EVALUATING POINT-LIGHT BIOLOGICAL MOTION IN MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

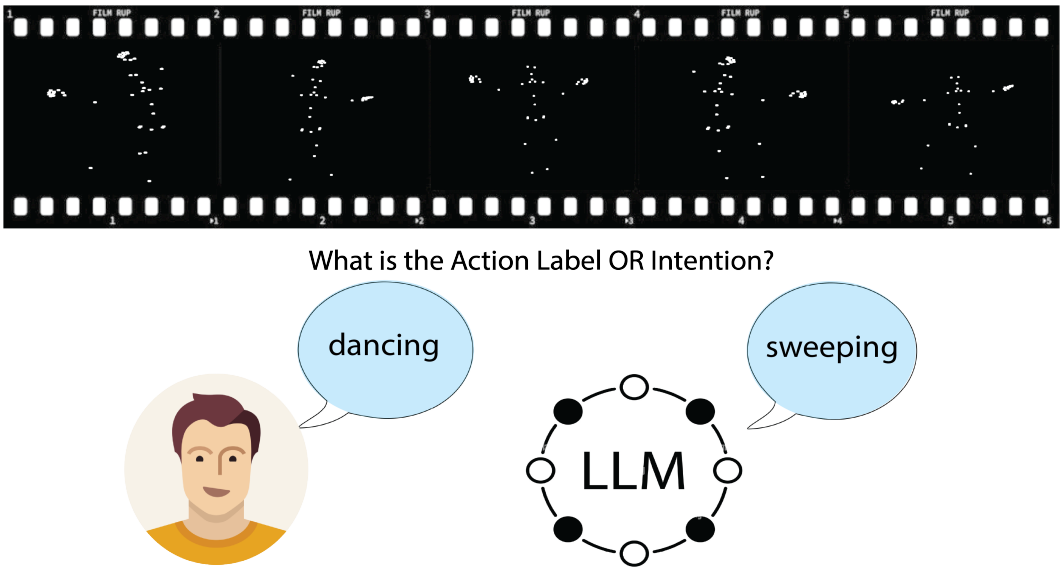

Figure 1: Point-light displays (PLDs) reveal a critical gap between human and artificial biological motion perception. While humans intuitively reconstruct 3D trajectories from sparse visual cues to infer biological motion and rich attributes from PLDs, current multimodal large language models rely on 2D feature aggregation that often fails to capture the necessary semantic and spatiotemporal grounding. The PLD reel depicts an agent dancing.

ABSTRACT

Humans can extract rich semantic information from minimal visual cues, as demonstrated by point-light displays (PLDs), which consist of sparse sets of dots localized to key joints of the human body. Multimodal large language models (MLLMs), despite demonstrating progress on various multimodal tasks, currently lack such structural and semantic abstraction required to interpret human motion. Since PLDs isolate body motion as the sole source of meaning, they present a key stimulus for testing the constraints of action understanding in these systems. Here we introduce ActPLD, the first benchmark to evaluate action processing in MLLMs from PLDs. Tested models include state-of-the-art proprietary and open-source systems on single-actor and socially interacting PLDs. Our results reveal consistently low performance across models, introducing fundamental gaps in action processing and spatiotemporal understanding.

1 INTRODUCTION

1.1 BACKGROUND

Humans can infer rich attributes from minimal visual kinematics, such as intentionality (Manera et al., 2010), identity (Loula et al., 2005), emotions (Atkinson et al., 2004), and even self-awareness (Kadambi et al., 2024) from point-light displays (PLDs) (Johansson, 1973), a set of moving dots localized to key joints on the human body.

This attunement toward sparse biological patterns emerges early in human development, present in newborns (Simion et al., 2008) and may even be acquired prenatally (Reid et al., 2017; Craighero, 2024). The underlying process is often attributed to a common mapping between perception and action in the human brain to facilitate action understanding (Prinz, 1997; Hommel et al., 2001), especially from visually sparse kinematics depicted by PLDs (Giese & Poggio, 2003).

Recent multimodal models aim to achieve artificial general intelligence by meeting or surpassing human-like metrics across sensory inputs including perceptual tasks. Indeed, recent models including GPT-4V (OpenAI, 2023), Gemini 2.5 Pro (Google DeepMind, 2025), CogVLM2-Video (Hong et al., 2024), Qwen2-VL (Wang et al., 2024a), and PLLaVA (Xu et al., 2024) and other vision language models (VLMs) have shown relatively impressive progress across multimodal tasks, including image captioning, scene recognition, and basic video processing.

However, MLLMs lack the temporal, embodied, and structural priors that humans use to understand actions. While they can process the visual information from frames, MLLMs struggle to build world models with the intuitive, four-dimensional (3D space + time). Their relative success in more isolated domains (e.g., image captioning) often does not transfer to spatiotemporal domains (Zhou et al., 2025; Hong et al., 2025) or semantic understanding. These and related tasks and are noted as core gaps between MLLM symbolic and human alignment.

Recent action benchmarks (e.g., MotionBench, (Hong et al., 2025)) demonstrate how state-of-the-art models struggle to identify simple motions from raw videos, particularly on spatiotemporal tasks like action sequencing, attributed to their rudimentary motion perception ability. Other benchmarks (e.g., LlavaAction, (Ye et al., 2025)) found that even basic distractor semantic labels impede action recognition performance.

