605.26494* [10.48550/arXiv.2605.26494](https://arxiv.org/abs/2605.26494) (2026). [2605.26494](https://arxiv.org/abs/2605.26494).
46. Qwen Team. Qwen3.5: Towards native multimodal agents. Qwen technical blog (2026).
47. Tong, Z., Song, Y., Wang, J. & Wang, L. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Adv. neural information processing systems* **35**, 10078–10093 (2022).
48. Li, K. *et al.* Videochat: Chat-centric video understanding. *Sci. China Inf. Sci.* **68**, 200102 (2025).
49. Zhao, J. *et al.* Humanomni: A large vision-speech language model for human-centric video understanding. *arXiv preprint arXiv:2501.15111* (2025).
50. Yu, J. *et al.* Coca: Contrastive captioners are image-text foundation models. *Transactions on Mach. Learn. Res.* (2022).
51. Bolya, D. *et al.* Perception encoder: The best visual embeddings are not at the output of the network. In *Advances in Neural Information Processing Systems*, vol. 38, 60884–60937 (2025).
52. Tarhan, L., De Freitas, J. & Konkle, T. Behavioral and neural representations en route to intuitive action understanding. *Neuropsychologia* **163**, 108048 (2021).
53. Graziano, M. S. & Aflalo, T. N. Mapping behavioral repertoire onto the cortex. *Neuron* **56**, 239–251 (2007).
54. U.S. Bureau of Labor Statistics. American time use survey: 2019 results (2020).
55. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. statistical mechanics: theory experiment* **2008**, P10008 (2008).
56. Brin, S. & Page, L. The anatomy of a large-scale hypertextual web search engine. *Comput. networks ISDN systems* **30**, 107–117 (1998).
57. Radford, A. *et al.* Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763 (PmLR, 2021).
58. Van Den Oord, A., Vinyals, O. *et al.* Neural discrete representation learning. *Adv. neural information processing systems* **30** (2017).

59. Song, J., Meng, C. & Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations* (2021).
60. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695, [10.1109/CVPR52688.2022.01042](https://doi.org/10.1109/CVPR52688.2022.01042) (2022).
61. Hertz, A. *et al.* Prompt-to-prompt image editing with cross attention control. In *International Conference on Learning Representations* (2023).
62. Hoerl, A. E. & Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67, [10.1080/00401706.1970.10488634](https://doi.org/10.1080/00401706.1970.10488634) (1970).
63. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
64. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Royal Stat. Soc. Ser. B* **57**, 289–300, [10.1111/j.2517-6161.1995.tb02031.x](https://doi.org/10.1111/j.2517-6161.1995.tb02031.x) (1995).
65. Glasser, M. F. *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).
66. Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis—connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4, [10.3389/neuro.06.004.2008](https://doi.org/10.3389/neuro.06.004.2008) (2008).
67. Gao, J. S., Huth, A. G., Lescroart, M. D. & Gallant, J. L. Pycortex: an interactive surface visualizer for fMRI. *Front. neuroinformatics* **9**, 23 (2015).

Acknowledgements

This work was supported in part by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB1010202); in part by the National Natural Science Foundation of China under Grant 62576336; and in part by the Beijing Natural Science Foundation under Grant L243016. We would like to extend our gratitude to the creators of the Spoken Moments in Time (S-MiT) and Moments in Time (MiT) datasets for providing the foundational video stimuli. We thank the authors of the Human Action Dataset (HAD-fMRI) for making their neuroimaging data publicly accessible, and Bockes *et al.* for sharing the human behavioural similarity judgements. We also thank Martin N. Hebart and the ViCCo Group for sharing the original code for the SPoSE algorithm. Finally, we acknowledge the developer communities behind Qwen and DeepSeek for providing accessible, high-performance foundation models.

Author contributions

C.D., Y.S., Y.L., and H.H. conceptualized and designed the research; C.D., Y.S., and Y.L. conducted the experiments and implemented the computational pipelines (SPoSE, fMRI encoding, and generative editing); C.D., Y.S., Y.L., S.Q., K.Z., Y.W., and C.H. performed the data collection and data preprocessing; C.D., Y.S., Y.L., C.H., and H.H. drafted the manuscript; all authors actively participated in the analysis and interpretation of the results; all authors reviewed and approved the final manuscript.

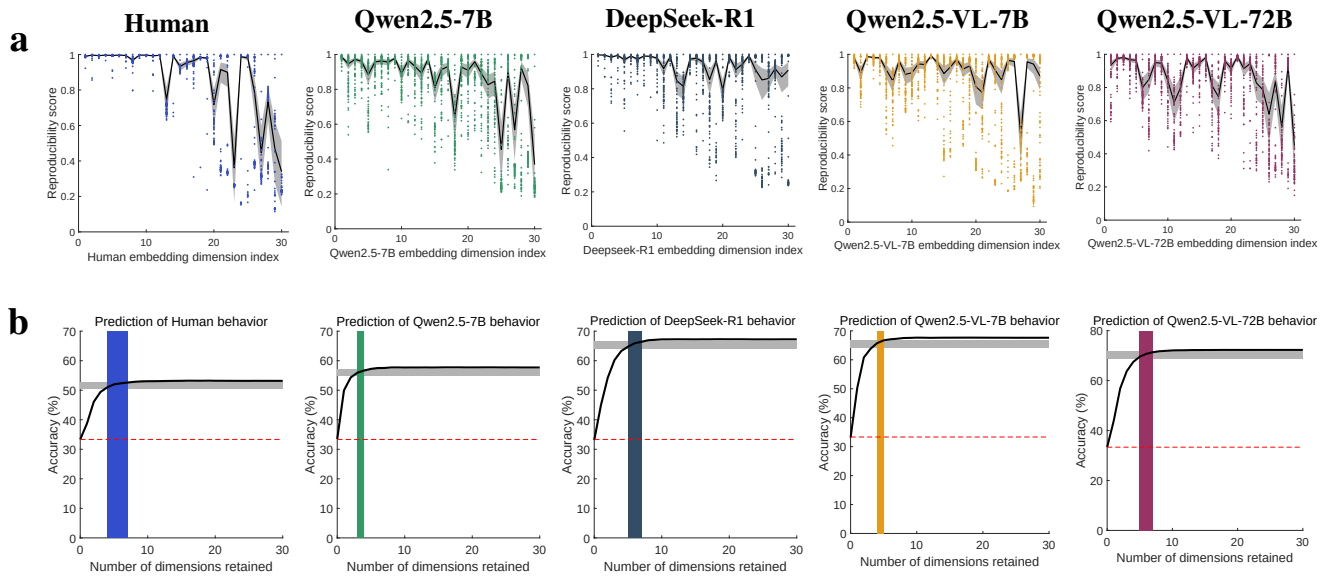
Competing interests

The authors declare no competing interests.

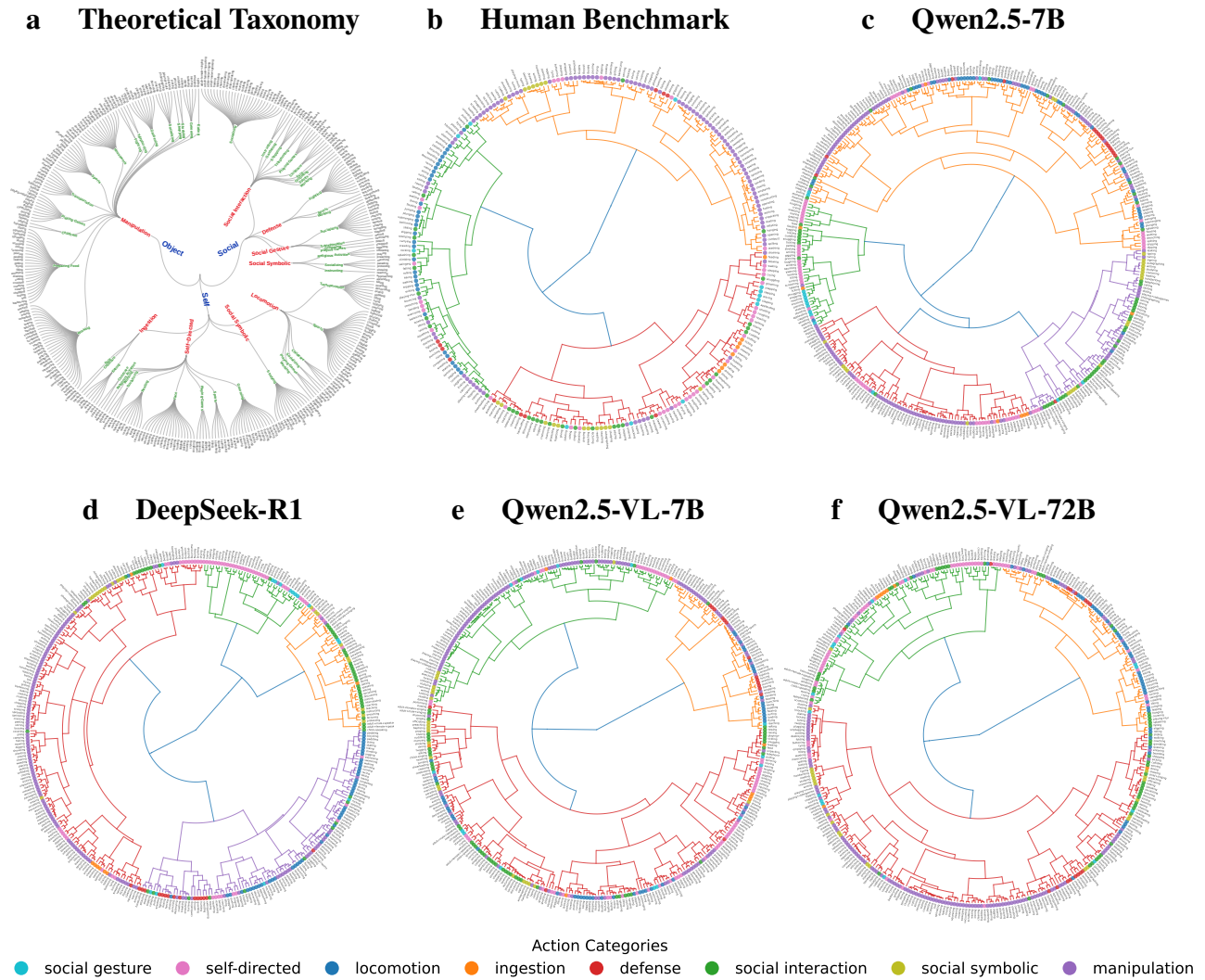
Additional information

Correspondence and requests for materials should be addressed to C.D. & H.H.