As MLLMs demonstrate growing competence in video understanding, several reasons motivate the increasing importance to also measure their performance on sparse, dynamic PLDs. First, PLDs are an excellent stimulus, as they are unconfounded by texture, background, or object affordances. Unlike naturalistic videos which provide contextual scaffolds, PLDs provide a deeper lens into action understanding by isolating motion. Second, spatiotemporal information from PLDs is rather unique, since humans seem to have dedicated neural resources for biological motion (Grosbras et al., 2012), which are intricately tied to our own body schema and body experience. Biological motion understanding is also crucial to build world models of the physical and social environment. Since humans understand actions through action experience (Rizzolatti & Sinigaglia, 2016) and navigation in the world, the comparison is therefore a key metric in the aim for artificial general intelligence. Finally, real-world perception often involves noisy or partial data, which PLDs emulate in a controlled manner. While no benchmark or study to date has measured how MLLMs perceive human motion from PLDs, cursory tests show that they often mistake human point-light walkers for constellations or rotating lines. Even minimal viewpoint shifts or occlusions appear to degrade performance.

1.2 Contributions

Here, we introduce ActPLD, the first benchmark to preliminarily assess MLLMs' action processing performance from human PLDs. Specifically we measure the following:

- **Action Classification:** Can the model identify the action from sparse kinematics?
- **Intent Inference:** Can the model attribute intentionality (purposeful behavior) to the dots?

We evaluate model performance via:

- **3 Alternative Forced Choice (3-AFC):** What is the correct choice out of 3 options?
- **Chain-of-Thought (CoT) Description Similarity:** How well do model-generated action descriptions match human labels?

1.3 MLLM architectures

MLLMs aim to unify perception and language understanding. MLLM architectures traditionally consist of three main components: modality encoders, large language models and a modality interface that connects the two. Modality encoders are specialized modules that transduce input data into a readable representation (typically vector embeddings) for the LLM. The LLM component is the

basis for the advanced reasoning faculties in MLLMs. The modality interface bridges the modality encoders with LLMs to facilitate communication between these two systems. A fourth component, a generator module, is sometimes also included in MLLMs to produce outputs in modalities beyond text (Brohan et al., 2023; Driess et al., 2023).

These components are functionally separable. Modality encoders typically tokenize information by generating embeddings analogous to characters or words or use feature-level fusion to extract relevant features across multiple modality encoders. Feature-level fusion uses specialized mechanisms like cross-attention (Alayrac et al., 2022), which internally aligns encoder output with internal LLM representations, or adapter modules to selectively prioritize relevant multimodal features.

1.4 Theoretical and Practical Gaps

In contrast to the feedforward and modularized processing described above, human action understanding depends on prior experience and is deeply shaped by bodily experiences and internal states, as well as predictive and forward modeling of these experiences.

By contrast, MLLMs lack access to either form of embodiment. They can parse language about physical states (e.g., "I'm tired") but may lack the internal representations to ground such language experientially. This lack of bodily grounding has been used to explain major limitations in their ability to reason about causality, intention, and motoric processing, notably in visually sparse inputs such as PLDs.

Beyond lacking internal bodily states, MLLMs also deal with other technical limitations (Mon-Williams et al., 2025). Video understanding requires integrating space, time, and multimodal signals in a continuous way, and current MLLMs perform poorly on video understanding benchmarks. Video captioning in these models typically occurs at the scene level, which overlooks more granular spatiotemporal details (Zhou et al., 2025). Benchmarks like MVBench (Li et al., 2024) and VLM4D (Zhou et al., 2025) consistently show that MLLMs lag far behind humans in temporal abstraction, motion understanding, and audio-visual integration, even for simple sequences. MotionBench (Hong et al., 2025) found that models achieve only 55–58% mean accuracy on motion processing tasks (e.g., motion recognition, repetition counting, relative location tracking, camera motion reasoning), often near random baselines on rapid or subtle movements. These failures are attributed to aggressive, potentially lossy visual compression needed for processing and shallow temporal fusion that disrupts the sequence of the videos. Even top-performing MLLMs may degrade sharply on temporally sensitive tasks, and small PLD perturbations (e.g., rotation, occlusion) may impair model predictions. To date, current benchmarks like VLM4D (Zhou et al., 2025) and OpenVLA (Chen et al., 2025a) have examined higher-level spatial-temporal reasoning in semantically rich scenes. However, these limitations could become even clearer on abstract generalization tasks like action understanding from PLDs, using motion as the primary cue.

1.5 Current Action Understanding Benchmarks

While no LLM benchmark has yet evaluated human PLDs, we broadly divide related action benchmarks into two sets to better describe the current landscape:

1.5.1 Action Understanding Benchmarks

These benchmarks evaluate action understanding on contextually rich video stimuli. We briefly describe these below:

- **MotionBench** (Hong et al., 2025) – fine-grained motion perception across 6 task types; reveals sub-60% accuracy for state-of-the-art VLMs.
- **MVBench** (Li et al., 2024) – multimodal video understanding suite, including 20 video tasks, including action recognition and reasoning. Models generally score below 30%.
- **LVBench** (Wang et al., 2024b) – focuses on long-video understanding, such as narrative consistency, event recall, and long-term/extended reasoning. The benchmark includes multi-hour videos with structured annotations. LVBench has become an important evaluation suite and demonstrates that MLLMs significantly underperform on these tasks. They also note substantial performance leaps as the number of evaluated frames increases.

- **LongVideoBench** (Wu et al., 2024) – focused on long-context model performance over extended temporal videos.
- **Video-MME** (Fu et al., 2024) – spans 900 videos (∼254 hours) and 2700 Q&A pairs across six visual domains. Includes both long-range (hour) and short-range videos. Model performance, particularly Gemini 1.5 Pro, showed relatively strong performance (∼75%).
- **MLVU** (Zhou et al., 2024) – long-range videos (3 min–2 hours) involving nine evaluation tasks across multiple genres and domains. Model performance achieves ∼64% accuracy on multiple choice tasks.
- **SPORTQA** (Xia et al., 2024) – includes sports knowledge and reasoning across three difficulty levels.
- **BioMotion Arena** (Chen et al., 2025c) – uses LLMs to generate artificial (simulated) biological motion animations, validated with human annotation to determine which motions were more biologically plausible. Model performance was greatest for Gemini 2.5 pro, though stayed relatively low across models.

1.5.2 Embodied Action Benchmarks

There is also active discussion around embodied agents using MLLMs, typically integrated with robotics. Most evaluations focus exclusively on external embodiment—robotic manipulation, navigation, and planning. Consistently, these embodied models show especially poor performance, and relative to humans, on embodied tasks. We briefly describe them below:

- **VisualAgentBench** (Liu et al., 2024) – designed to train and evaluate visual MLLMs on different tasks on datasets related to embodiment, graphical user interface, and visual design. Even the strongest proprietary models performed poorly (highest average performance was GPT-4o ∼36.2%).
- **ELLMER (Embodied LLM-enabled Robot)** (Mon-Williams et al., 2025) – couples GPT-4 with a sensorimotor loop of vision and force feedback to carry out complex tasks (e.g., making coffee, drawing on plates). Despite high-level reasoning success, the system still relies on pre-curated knowledge bases and hard-coded loops to compensate for the absence of internal control models or homeostatic feedback.
- **EmbodiedEval** (Cheng et al., 2025) – evaluated MLLMs across embodied tasks, including attribute question answering, spatial question answering, navigation, object interaction, and social interactions across a range of diverse interactions and scenes. Success rates across tasks remained consistently low. The best-performing model (GPT-4o) achieved only a 25% success rate in contrast to humans (near-perfect 97.26%).
- **EmbodiedBench** (Yang et al., 2025) – evaluated MLLMs on 1,128 testing tasks (ranging from high-level to low-level semantic tasks) across four environments. They found that MLLMs excel at high-level tasks but struggle with lower-level tasks, with the best model, GPT-4o, scoring only 28.9% on average.
- **ECBench** (Dang et al., 2025) – an embodied benchmark focused on egocentric perception, covering three sets: static scenes, dynamic scenes, and hallucinations. Across MLLMs, GPT-4o showed highest mean performance (∼50.35%), though significantly below human performance (∼94.96%). Notably, the top-performing embodied models (grounded in situational understanding) performed especially poorly (mean performance ∼21–35%).

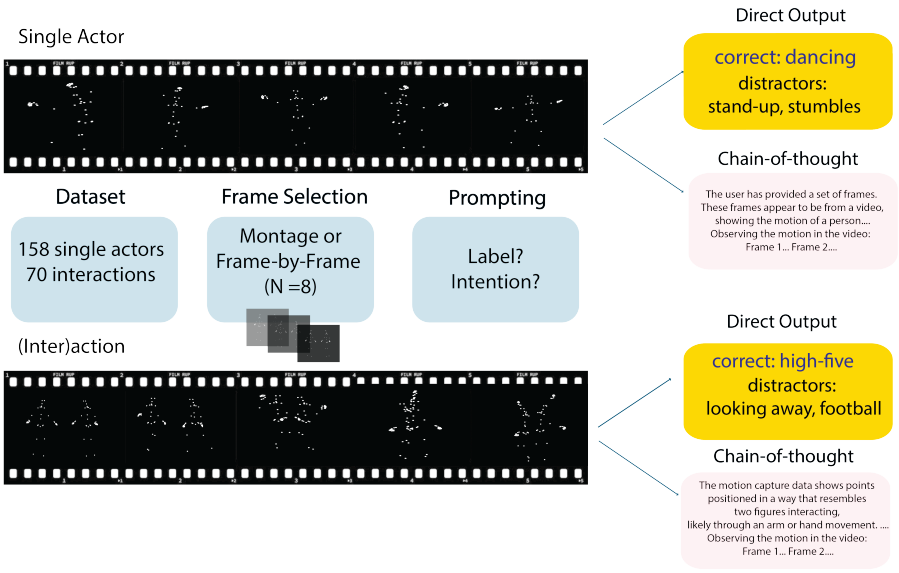

Figure 2: ActPLD Model Processing depicting single actor PLD dancing (top) and high-fiving social interaction (bottom). A total of 228 actions were used to construct the benchmark, with 158 depicting a single actor PLD and 70 depicting social interacting PLDs. The most relevant frames were selected by model parsing and models were prompted to infer the action label or intention. Sample direct output and chain-of-thought descriptions are shown on the right.

1.6 ACTPLD

Building on the prior lines of research, we construct ActPLD, designed to evaluate point-light biological motion on current proprietary and open-source MLLMs. We evaluate models using visually sparse point-light displays, which eliminate visual confounds like texture, background, or objects. We focus on action understanding, evaluated on action recognition and intentionality inferences. We evaluate both single-actor and social interactions in PLDs. Our evaluation setup also includes Chain-of-Thought prompting and 3-Alternative Forced Choice (3-AFC).

All PLDs are generated from real human motion-capture data covering 158 individual actions (CMU dataset: `https://mocap.cs.cmu.edu/`) and 70 social interactions (34 from CMU and 36 from (Manera et al., 2010)). The actions span large-scale movements and interactions to subtle gestures. As described above, our evaluation is measured by independent MLLMs (GPT-5 and Gemini 2.5-Pro) using descriptions from Chain-of-Thought prompting and multiple choice (3 AFC).

2 TASKS

2.1 EVALUATION METRICS

We designed two core evaluation tasks:

- **Action Classification** – Can the model correctly identify the depicted motion?
- **Intention Inference** – Can it attribute purposeful behavior to the motion pattern?