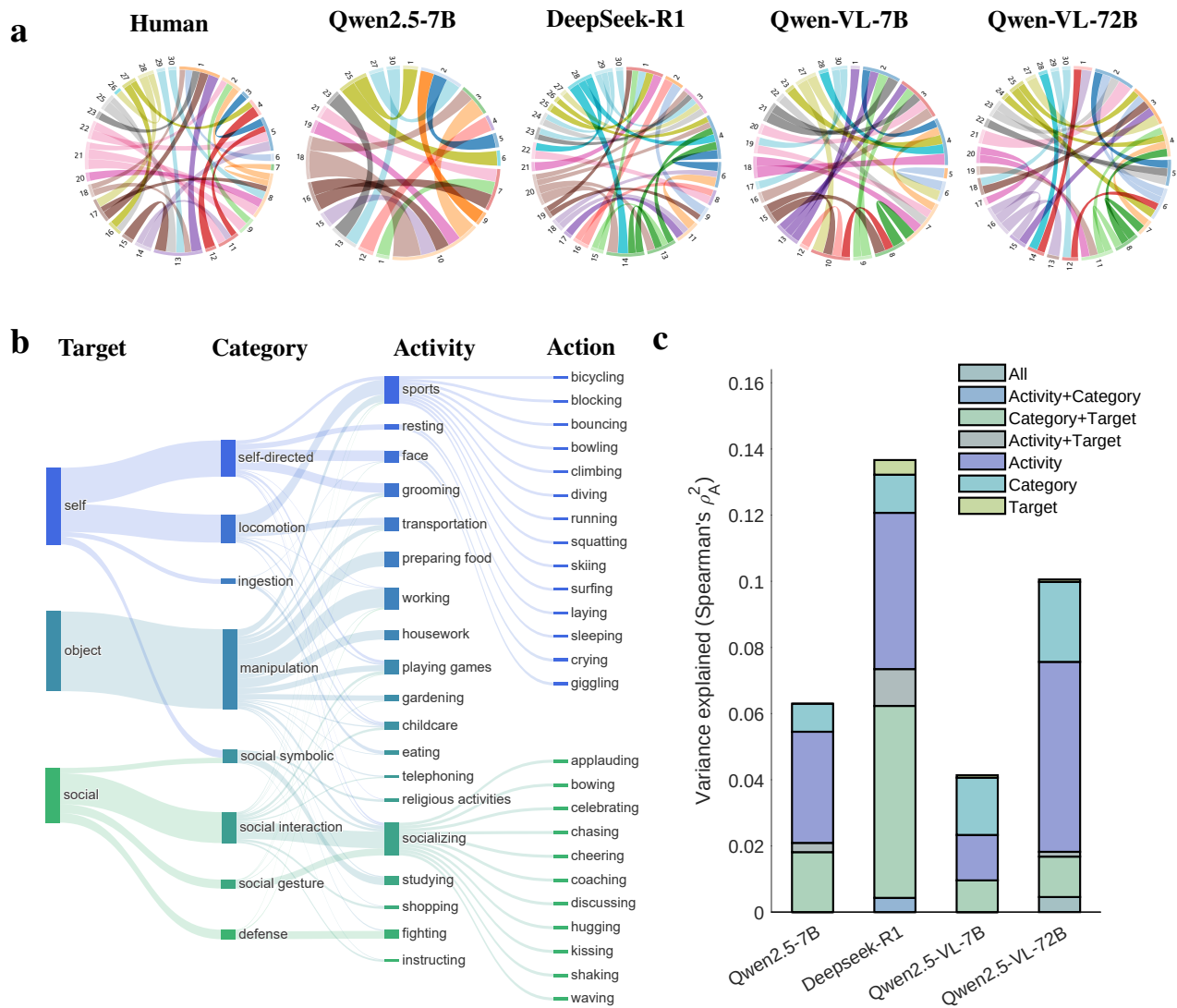
Extended data



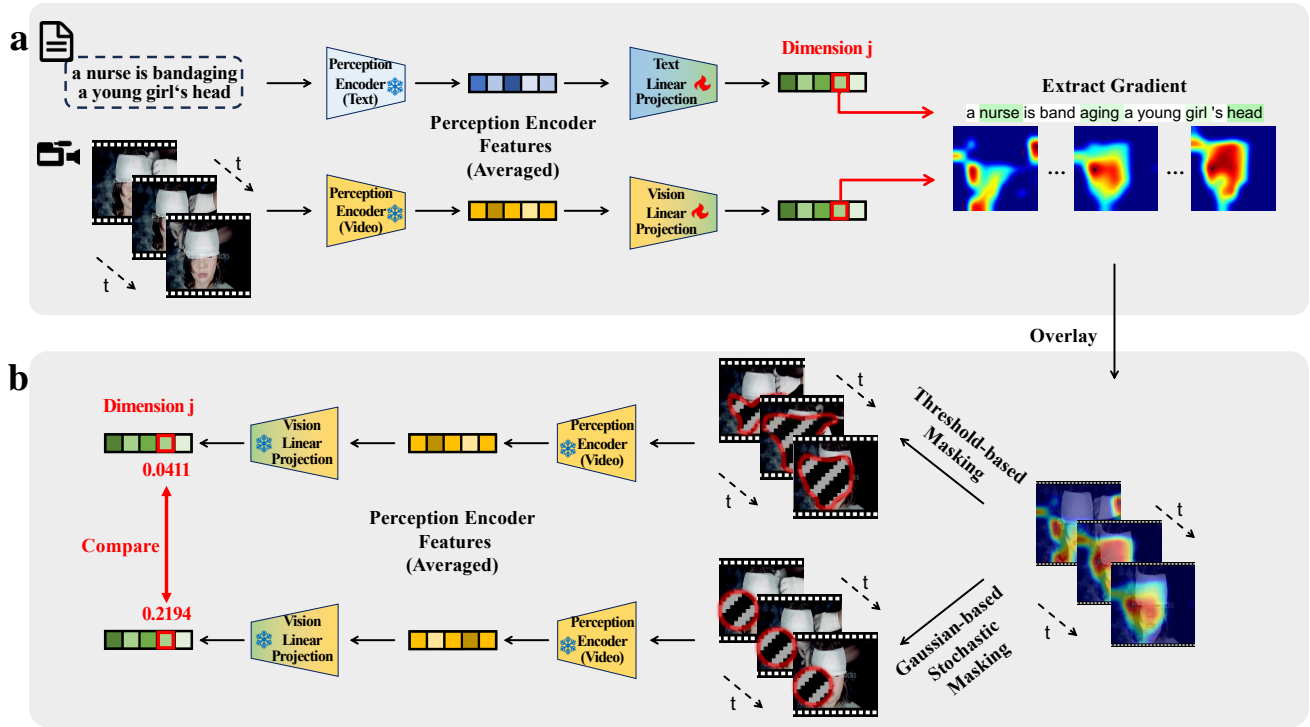
Extended Data Fig. 1. Stability validation and dimensionality profiles of the extracted action spaces. **a**, Reproducibility of individual SPoSE dimensions evaluated across 65 independent random initialization seeds. For each derived dimension, the best-matching counterpart in the remaining 64 runs was identified via Pearson correlation. Solid lines represent the mean reproducibility score (calculated via Fisher z -transformation), shaded regions denote 95% CIs, and scatter points represent individual run matches. This confirms that the extracted representations capture stable geometric structures rather than algorithmic noise. **b**, Dimensionality saturation curves evaluating predictive fidelity. Dimensions were progressively retained based on their ranked explanatory weights to predict held-out triplet choices. The red dashed line indicates chance-level performance (33.3%), and the gray shaded bands define the threshold of retaining 95% to 100% of peak predictive accuracy. Colored vertical bars mark the minimum number of dimensions required to cross this 95% threshold.



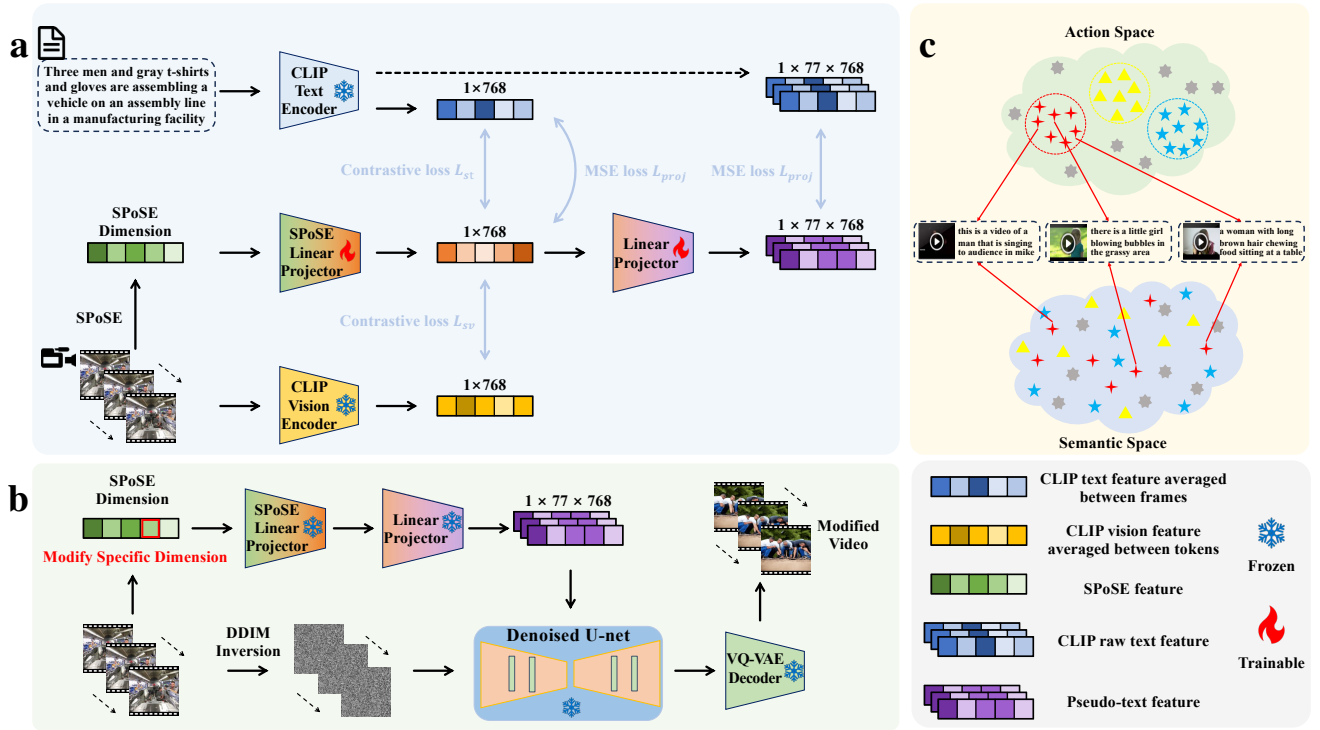
Extended Data Fig. 2. Hierarchical semantic organisation of action spaces captured by radial dendrograms. **a**, The theoretical taxonomic organisation of the 355 action categories utilized in this study, illustrating the predefined action taxonomy (inner rings, e.g., *Socializing*, *Sports*) to specific action leaves. **b–f**, Radial dendrograms illustrating the hierarchical clustering structures derived from the latent SPoSE embeddings of the human benchmark (**b**), LLMs (**c**, **d**), and MLLMs (**e**, **f**). The trees were constructed using Ward’s agglomerative clustering method applied to the cosine dissimilarity matrix ($1 - \text{similarity}$) of the action embeddings. Leaf nodes are color-coded by their annotated semantic superclasses. The coherent branching patterns across all artificial models demonstrate that the AI systems show category-related branching patterns and semantic proximity of human actions without explicit hierarchical supervision. Notably, the clustering branches in DeepSeek-R1 and Qwen-VL-72B exhibit a qualitative similarity with the theoretical taxonomy and the human benchmark.



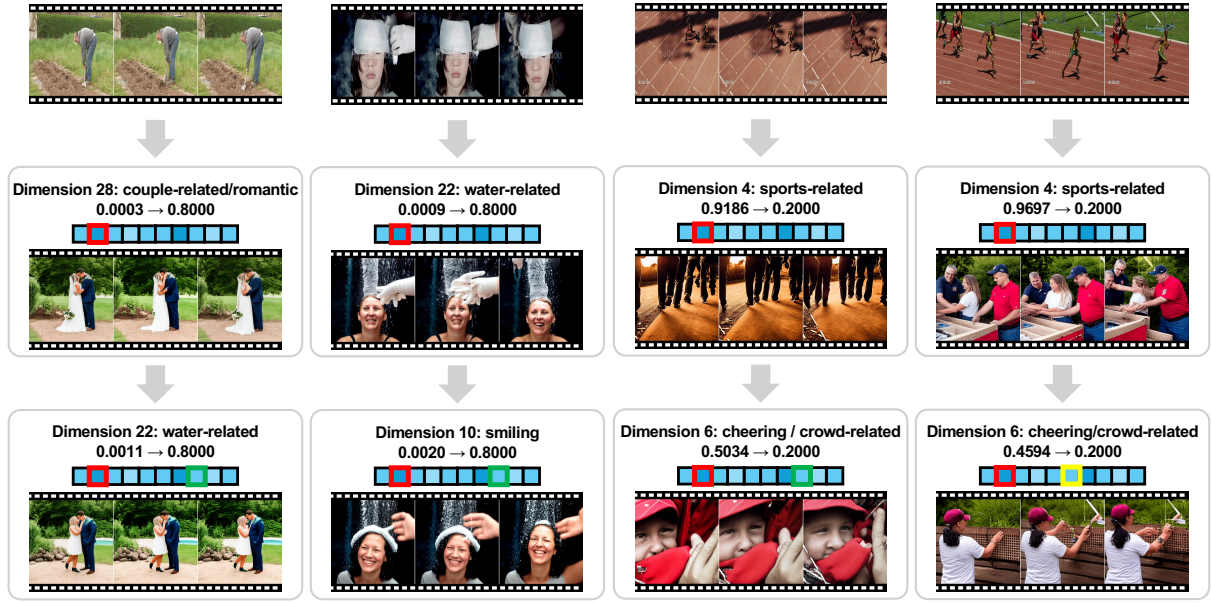
Extended Data Fig. 3. Complementary topological visualizations, hierarchical taxonomy, and variance partitioning of semantic drivers. **a**, Chord diagrams providing a dense, circular visualization of the representational network geometries. The nodes represent individual action dimensions, and the chord thicknesses correspond to the strength of inter-dimensional correlations. This complementary perspective highlights the dense, hub-like interconnectivity present in larger models (e.g., DeepSeek-R1 and Qwen-VL-72B), which structurally approximate the integrated complexity of the human cognitive graph, distinguishing them from the sparser connectivity of smaller models. **b**, Hierarchical categorization of the 355 naturalistic human actions. The alluvial diagram delineates the multi-level taxonomy flowing from three broad action targets (*self*, *object*, *social*) through intermediate functional categories (e.g., *manipulation*, *social interaction*) down to fine-grained daily activities (e.g., *sports*, *grooming*) and individual action exemplars. Flow width reflects the proportional density of actions assigned to each respective branch. **c**, Variance partitioning analysis disentangling the unique and shared contributions of the aforementioned semantic features (*target*, *category*, and *activity*) to the models' similarity judgements. The total explained variance (evaluated via out-of-sample Spearman's ρ_A^2) is decomposed using the inclusion-exclusion principle. The stacked bars demonstrate that the advanced behavioural consistency of LLMs and MLLMs is disproportionately driven by fine-grained *category* and *activity* semantics, paired with a substantial degree of shared interaction variance. This indicates that AI systems resolve action perception by heavily prioritizing high-level contextual routines and functional goals, providing an explanation for why foundation models substantially outperform traditional pixel-matching vision encoders.