We assess model performance using:

- **3 Alternative-Forced Choice (3-AFC) Accuracy** – Three-alternative forced choice recognition.
- **Chain-of-Thought Consistency** – Coherence of intermediate reasoning steps compared to human reasoning paths

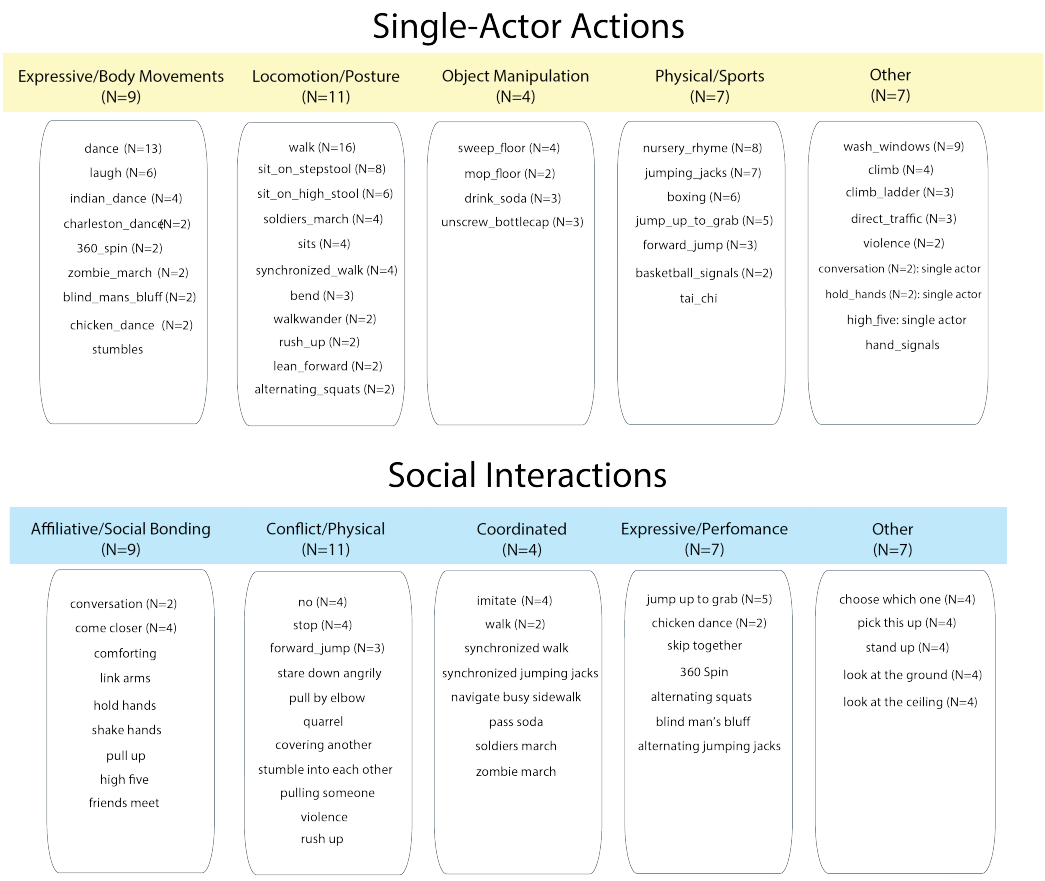

Figure 3: Actions used for benchmarking. We divide into 5 different semantic categories for each group (single actor vs social interactions). All single-actor actions were taken from the CMU motion capture database (`https://mocap.cs.cmu.edu/`). Social interactions were taken from CMU and (Manera et al., 2010).

2.2 Models Evaluated

We benchmark both proprietary and open-source top-ranked MLLMs: GPT-5, GPT-4o, Gemini 2.5 Pro, Claude-Sonnet-4, Claude Opus 4.1, and Qwen2.5-VL-72B.

2.3 Benchmark Construction

PLD videos were sourced from the open-source Carnegie Mellon Motion Capture (CMU) dataset and (Manera et al., 2010). 158 everyday actions were taken from the CMU database for the single-actor tasks. Thirty-six interactions from (Manera et al., 2010) were used for the social interactions, and of the 70 total social interactions, the remaining 34 interactions were taken from the CMU dataset. See Table 1 for a summary and Figure 3 for all the action labels. PLDs were created by displaying and capturing the videos using BioMotionToolbox (van Boxtel & Lu, 2013) in MATLAB R2024a and Psychtoolbox.

2.4 Frame Extraction

For each video, eight consecutive frames (128 x 128 px) were extracted from a central portion of the script. This was determined by trimming the first and last 10% of frames to avoid any artifacts (e.g., T-pose) and computing a middle window large enough to fit eight consecutive frames. If the model's response was uninformative, the sampling window was shifted earlier or later in the clip on subsequent attempts (up to 3 retries). For Gemini, the eight frames were combined into a 4x2 grid montage for better compatibility. For all other providers (Anthropic, OpenAI, Qwen, and Replicate), the frames were individually inputted as base64-encoded images.

Table 1: Benchmark Construction Summary

Action Category	Dataset	Total
Single-Actor	Carnegie Mellon Motion (CMU) Capture Database	158
Interactive	(N=36) from (Manera et al., 2010); (N=34) from CMU Motion Capture Database	70

2.5 Model Evaluation

Three-alternative Forced Choice (3AFC) Around 60 intention/action labels were further organized into semantic groups (e.g., sit down, squat down, sit on stool) to prevent overly similar distractors appearing in the 3AFC options. For each video, the ground-truth label was paired with two distractor labels, sampled from all possible labels/intentions while avoiding semantically overlapping groups and consistent across models. The label was parsed directly from the video filename. The three response options were shuffled on each trial to form a 3AFC multiple-choice set. Model performance was evaluated as the percentage of correct predictions across valid trials. Any errors (e.g., due to API failures or timeouts) were excluded. Accuracy was then computed as the mean correct values over the remaining rows.