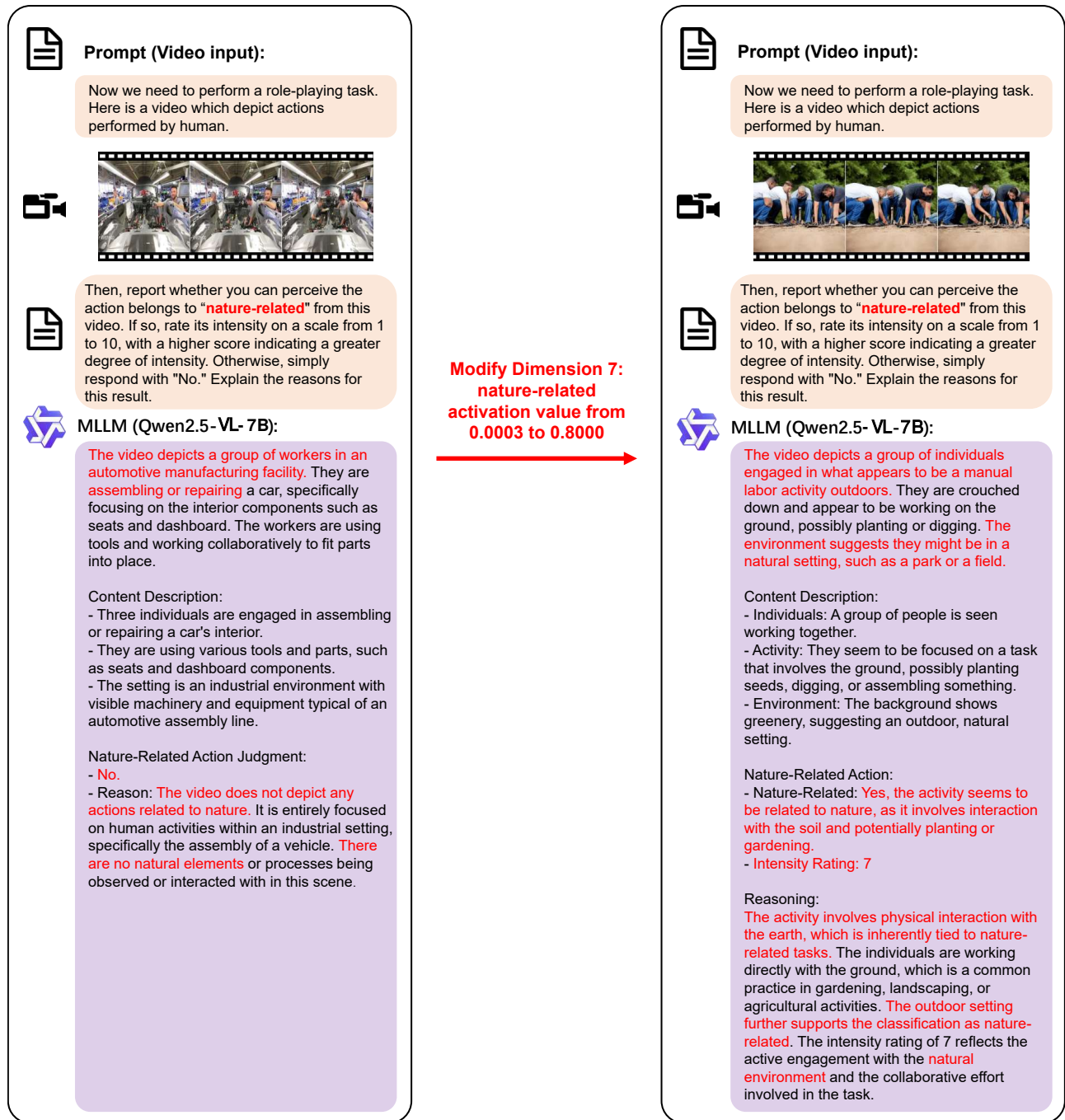
Extended Data Fig. 4. Schematic of the feature attribution and spatial perturbation framework. a, Feature Attribution: To identify the input features driving specific action dimensions (Dimension j), we applied gradient-based attribution. Text and video inputs are processed by frozen Perception Encoders and mapped to the SPoSE action space via trainable linear projections. By computing the gradient of the target dimension j with respect to the inputs, modality-specific attributions are generated. For a *bandaging* action, Grad-ECLIP highlights relevant textual tokens (e.g., “nurse”, “head”), while Grad-CAM produces spatial saliency heatmaps localizing the active visual regions (e.g., the patient’s head and the nurse’s hands). **b, Masking Assessment:** To validate the relevance of the visual regions identified by Grad-CAM, a frame-wise spatial masking protocol was employed. High-activation regions were occluded using threshold-based targeted masking (top path), while Gaussian-based stochastic masking served as a spatial control (bottom path). Re-evaluating the masked videos through the frozen encoding pipeline demonstrates that targeted occlusion yields a substantially larger reduction in the target dimension’s predicted activation (dropping to 0.0411) compared to the stochastic control (0.2194). This differential reduction confirms the auxiliary projection’s sensitivity to the specific salient regions localized by Grad-CAM.

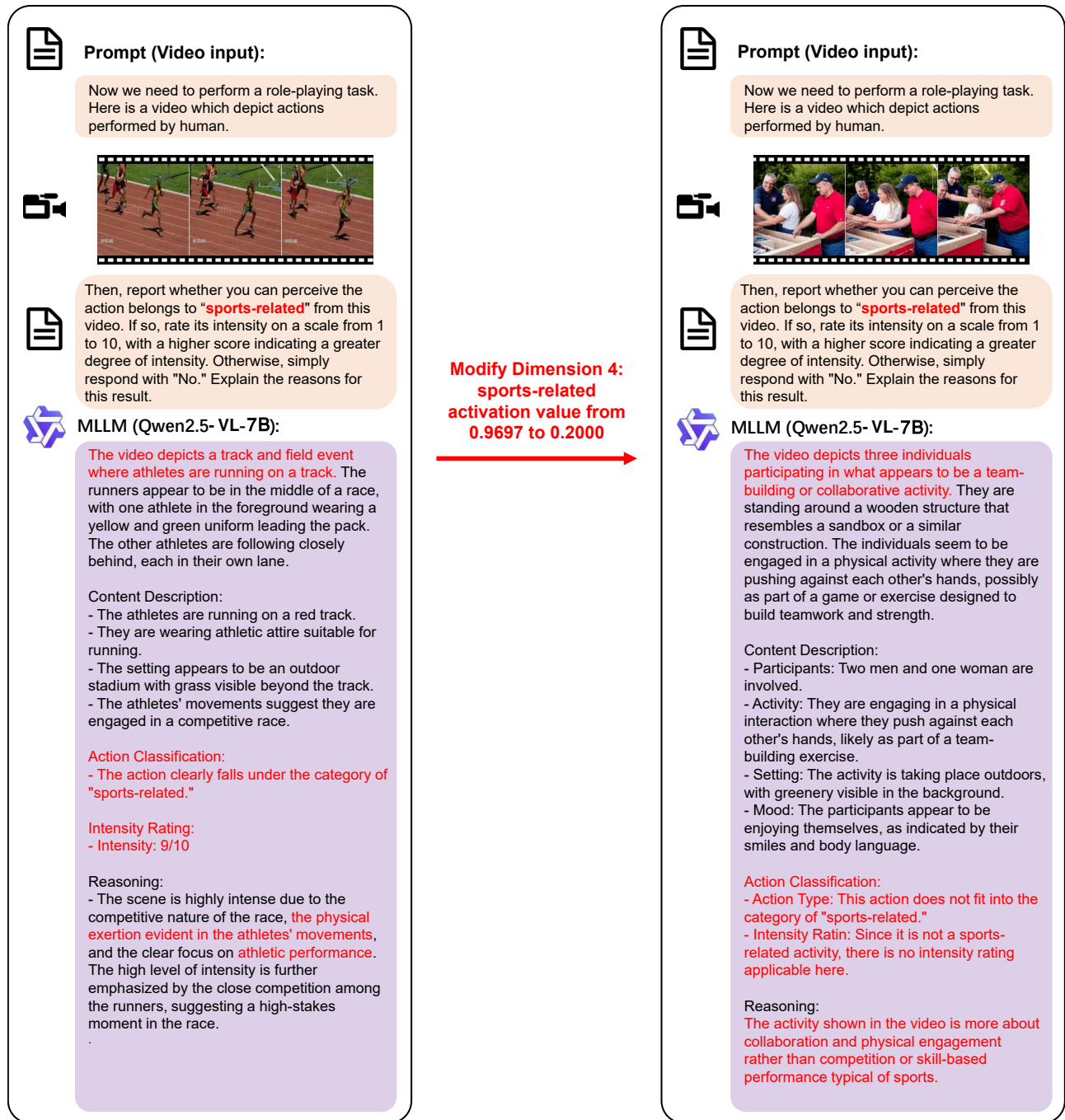


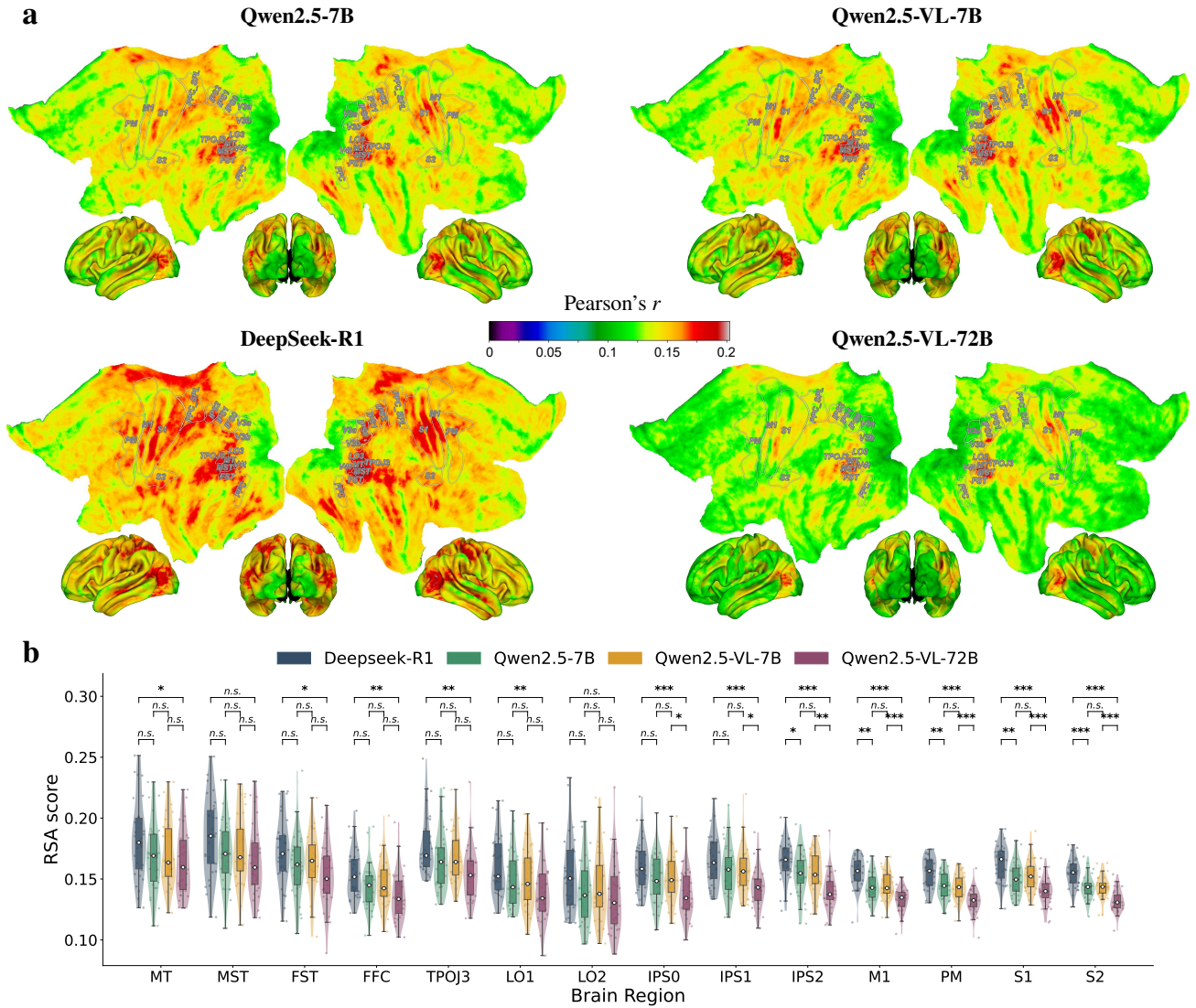
Extended Data Fig. 5. Architecture of the cross-space Action Projector for dimension-guided video synthesis and modulation. **a, Training pipeline.** A multi-objective alignment strategy maps the 30-dimensional SPoSE action embedding into the multimodal latent space of a pre-trained model (i.e., CLIP). The trainable linear projection modules (indicated by red fire icons) are optimized using contrastive losses (L_{st} , L_{sv}) to align with features extracted from frozen CLIP textual and visual encoders (indicated by blue snowflake icons). Auxiliary mean squared error (MSE) losses (L_{proj}) enforce fine-grained alignment with token-level text features ($1 \times 77 \times 768$). **b, Inference pipeline.** The source video undergoes DDIM inversion to extract a latent noise tensor, preserving the original temporal and structural priors. Concurrently, targeted modifications are applied to specific SPoSE dimension values. The trained (now frozen) Action Projector transforms this modulated embedding into a $1 \times 77 \times 768$ pseudo-text conditioning tensor. Conditioned on this tensor, a frozen Denoising U-Net synthesizes modified latent frames, which are subsequently decoded into pixel space by a VQ-VAE decoder. **c, Conceptual schematic of cross-space mapping.** By aligning the behaviorally derived Action Space with the generative Semantic Space, distinct behavioral clusters (e.g., *singing* versus *chewing*) are mapped to navigable generative coordinates. This linkage enables the direct translation of low-dimensional behavioral axes into conditioning signals for semantic video modulation.