Chain-of-Thought Description Matching A key focus of the work was to also evaluate the consistency between model accuracy and their reasoning. To determine whether a model's response correctly identified the ground truth action, we used an independent LLM (Gemini 2.5 Pro) to evaluate the model's thought log. Each response was read in full to assess whether the overall description supported the true label. We treated responses as negation-aware: if a label was mentioned in a negated form (e.g., "not sitting"), this counted against it, unless the rest of the description provided strong positive evidence. Positive cue matching was prioritized, supporting phrases such as 'is,' 'shows,' 'depicts,' 'represents,' 'looks like,' 'best described as,' 'demonstrates,' 'doing,' 'most consistent with,' 'matches,' 'resembles,' or 'most likely' appeared near the true label (same sentence or within $\sim$80 characters). If a true label was not explicitly named, we applied a descriptive fallback: the response was considered correct if it clearly described defining features characteristic of the true action (e.g., "flapping arms" for chicken dance), provided that competing labels were not positively supported. Synonyms and paraphrases were normalized to account for minor wording differences (e.g., pick up vs. pick this up). A response was marked as a match if: (i) the true label had more positive than negative evidence; (ii) the true label had at least one positive cue while competing labels were unsupported or negated; or (iii) the descriptive fallback condition applied. Otherwise, the response was marked as not a match. Any errors were excluded from percentage calculations.

3 Results

3.1 Single Actor Evaluation Results

Human (N=2) ground truth performance for single-actor action inference was 93.33% (near-ceiling) and significantly above 3AFC chance (33.33%). While all models performed above chance, performance across all models remained poor, with a range of $\sim$34–41%. GPT-4o scored the highest mean performance for single actors (M=40.19%), Claude Sonnet-4 scored the lowest on average (M=34.02%). Gemini 2.5 Pro uniquely showed poorest performance on single actor accuracy for 3AFC, but significantly higher performance for the CoT Description Matching Task (Figure 4). All model results are reported in Table 2 and individual action results in Appendix A.1.

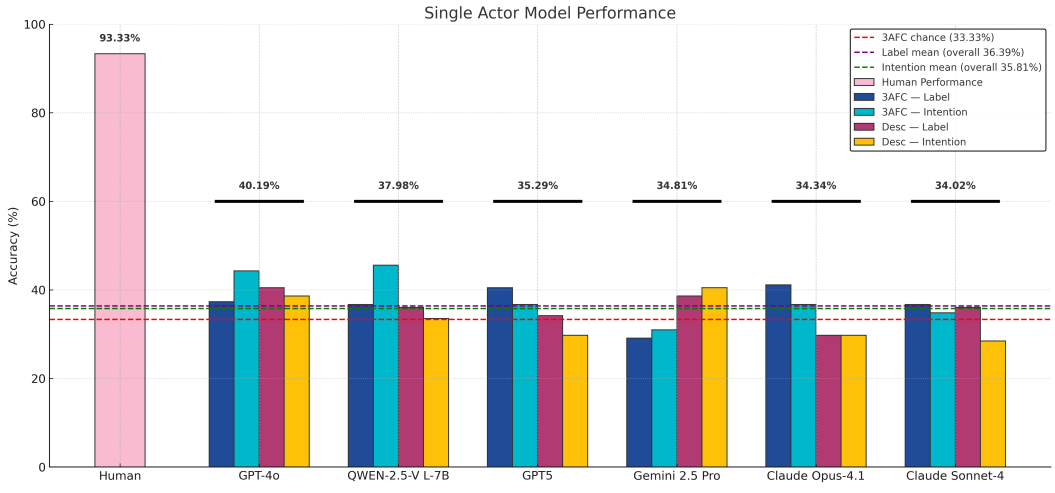

Figure 4: Model Performance for single-actor PLDs evaluated by 3-alternative forced-choice (3AFC) and description matching tasks, and separated into label classification and intention classification. Each model's performance is shown as four grouped bars: 3AFC label (dark blue), 3AFC intention (light blue), description matching label (fuchsia), and description matching intention (yellow). Models are ordered from left to right by their overall mean accuracy. The pink bar (far left) reflects mean human performance (N=2; ∼93.33%). A solid black line spans each model's group of bars, with the bold percentage above indicating the model's mean accuracy across all four conditions. The dashed red line indicates chance performance (33.3%), while the dashed green and purple lines indicate the mean accuracies across models for label and intentions, respectively.

Table 2: Single Actor Performance Results

Model	Type	3AFC	Description Matching
Claude Opus-4.1	Label	41.14	29.75
Claude Opus-4.1	Intention	36.71	29.75
Claude Sonnet-4	Label	36.71	36.08
Claude Sonnet-4	Intention	34.81	28.48
GPT-4o	Label	37.34	40.51
GPT-4o	Intention	44.30	38.61
Gemini 2.5 Pro	Label	29.11	38.61
Gemini 2.5 Pro	Intention	31.01	40.51
Qwen-2.5-VL-7B	Label	36.71	36.08
Qwen-2.5-VL-7B	Intention	45.57	33.54
GPT-5	Label	40.51	34.18
GPT-5	Intention	36.71	29.75

3.2 Social Interaction Evaluation Results

Human (N=2) ground truth performance for single-actor action inference was 93.5% (near-ceiling) and significantly above 3AFC chance (33.33%). All models similarly performed above chance performance on 3AFC (33.33%) for social interactions. While model performance across all models still remained poor, with a range of ∼28–50%, GPT-5 and Gemini scored significantly higher on Social Interactions than for Single Actors. Gemini 2.5 Pro scored the highest mean performance for Social Interactions (M=49.95%), Claude Opus-4.1 scored the lowest (M=28.57%). Note that Gemini 2.5 Pro showed the poorest performance on single actor accuracy for 3AFC, but the highest 3AFC accuracy for social interactions. All model results are reported in Table 3 and individual interaction results in Appendix A.1.