Extended Data Fig. 6. Sequential and bidirectional generative modulation of action dimensions. This figure demonstrates multi-step conditional video generation across four distinct action sequences (columns). **Top row:** Original source videos. **Middle row (Step 1):** Visual outputs following the initial dimensional manipulation. This step illustrates both up-regulation (e.g., increasing *couple-related* or *water-related* from near zero to 0.8000, left two columns) and down-regulation (e.g., decreasing *sports-related* from > 0.9 to 0.2000, right two columns) of specific SPoSE dimensions. Down-regulating the *sports-related* dimension successfully suppresses athletic track cues. **Bottom row (Step 2):** Visual outputs following a subsequent, additive dimension edit. This illustrates sequential composability, such as adding *water-related* features to the newly generated romantic couple (column 1) or adding the *smiling* attribute to the water scene (column 2). These interventions confirm that the extracted dimensions function as continuous, bidirectional, and composable conditioning variables for semantic synthesis and modulation.



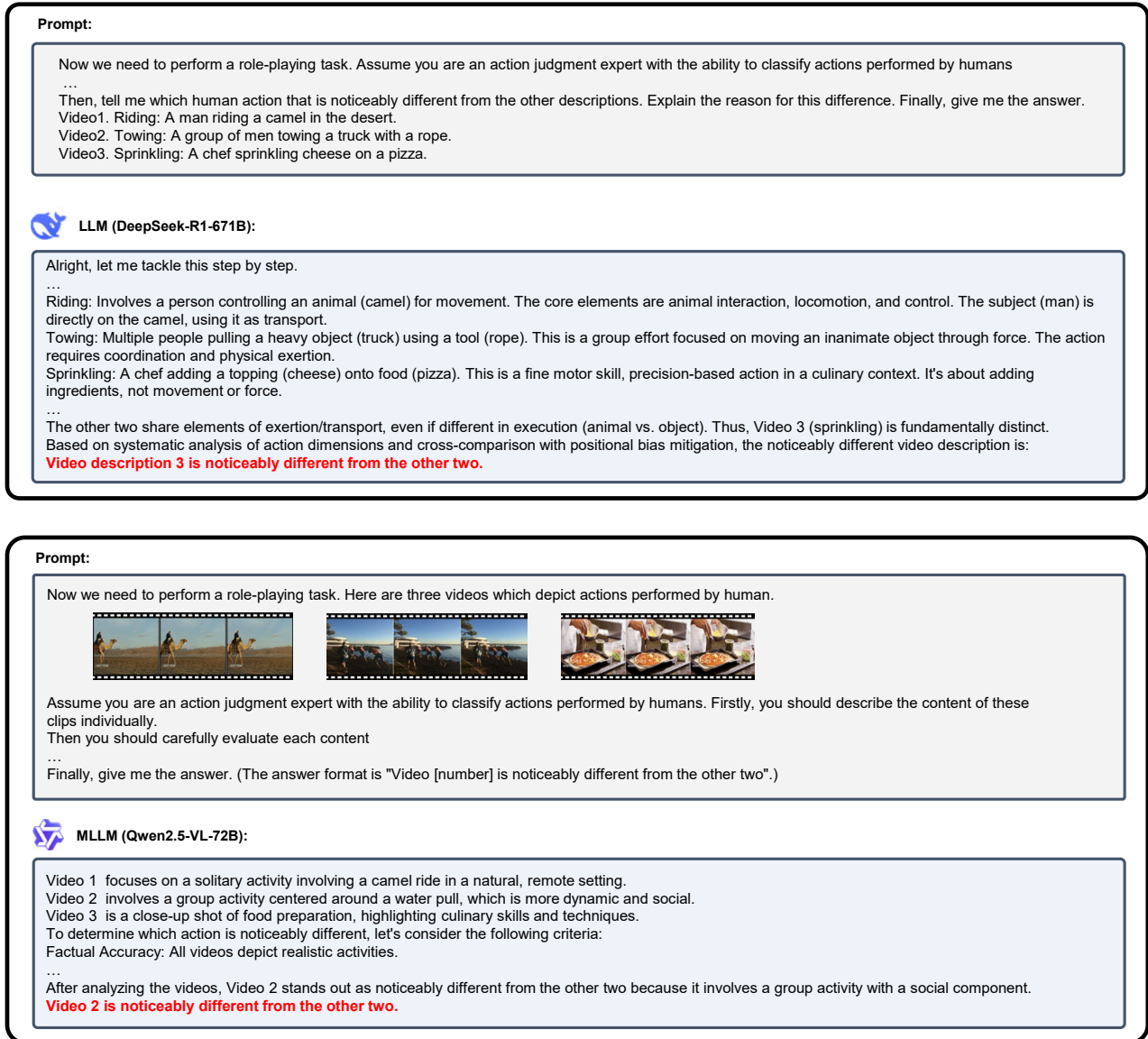




Extended Data Fig. 9. Searchlight representational similarity analysis (RSA) of cortical action geometries. This analysis provides second-order geometric validation, complementing the voxel-wise encoding models. **a**, Flattened cortical maps displaying group-level searchlight RSA scores (Pearson's r , averaged across $n = 30$ subjects). These maps quantify the direct geometric alignment between the model-derived semantic spaces and local multi-voxel activity patterns. The text-only DeepSeek-R1 model demonstrates visually broader and stronger representational correspondence, particularly expanding across the parietal and sensorimotor networks. **b**, ROI-averaged searchlight RSA performance. Violin plots show the full distribution of individual subject values, overlaid with boxplots (median, IQR, and $1.5 \times$ IQR whiskers). White-filled circles indicate medians. Brackets denote pairwise statistical comparisons (two-tailed Welch's t-test, Benjamini-Hochberg FDR corrected; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; n.s., not significant). In this test of representational geometry, the text-derived DeepSeek-R1 embedding significantly outperforms the multimodal Qwen2.5-VL-72B embedding across the intraparietal sulcus (IPS0–IPS2) and the entire evaluated sensorimotor network (M1, PM, S1, S2).



Extended Data Fig. 10. Model-specific regional neurosemantic tuning profiles across diverse computational architectures. Complementing the main text analysis, this figure presents the regional tuning profiles derived from the text-only DeepSeek-R1 (a), the text-only Qwen2.5-7B (b), and the multimodal Qwen2.5-VL-72B (c). Donut plots illustrate the top three action dimensions with the largest positive fractional ridge regression weights within five targeted ROIs (MT, MST, FST, FFC, and TPOJ3). Below each plot, the four representative HAD videos exhibiting the highest model-predicted regional similarity are displayed. Notably, despite differences in the specific semantic labels extracted by each individual foundation model, the mapped regional tuning profiles exhibit functional consistency. Across architectures, the FST region consistently tunes to fine-motor manipulation (e.g., *hand-related*, *grooming / nursing*, *writing / drawing*), while MT and TPOJ3 encode multi-agent dynamics and broader contextual activities (e.g., *interaction / crowd-related*, *entertaining*, *sports-related*). This cross-model consistency further confirms that the extracted semantic spaces capture anatomically grounded features of human action perception.



Supplementary Fig. 1. Representative prompt structures and modality-dependent reasoning in the odd-one-out paradigm. This figure illustrates the experimental setup and the reasoning elicited from the foundation models for an identical action triplet (*Riding*, *Towing*, *Sprinkling*). **Top:** The text-only LLM (DeepSeek-R1-671B) evaluates textual descriptions. Its reasoning is driven by abstract functional and kinematic semantics; it groups *Riding* and *Towing* based on shared elements of “exertion/transport,” selecting Video 3 (*Sprinkling*) as the outlier due to its distinct reliance on “fine motor skill” and “precision.” **Bottom:** The visually grounded MLLM (Qwen2.5-VL-72B) directly processes sequences of video frames. Its reasoning is heavily influenced by visual scene composition and social cues; it groups the solitary and close-up actions, selecting Video 2 (*Towing*) as the outlier because it distinctly involves a “group activity with a social component.” These qualitative examples demonstrate that while both architectures utilize coherent, step-by-step logical deduction, their distinct input modalities can naturally lead to divergent, yet internally valid, semantic categorizations of the same action concepts.

Supplementary Table 1. List of actions selected for this work. This table details all 355 action classes included in the experiment. Actions that overlap with the human behavioural dataset are denoted by a green checkmark (✓).

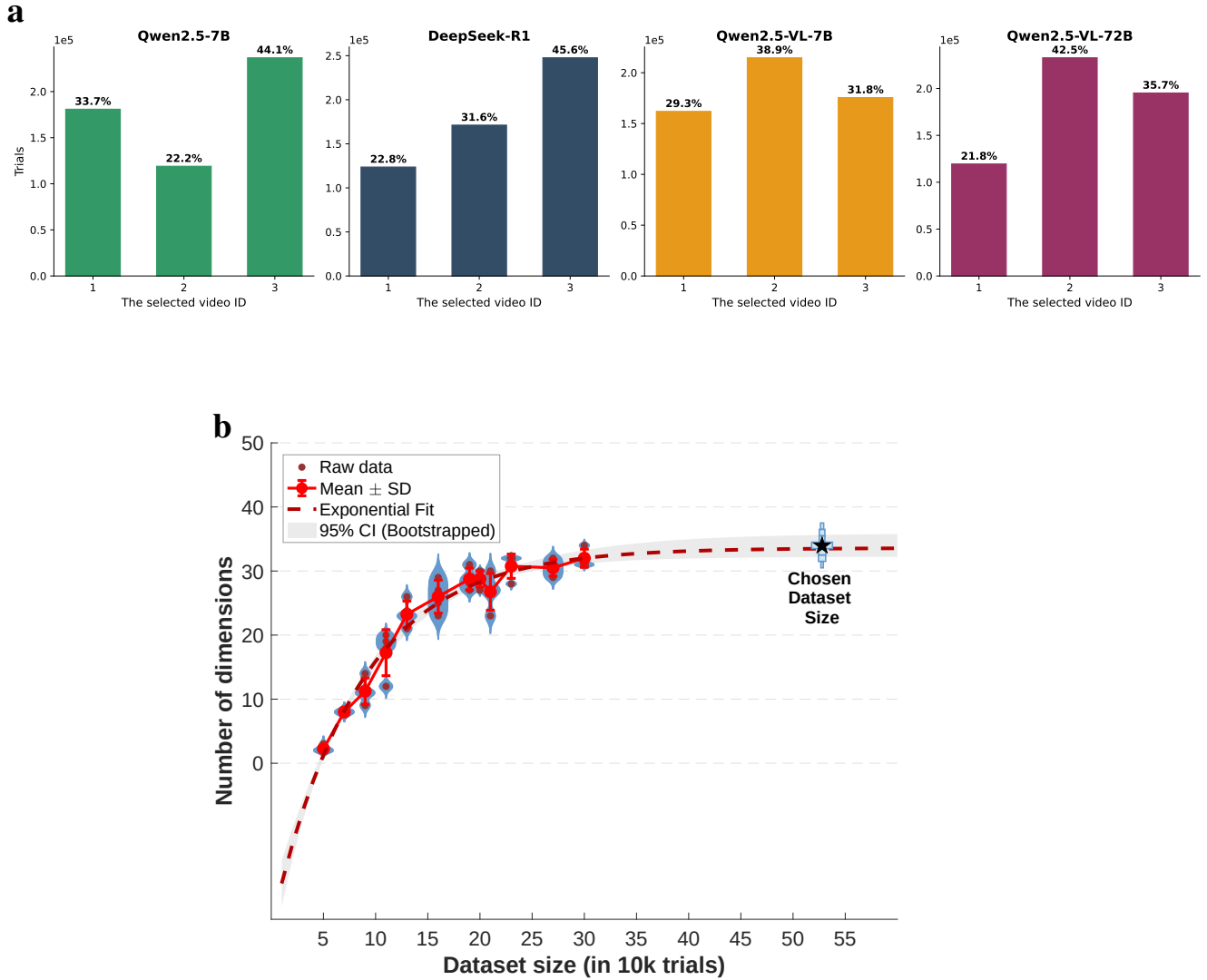
No. Action	No. Action	No. Action	No. Action	No. Action	No. Action
1 adult female singing	61 clinging	121 flicking ✓	181 massaging ✓	241 running ✓	301 steering ✓
2 adult female speaking	62 clipping ✓	122 flipping ✓	182 measuring ✓	242 sailing ✓	302 stirring ✓
3 adult male singing	63 closing ✓	123 floating ✓	183 mopping ✓	243 saluting ✓	303 stitching ✓
4 adult male speaking	64 coaching ✓	124 flossing	184 mowing	244 sanding ✓	304 stomping ✓
5 aiming ✓	65 colliding	125 flowing	185 officiating	245 sawing ✓	305 stopping
6 applauding ✓	66 combing ✓	126 flying ✓	186 opening ✓	246 scooping	306 stretching ✓
7 approaching	67 combusting ✓	127 folding ✓	187 operating ✓	247 scratching ✓	307 stroking
8 arresting ✓	68 competing ✓	128 frowning ✓	188 packaging ✓	248 screwing ✓	308 studying ✓
9 ascending ✓	69 constructing ✓	129 frying ✓	189 packing ✓	249 scrubbing	309 submerging ✓
10 asking ✓	70 cooking ✓	130 fueling ✓	190 paddling	250 selling	310 surfing ✓
11 assembling ✓	71 coughing ✓	131 gambling	191 painting ✓	251 serving ✓	311 sweeping ✓
12 attacking	72 covering	132 gardening ✓	192 parading ✓	252 sewing ✓	312 swerving
13 autographing ✓	73 cracking ✓	133 gasping	193 paying ✓	253 shaking ✓	313 swimming ✓
14 baking ✓	74 crafting ✓	134 giggling	194 pedaling ✓	254 shaving	314 swinging
15 balancing ✓	75 cramming ✓	135 giving ✓	195 peeling ✓	255 shivering	315 talking
16 bandaging	76 crashing	136 grilling ✓	196 performing ✓	256 shooting ✓	316 taping ✓
17 baptizing ✓	77 crawling ✓	137 grinning	197 petting	257 shopping ✓	317 tapping ✓
18 barbecuing ✓	78 crouching	138 gripping ✓	198 photographing ✓	258 shouting ✓	318 tattooing ✓
19 bathing ✓	79 crushing ✓	139 grooming ✓	199 picking ✓	259 shoveling ✓	319 teaching ✓
20 bending ✓	80 crying ✓	140 guarding ✓	200 piling	260 showing	320 tearing
21 bicycling ✓	81 cuddling ✓	141 hammering ✓	201 piloting ✓	261 shredding ✓	321 telephoning ✓
22 biking	82 cutting ✓	142 handcuffing ✓	202 pitching	262 shrugging ✓	322 texting
23 biting ✓	83 dancing ✓	143 handwriting ✓	203 placing	263 signing ✓	323 throwing
24 blocking	84 descending ✓	144 hanging ✓	204 planting ✓	264 singing ✓	324 tickling ✓
25 blowing ✓	85 destroying	145 hiking ✓	205 playing ✓	265 sitting ✓	325 tilting
26 boarding ✓	86 dialing	146 hitchhiking ✓	206 playing fun ✓	266 skating ✓	326 towing
27 boating ✓	87 digging ✓	147 hitting ✓	207 playing music ✓	267 sketching ✓	327 trimming ✓
28 boil	88 dining ✓	148 howling ✓	208 playing sports ✓	268 skiing ✓	328 tripping ✓
29 boiling	89 dipping ✓	149 hugging ✓	209 playing videogames	269 skipping	329 tucking
30 bouncing ✓	90 discussing ✓	150 hunting ✓	210 plugging ✓	270 slapping ✓	330 tuning ✓
31 bowing ✓	91 diving ✓	151 imitating	211 plunging ✓	271 sleeping ✓	331 turning
32 bowling	92 dragging	152 inflating ✓	212 pointing ✓	272 slicing ✓	332 twisting ✓
33 boxing ✓	93 draining ✓	153 injecting ✓	213 poking	273 sliding ✓	333 tying ✓
34 breaking ✓	94 drawing ✓	154 instructing	214 pouring ✓	274 slipping ✓	334 typing ✓
35 brushing ✓	95 drenching	155 interviewing ✓	215 praying ✓	275 smashing ✓	335 unfolding
36 bubbling ✓	96 dressing ✓	156 ironing	216 preaching ✓	276 smelling ✓	336 unloading ✓
37 building ✓	97 drilling ✓	157 jogging ✓	217 pressing	277 smiling ✓	337 unpacking ✓
38 bulldozing	98 drinking ✓	158 joining	218 protesting ✓	278 smoking ✓	338 unplugging
39 burning	99 dripping	159 juggling ✓	219 pulling ✓	279 snapping ✓	339 vacuuming ✓
40 burying ✓	100 driving ✓	160 jumping ✓	220 punching ✓	280 sneezing	340 waking ✓
41 buttoning ✓	101 dropping	161 kicking ✓	221 punting ✓	281 sniffing ✓	341 walking ✓
42 buying ✓	102 drumming	162 kissing ✓	222 pushing ✓	282 snuggling ✓	342 washing ✓
43 calling ✓	103 drying ✓	163 kneeling	223 putting	283 socializing	343 watering ✓
44 camping ✓	104 dunking	164 knitting	224 queuing	284 sowing ✓	344 waving ✓
45 carrying ✓	105 dusting ✓	165 knocking ✓	225 racing ✓	285 speaking ✓	345 waxing
46 carving ✓	106 eating ✓	166 landing	226 rafting ✓	286 spilling ✓	346 weeding ✓
47 catching ✓	107 emptying ✓	167 laughing ✓	227 raising ✓	287 spinning ✓	347 welding
48 celebrating ✓	108 entering ✓	168 launching	228 reaching	288 spitting	348 wetting ✓
49 chasing	109 erasing	169 laying	229 reading ✓	289 splashing ✓	349 whistling ✓
50 cheering ✓	110 exercising ✓	170 leaking	230 removing	290 spraying ✓	350 winking
51 cheerleading ✓	111 exiting ✓	171 leaning	231 repairing ✓	291 spreading ✓	351 working ✓
52 chewing ✓	112 extinguishing	172 leaping	232 resting	292 sprinkling ✓	352 wrapping ✓
53 child singing	113 falling ✓	173 lecturing ✓	233 riding ✓	293 sprinting ✓	353 wrestling ✓
54 child speaking	114 feed	174 licking ✓	234 rinsing ✓	294 squatting ✓	354 writing ✓
55 chopping ✓	115 feeding ✓	175 lifting ✓	235 rising ✓	295 squeezing	355 yawning ✓
56 clapping ✓	116 fencing	176 loading ✓	236 roaring	296 squinting ✓	
57 clawing ✓	117 fighting ✓	177 locking	237 rocking ✓	297 stacking ✓	
58 cleaning ✓	118 filling ✓	178 manicuring	238 rolling ✓	298 standing ✓	
59 clearing	119 filming ✓	179 marching ✓	239 rowing ✓	299 starting	
60 climbing ✓	120 fishing ✓	180 marrying ✓	240 rubbing	300 stealing	

Supplementary Table 2. Action dimensions derived from LLMs, MLLMs, and human behavioural data. Each dimension is labelled according to the top nine actions with the highest activation values, determined through human annotation and model-generated suggestions. Models include Qwen2.5-7B, DeepSeek-R1 (LLMs), and Qwen2.5-VL-7B, Qwen2.5-VL-72B (MLLMs).

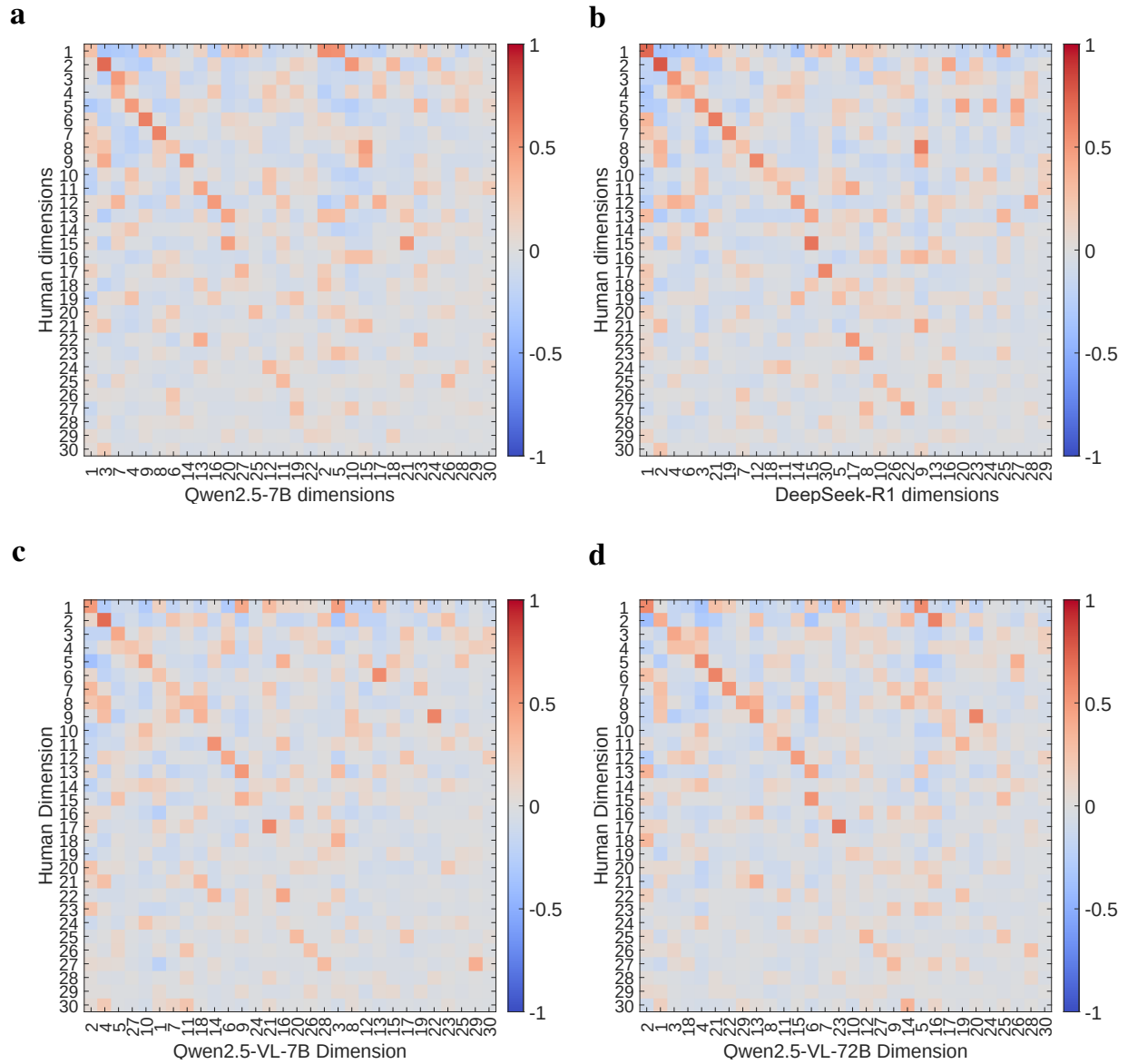
Dim.	Qwen2.5-7B	DeepSeek-R1	Qwen2.5-VL-7B	Qwen2.5-VL-72B	Human
1	housework-related	housework-related	eating	routine-related	craft
2	hand-related	sports-related	labor-related	hand-related	sport
3	entertaining-related	portrait-related	hand-related	business-related	speaking
4	portrait-related	speaking / talking	sports-related	smiling	interaction
5	repairing	hand-related	business-related	labor-related	mouth
6	moving / walking	interaction / crowd-related	cheering / crowd-related	handicraft-related	food
7	speaking / talking	moving / walking	nature-related	nature-related	nature/outdoor
8	planting / nature-related	danger-related	playing	resting / relaxing	vehicle
9	cooking	driving	handicraft-related	violence-related	water
10	sports-related	communication / finger-related	smiling	full-body action	child
11	art / performance-related	baby-related	moving	massaging	body
12	intimate-related	water-related	resting / relaxing	performance-related	crowd
13	grooming / nursing	art / performance-related	cooking	extreme sports-related	hand
14	water-related	cheering	massaging	exercise-related	female
15	driving	writing / drawing	child-related / playing	cleaning	drawing
16	cheering	ball sports-related	cleaning	sports-related	aiming
17	religion-related	grooming / nursing	performance-related	child-related / playing	cleaning
18	ball sports-related	child-related / playing	extreme sports-related	interaction-related	knotting
19	violence-related	planting / nature-related	planting	grooming / nursing	gestures
20	handicraft-related	mouth-related / blowing	electrical device-related	water-related	force
21	writing / drawing	cooking	sweeping	cooking	winter sports
22	child-related / playing	violence-related	exercise-related	planting	body care
23	self-caring	relaxing	water-related	cleaning	labor / constructing
24	resting	facial / negative-related	learning	baby-related	animals
25	interaction / crowd-related	repairing	sound-related	learning / working	music
26	communication / finger-related	purchasing / trade-related	purchasing / trade-related	eating / drinking	entering/leaving
27	cleaning	eating / drinking	couple-related / romantic	purchasing / trade-related	violence
28	facial / negative-related	religion-related	violence-related	movement-related	foot/stomping
29	ceremony-related	resting	interaction-related	driving	handicraft
30	baby-related	cleaning	religion-related	religion-related	sport/locomotion