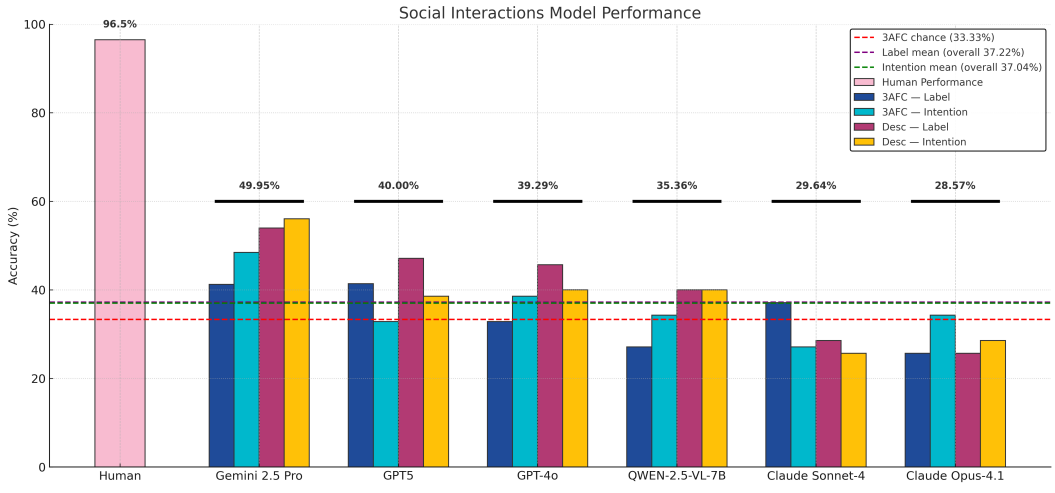

Figure 5: Social Interactions Model Performance for PLDs evaluated by 3-alternative forced-choice (3AFC) and description matching tasks and separated into label classification and intention classification. Each model's performance is shown as four grouped bars: 3AFC label (dark blue), 3AFC intention (light blue), description matching label (fuchsia), and description matching intention (yellow). Models are ordered from left to right by their overall mean accuracy. The pink bar (far left) reflects mean human performance (N=2; ∼96.5%). A solid black line spans each model's group of bars, with the bold percentage above indicating the model's mean accuracy across all four conditions. The dashed red line indicates chance performance (33.3%), while the dashed green and purple lines indicate the mean accuracies across models for label and intentions, respectively.

Table 3: Social Interactions Evaluation Results (%)

Model	Type	3AFC	Description Matching
Claude Opus-4.1	Label	25.71	25.71
Claude Opus-4.1	Intention	34.29	28.57
Claude Sonnet-4	Label	37.14	28.57
Claude Sonnet-4	Intention	27.14	25.71
GPT-4o	Label	32.86	45.71
GPT-4o	Intention	38.57	40.00
Gemini 2.5 Pro	Label	41.27	53.97
Gemini 2.5 Pro	Intention	48.48	56.06
Qwen-2.5-VL-7B	Label	27.14	40.00
Qwen-2.5-VL-7B	Intention	34.29	40.00
GPT-5	Label	41.43	47.14
GPT-5	Intention	32.86	38.57

3.3 Comparison of Evaluations

Finally, we compared the consistency of model evaluation results by performing a Spearman correlation on the 3AFC and Description Matching Accuracy Values. The models showed a very strong performance relationship for social interactions, Spearman's $\rho = .725$, $p = .008$, 95% CI: [.240, .920]. However, no significant consistency relationship was observed for the single actor PLD, $\rho = -.112$, $p = .730$.

Table 4: Social interactions vs single actor overall performance across models

	Single Actor	Social Interactions
Mean 3-AFC	37.55	35.10
Mean Description Matching	34.65	39.17
Overall Means	36.10	37.14

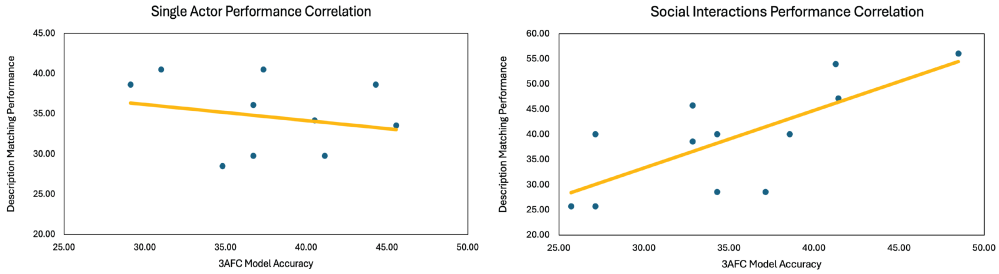

Figure 6: Relationship between 3AFC performance (x-axis) and CoT Description Matching Performance (y-axis). Left plot for single actor shows no significant relationship ($p = .730$); right plot shows a strong positive association ($p = .008$).

4 Discussion

Recognizing and inferring actions from minimal visual data, as in the case of point-light displays (PLDs), is a hallmark of human action understanding. From PLDs, humans can infer body structure from sparse two-dimensional projections and map these projections to abstracted action categories. To date, however, it remains unknown whether such action inference ability is similarly present in current large language models.

Here, we introduce ActPLD, a benchmark constructed to evaluate human PLDs for action understanding in MLLMs. Across state-of-the-art MLLMs, we observed consistently poor performance for PLDs, with accuracy scores hovering just above chance levels ($\sim$28–41% for single-actors, 25–49% for social interactions). Socially interacting PLDs were overall better identified and described than the single-actor PLDs and were consistent between 3AFC and description matching evaluations (Spearman's $\rho = .725$, $p = .008$). However, the single actor results did not reveal a consistent relationship ($\rho = -.112$, $p = .730$). Our findings reinforce that verbosity or relevance in thought logs do not guarantee correct final answers, especially in sparse perceptual settings. This mismatch contributes to the critical gap between the appearance of reasoning and the correct substance of the response, and in particular tethered to the correct perceptual judgment.