Supplementary Table 3. Quantitative comparison of activation reduction under targeted and random spatial masking strategies. This table details the changes in the representation values for specific action dimensions when subjected to saliency-guided targeted masking compared to random spatial masking. The original activation values establish the baseline for the Qwen and DeepSeek architectures. The columns for targeted and random masking display the absolute updated activation values, accompanied by the relative percentage decrease from the baseline enclosed in parentheses. A higher percentage indicates a more substantial reduction in the activation of the dimension, with larger reductions indicating greater sensitivity of the predicted score to the masked region. Negative reduction percentages indicate an increase in the activation value relative to the baseline.

Model	Video ID	Dim.	Original Activation	Targeted Masking Activation (↓%)	Random Masking Activation (↓%)
Qwen2.5-VL-7B	2	17	1.000	0.344 (65.57%)	0.423 (57.69%)
	15	14	0.451	0.041 (90.89%)	0.158 (64.96%)
	196	14	0.309	0.138 (55.17%)	0.164 (46.75%)
	229	3	0.273	0.218 (20.26%)	0.241 (11.68%)
	229	13	0.686	0.098 (85.70%)	0.233 (66.08%)
DeepSeek-R1	2	20	0.471	0.126 (73.30%)	0.134 (71.54%)
	15	17	0.596	0.267 (55.12%)	0.250 (58.04%)
	196	5	0.606	0.324 (46.47%)	0.334 (44.76%)
	229	21	0.169	0.092 (45.86%)	0.170 (−0.67%)

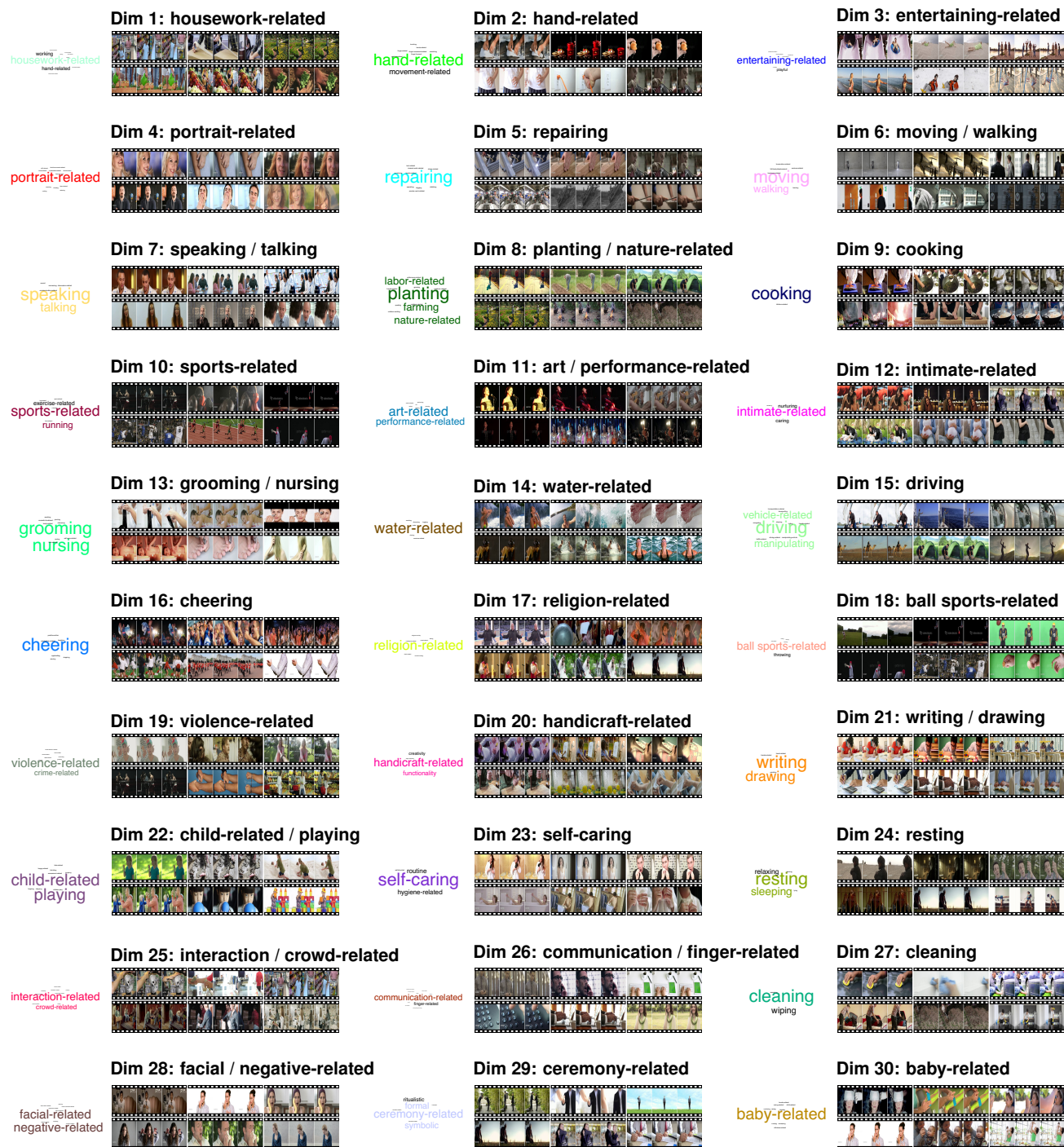


Supplementary Fig. 2. Assessment of positional response distributions and dataset size saturation. **a**, Distributions of the selected odd-one-out positions (Video IDs 1, 2, or 3) across the four primary foundation models. While text-only LLMs exhibit a recency bias (favoring the third position) and MLLMs show a slight preference for the second position, the responses remain broadly distributed across all available options. This confirms that despite inherent positional heuristics common in autoregressive models, the systems did not collapse into deterministic positional responding, ensuring that the derived embeddings are driven by semantic content comparisons. **b**, Extrapolation of recoverable semantic dimensions as a function of dataset size (in units of 10,000 trials). The number of stable SPoSE dimensions (y-axis) is plotted against sub-sampled training volumes (x-axis). Violin plots and scatter points show the distribution of dimensions at each step, with error bars denoting the mean $\pm$ standard deviation. The exponential fit (red dashed line) and its bootstrapped 95% CI (gray shaded region) demonstrate that the extracted dimensionality saturates at approximately 30 dimensions around 300,000 trials. The total chosen dataset size (approximately 550,000 trials, indicated by the black star) safely exceeds this saturation threshold.

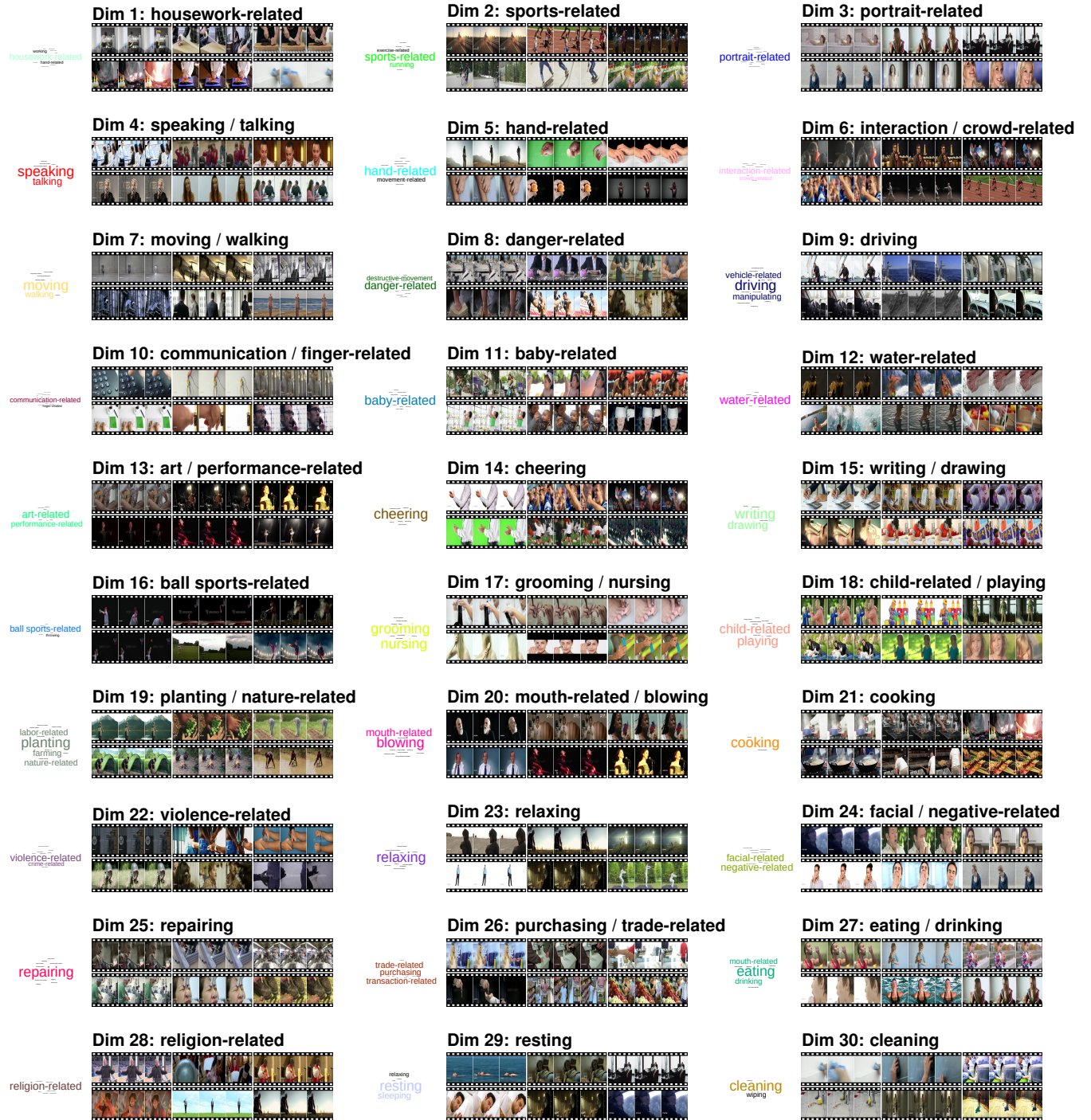


Supplementary Fig. 3. Dimension-wise representational alignment between human and artificial action spaces.

Heatmaps display the Pearson cross-correlation matrices comparing the action-loading profiles of the 30 SPOSE dimensions extracted from the human benchmark against those derived from Qwen2.5-7B (a), DeepSeek-R1 (b), Qwen2.5-VL-7B (c), and Qwen2.5-VL-72B (d). To facilitate visual comparison, the columns (artificial model dimensions) have been computationally reordered to match their most highly correlated human counterparts. The prominent red diagonal across all four matrices demonstrates a strong relationship, indicating that for the majority of human semantic axes, there exists a corresponding, highly aligned dimension within the foundation models. Conversely, the presence of off-diagonal elements (lighter red and blue patches) reflects systemic divergences, where artificial models exhibit residual cross-talk or conflation between semantic dimensions in the human cognitive space.



Supplementary Fig. 4. Visualization of learned action dimensions for Qwen2.5-7B. For each dimension, we present its corresponding action label(s) and the six video frames that received the highest weight.



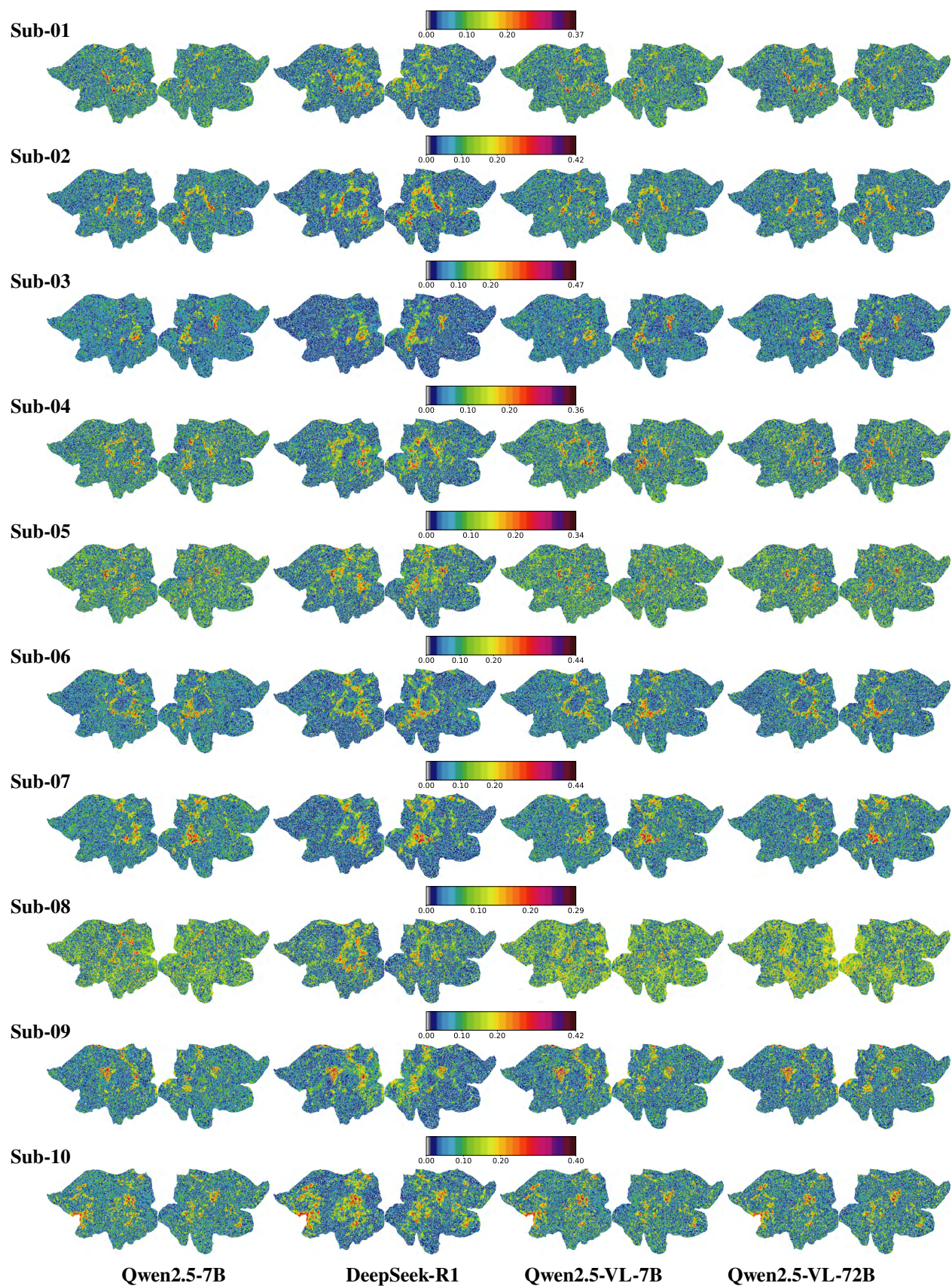
Supplementary Fig. 5. Visualization of learned action dimensions for DeepSeek-R1. For each dimension, we present its corresponding action label(s) and the six video frames that received the highest weight.



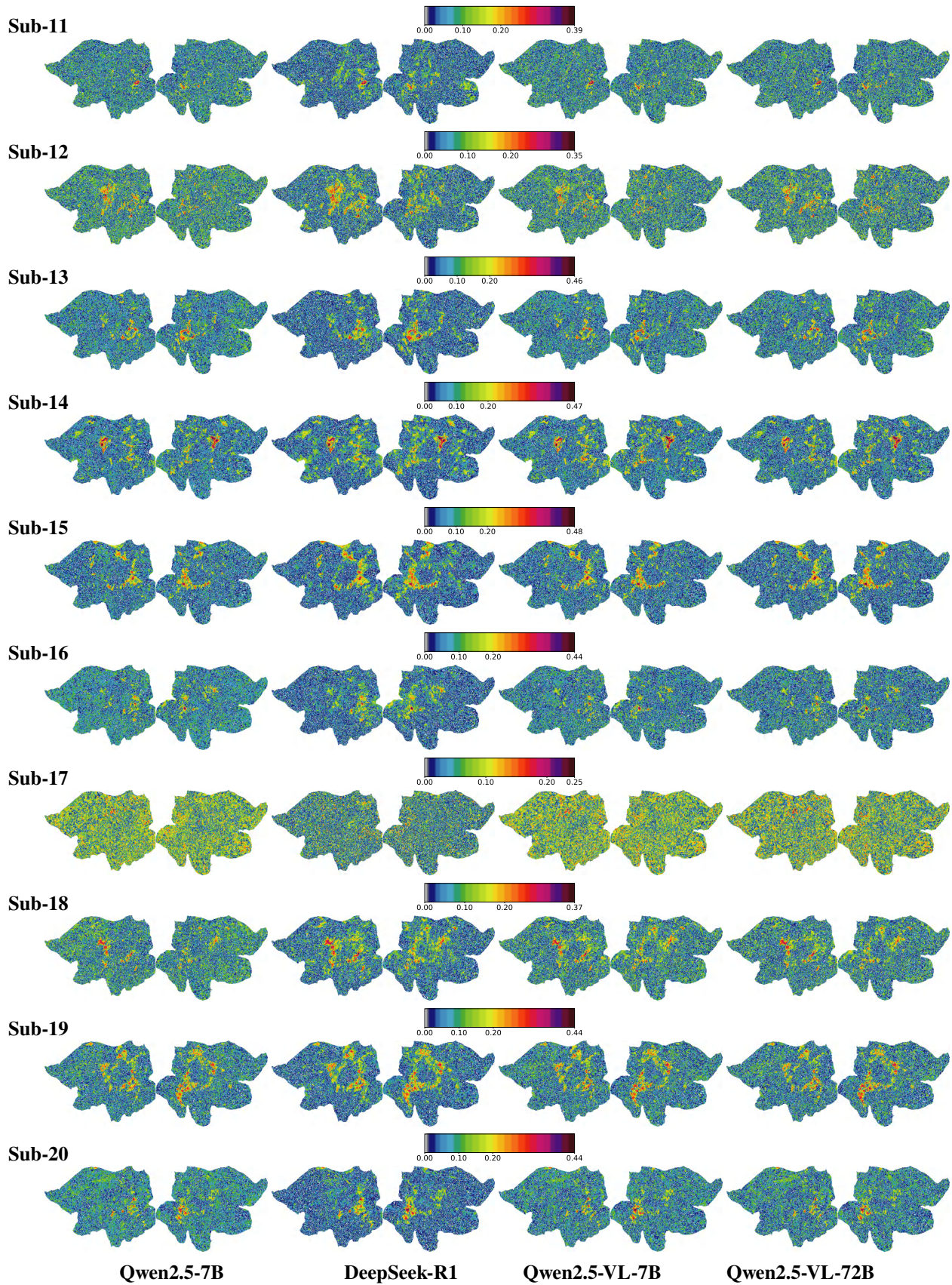
Supplementary Fig. 6. Visualization of learned action dimensions for Qwen2.5-VL-7B. For each dimension, we present its corresponding action label(s) and the six video frames that received the highest weight.



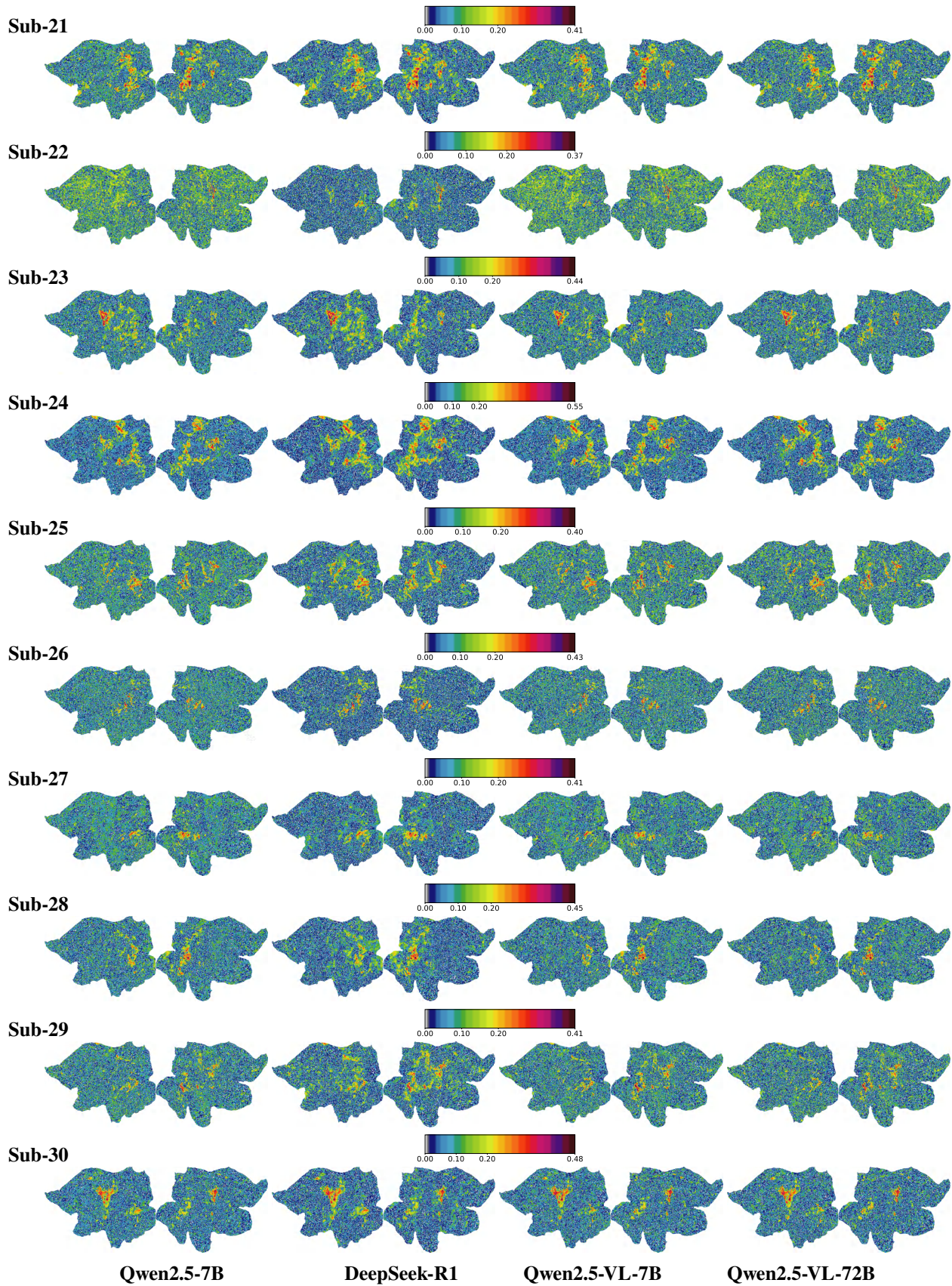
Supplementary Fig. 7. Visualization of learned action dimensions for Qwen2.5-VL-72B. For each dimension, we present its corresponding action label(s) and the six video frames that received the highest weight.