One explanation for these findings pertains to the minimalistic nature of these stimuli. Single actor PLDs force models to rely more heavily on integrating individual joint trajectories from sparse dots, which appears to be a current technical limitation (Zhou et al., 2025). Indeed, previous embodied benchmarks, such as EmbodiedBench (Yang et al., 2025), show that these models struggle on lower-level action tasks, while performing better on higher-level semantic tasks. The null correlation between 3AFC and description matching for single actors could further indicate that models are employing inconsistent reasoning strategies, potentially relying on different visual features or temporal windows depending on the evaluation format. On the other hand, for social interactions, the additional contextual information provided by the actors provides richer visual and structural priors, which may help maintain evaluation performance.

Several technical bottlenecks further compound the reasoning difficulties in these models. Current MLLMs are constrained by sampling limitations (typically $\sim$4–16 frames per video sequence). While humans can extract action information from even single frames with sufficient spatial context, downsampled PLD frames may contain insufficient positional information for meaningful inference. Additional technical constraints include poor cross-modal fusion ability, in reference to how models tightly integrate information from multimodal input types into a unified representation.

Of all tested models, Gemini 2.5 Pro showed a clear dissociation: it achieved the highest accuracy on social interactions (49.95%) but was ranked among the weakest on single-actor 3AFC, while at the same time performing relatively well on single-actor description matching. This could suggest an architectural trade-off in cross-modal fusion ability, which normally presents as an advantage relative to other models, but could work against it in sparse visual settings. Specific to single-actor PLDs, the model may lean too heavily on linguistic priors and could over-elaborate its interpretations. To this end, our results confirmed plausible free-text descriptions in Gemini, but ultimately reached the incorrect final solutions with predefined labels.

Our results together suggest that current approaches in MLLMs are insufficient for human action understanding. This is part of a broader set of challenges faced by multimodal systems. As recently surveyed (Chen et al., 2025b), the field still struggles with scalability and generalization, particularly with spatiotemporal information. These implications extend beyond action recognition and ultimately to core questions about achieving embodied, human-like artificial general intelligence.

4.1 Future Directions

- **Improvements to spatiotemporal processing:** current models face limitations in downsampling (∼4–16 frames per clip), which loses informative temporal information. Implementing recurrent and dynamic temporal memory or hierarchical attention mechanisms may preserve and prioritize informative movement trajectories.
- **Body schema problem:** humans have a structured prior of a skeletal model. We can automatically infer invisible joints, limb lengths, etc. Current MLLMs lack this inductive prior entirely. To MLLMs, PLDs are not uniquely prioritized, as a collection of independent pixel trajectories. It is therefore important to develop architectures that can properly represent human body models from sparse observations, as is represented in humans. This might require adding kinematic experience or constraints to these models (e.g., via external sensors) and/or attention mechanisms that capture joint dependencies or even body related modules.
- **Expanding the training dataset** Training on (a) sparse motion like PLDs could improve performance, but risk overfitting and increase reliance on the input data. An alternative could be (b) to focus on compositional generalization beyond PLDs, such as training on other low-fidelity modalities, including stick figures, motion trajectories, cartoons, sensor data, etc., to accommodate generalization to PLDs and other human motion.
- **Modeling internal states and dynamics (i.e., integrating motor experience):** Human action understanding is based on motor experience and simulation mechanisms. MLLMs could integrate and strengthen motor features (or motor-derived priors) via reinforcement learning or movement trajectory simulations that could add experiential grounding. This could improve model reasoning, particularly related to ongoing action dynamics.
- **Training for prediction:** Training models on prediction tasks (e.g., anticipating action outcomes) could help to internalize temporal dynamics of the actions. Some approaches could include engaging in counterfactual simulations (reasoning about alternative futures), next-frame prediction, or inferring action or social outcomes.

5 Conclusion

We introduced ActPLD, the first benchmark to preliminarily evaluate action understanding in multimodal large language models (MLLMs) using human point-light displays (PLDs) to assess whether models can extract biological structure and meaning from minimal biological cues. Across state-of-the-art proprietary and open-source systems, performance on both single-actor and social interaction PLDs remained close to chance, indicative of critical limitations in current model architectures.

LLM Disclosure

In accordance with the ICLR 2026 policy, we disclose that LLMs were used in the following ways (1) for testing and validating parts of the analysis, as the focus of this paper (2) for assistance with figure plotting.

REFERENCES

J.-B. Alayrac, J. Donahue, P. Luc, and et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022. arXiv:2204.14198.

A. P. Atkinson, W. H. Dittrich, A. J. Gemmell, and A. W. Young. Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception*, 33(6):717–746, 2004. doi: 10.1068/p5096.

A. Brohan, N. Brown, J. Carbajal, and et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

K. Chen, X. Liu, J. Wang, and et al. Openvla: A 7b-parameter open-source vision-language-action model. *Information Fusion*, 2025a. doi: 10.1016/j.inffus.2025.103198.

S. Chen, Z. Wu, K. Zhang, C. Li, B. Zhang, F. Ma, and Q. Li. Exploring embodied multimodal large models: Development, datasets, and future directions. *Information Fusion*, pp. 103198, 2025b. doi: 10.1016/j.inffus.2025.103198.

Z. Chen, L. Deng, Z. Chen, K. Zhang, Q. Jia, Y. Tian, Y. Zhu, and G. Zhai. Can large models fool the eye? a new turing test for biological animation. *arXiv preprint arXiv:2508.06072*, 2025c. URL `https://arxiv.org/abs/2508.06072`.

Z. Cheng, Y. Tu, R. Li, S. Dai, J. Hu, S. Hu, J. Li, Y. Shi, T. Yu, W. Chen, L. Shi, and M. Sun. Embodiedeval: Evaluate multimodal llms as embodied agents. *arXiv preprint arXiv:2501.11858*, 2025.