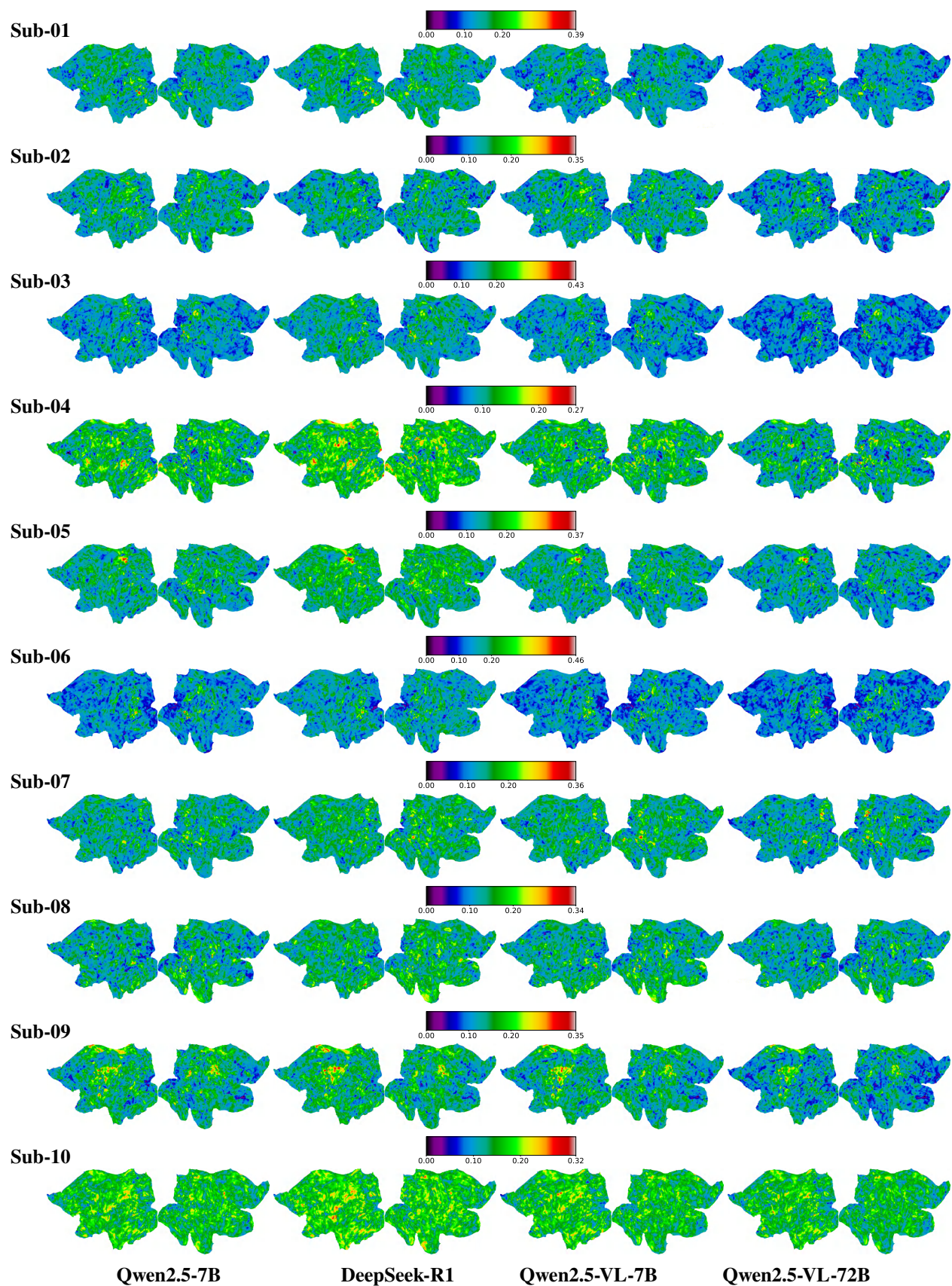
Supplementary Fig. 8. Results of voxel-wise encoding for individual subjects (Subjects 1-10).



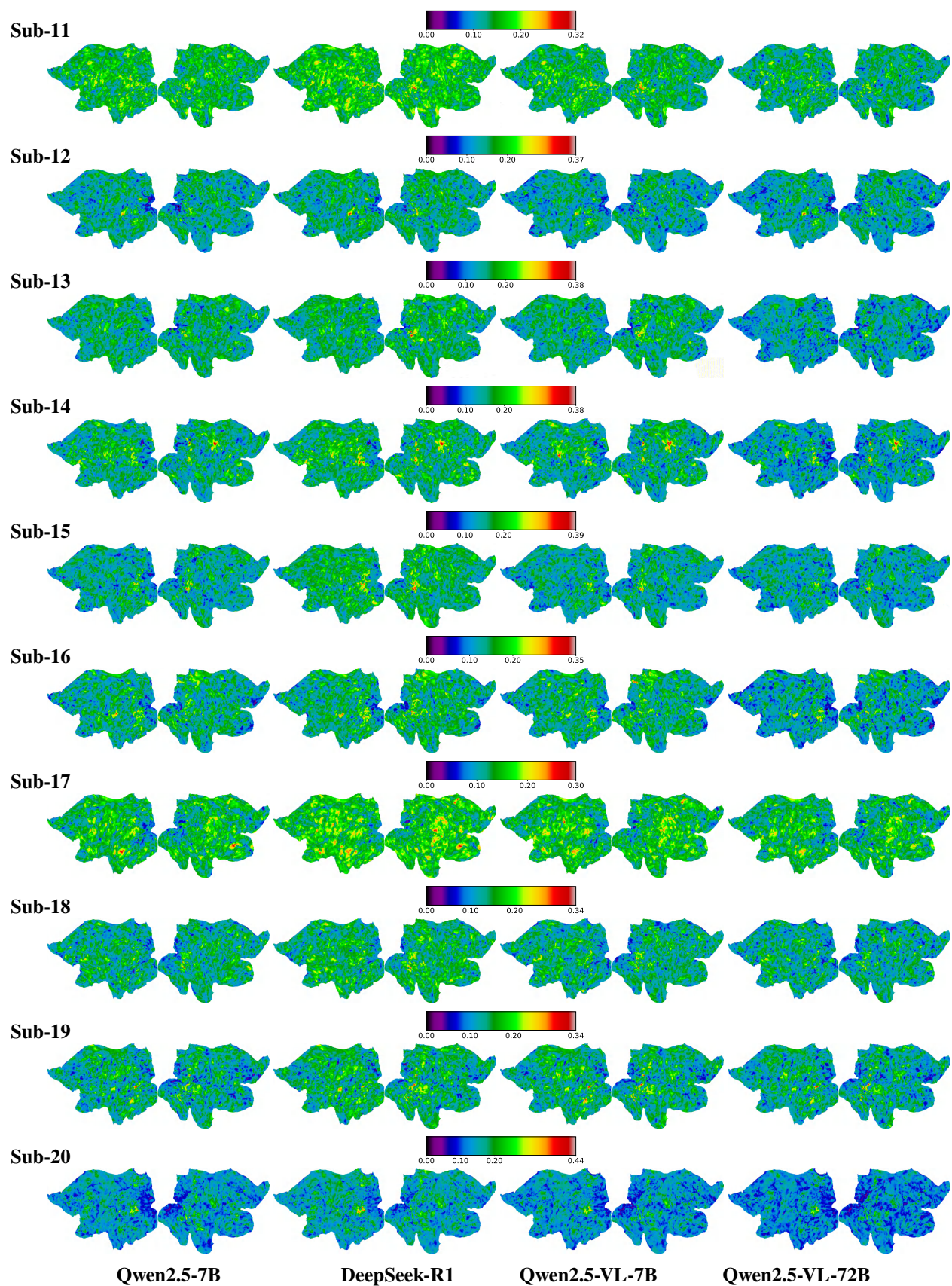
Supplementary Fig. 9. Results of voxel-wise encoding for individual subjects (Subjects 11-20).



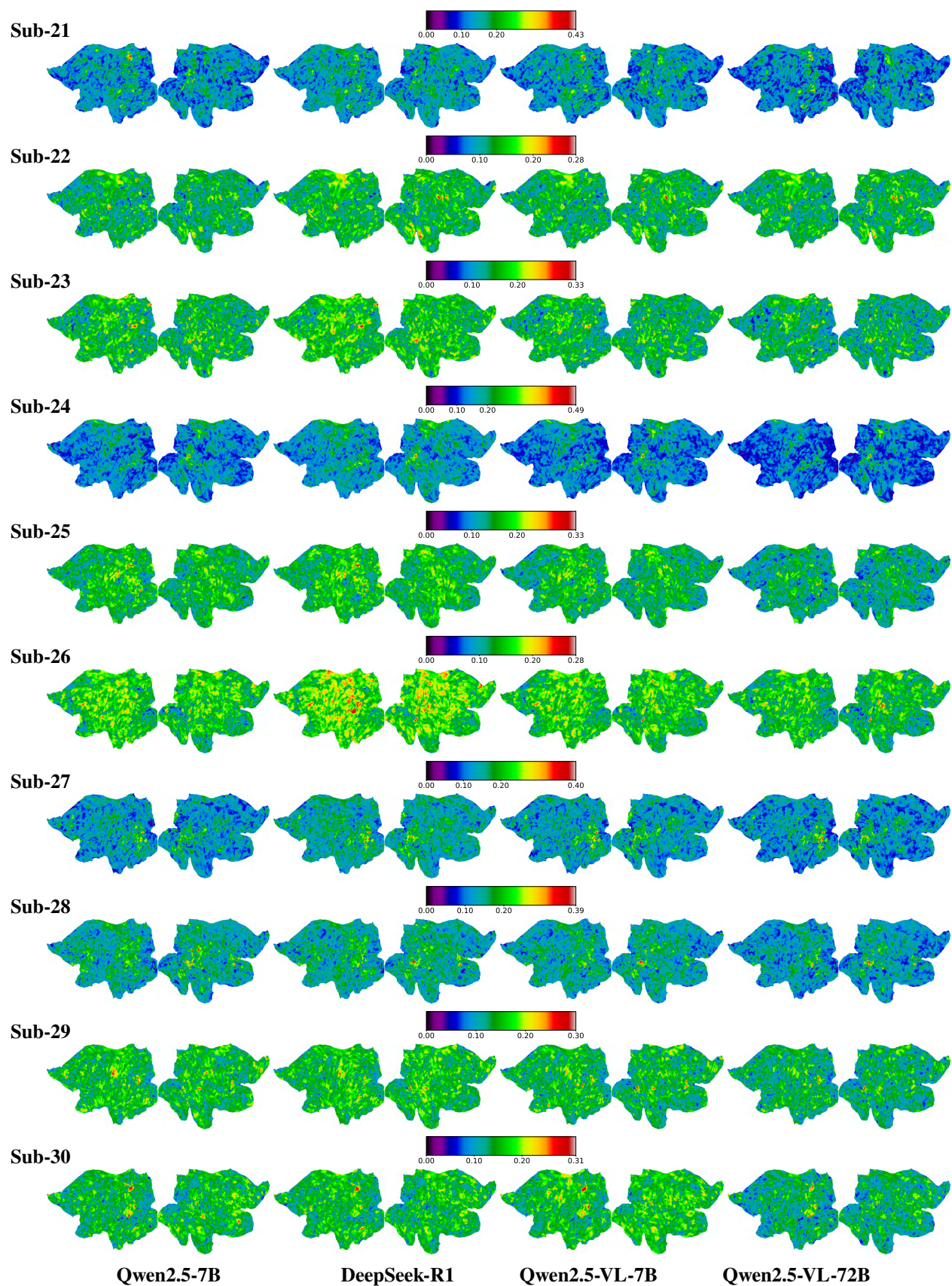
Supplementary Fig. 10. Results of voxel-wise encoding for individual subjects (Subjects 21-30).



Supplementary Fig. 11. Results of searchlight RSA for individual subjects (Subjects 1-10).



Supplementary Fig. 12. Results of searchlight RSA for individual subjects (Subjects 11-20).



Supplementary Fig. 13. Results of searchlight RSA for individual subjects (Subjects 21-30).