L. Craighero. An embodied approach to fetal and newborn perceptual and sensorimotor development. *Brain and Cognition*, 179:106184, 2024.

R. Dang, Y. Yuan, W. Zhang, Y. Xin, B. Zhang, L. Li, L. Wang, Q. Zeng, X. Li, L. Bing, and et al. Ecbench: Can multi-modal foundation models understand the egocentric world? a holistic embodied cognition benchmark. *arXiv preprint arXiv:2501.05031*, 2025.

D. Driess, M. Ewerton, M. Shridhar, and et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

C. Fu, R. Zhang, Y. Li, Q. Wu, D. Lin, X. Wang, and Y. Shan. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

M. A. Giese and T. Poggio. Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience*, 4(3):179–192, 2003. doi: 10.1038/nrn1057.

Google DeepMind. Gemini 2.5: Our most intelligent models are getting even better. `https://blog.google/technology/google-deepmind/google-gemini-updates-io-2025/`, 2025.

M.-H. Grosbras, S. Beaton, and S. B. Eickhoff. Brain regions involved in human movement perception: A quantitative voxel-based meta-analysis. *Human Brain Mapping*, 33(2):431–454, 2012. doi: 10.1002/hbm.21222.

B. Hommel, J. Müsseler, G. Aschersleben, and W. Prinz. The theory of event coding (tec): A framework for perception and action planning. *Behavioral and Brain Sciences*, 24(5):849–878, 2001. doi: 10.1017/S0140525X01000103.

W. Hong, W. Wang, M. Ding, and et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024.

W. Hong, Y. Cheng, Z. Yang, W. Wang, L. Wang, X. Gu, S. Huang, Y. Dong, and J. Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211, 1973. doi: 10.3758/BF03212378.

A. Kadambi, Q. Xie, and H. Lu. Individual differences and motor planning influence self-recognition of actions. *PLOS ONE*, 19(7):e0303820, 2024. doi: 10.1371/journal.pone.0303820.

K. Li, Y. Wang, Y. He, and et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22195–22206, 2024. arXiv:2311.17005.

X. Liu, T. Zhang, Y. Gu, I. L. Iong, Y. Xu, X. Song, and J. Tang. Visualagentbench: Towards large multimodal models as visual foundation agents. *arXiv preprint arXiv:2408.06327*, 2024.

F. Loula, S. Prasad, K. Harber, and M. Shiffrar. Recognizing people from their movement. *Journal of Experimental Psychology: Human Perception and Performance*, 31(1):210–220, 2005. doi: 10.1037/0096-1523.31.1.210.

V. Manera, B. Schouten, C. Becchio, and et al. Inferring intentions from biological motion: A stimulus set of point-light communicative interactions. *Behavior Research Methods*, 42(1):168–178, 2010. doi: 10.3758/BRM.42.1.168.

M. Mon-Williams, L. B. Smith, R. Henson, and et al. Embodied large language models enable robots to complete complex tasks in unpredictable environments. *Nature Machine Intelligence*, 2025. doi: 10.1038/s42256-025-01005-x.

OpenAI. Gpt-4v(ision) system card. `https://openai.com/index/gpt-4v-system-card/`, 2023.

W. Prinz. Perception and action planning. *European Journal of Cognitive Psychology*, 9(2):129–154, 1997. doi: 10.1080/713752551.

V. M. Reid, K. Dunn, R. J. Young, J. Amu, T. Donovan, and N. Reissland. The human fetus preferentially engages with face-like visual stimuli. *Current Biology*, 27(12):1825–1828, 2017. doi: 10.1016/j.cub.2017.05.044.

G. Rizzolatti and C. Sinigaglia. The mirror mechanism: a basic principle of brain function. *Nature Reviews Neuroscience*, 17(12):757–765, 2016. doi: 10.1038/nrn.2016.135.

F. Simion, L. Regolin, and H. Bulf. A predisposition for biological motion in the newborn baby. *Proceedings of the National Academy of Sciences*, 105(2):809–813, 2008. doi: 10.1073/pnas.0707021105.

J. J. A. van Boxtel and H. Lu. A biological motion toolbox for reading, displaying, and manipulating motion capture data in research settings. *Journal of Vision*, 13(12):7, 2013. doi: 10.1167/13.12.7.

P. Wang, S. Bai, S. Tan, and et al. Qwen2-vl: Enhancing vision–language models' perception at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

W. Wang, Z. He, W. Hong, Y. Cheng, X. Zhang, J. Qi, and J. Tang. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024b.

H. Wu, D. Li, B. Chen, and J. Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024.

H. Xia, Z. Yang, Y. Wang, and et al. Sportqa: A benchmark for sports understanding in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 5061–5081. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.283.

L. Xu, Y. Zhao, D. Zhou, Z. Lin, S. K. Ng, and J. Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024.

R. Yang, H. Chen, J. Zhang, M. Zhao, C. Qian, K. Wang, Q. Wang, T. V. Koripella, M. Movahedi, M. Li, H. Ji, H. Zhang, and T. Zhang. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025.

S. Ye, H. Qi, A. Mathis, and M. W. Mathis. Llavaction: evaluating and training multi-modal large language models for action recognition. *arXiv preprint arXiv:2503.18712*, 2025.

J. Zhou, Y. Shu, B. Zhao, B. Wu, Z. Liang, S. Xiao, M. Qin, X. Yang, Y. Xiong, B. Zhang, T. Huang, and Z. Liu. Mlvu: Benchmarking multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024.

S. Zhou, A. Vilesov, X. He, Z. Wan, S. Zhang, A. Nagachandra, D. Chang, D. Chen, X. E. Wang, and A. Kadambi. Vlm4d: Towards spatiotemporal awareness in vision language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. arXiv:2508.02095.

A Appendix

A.1 Individual Action Analysis

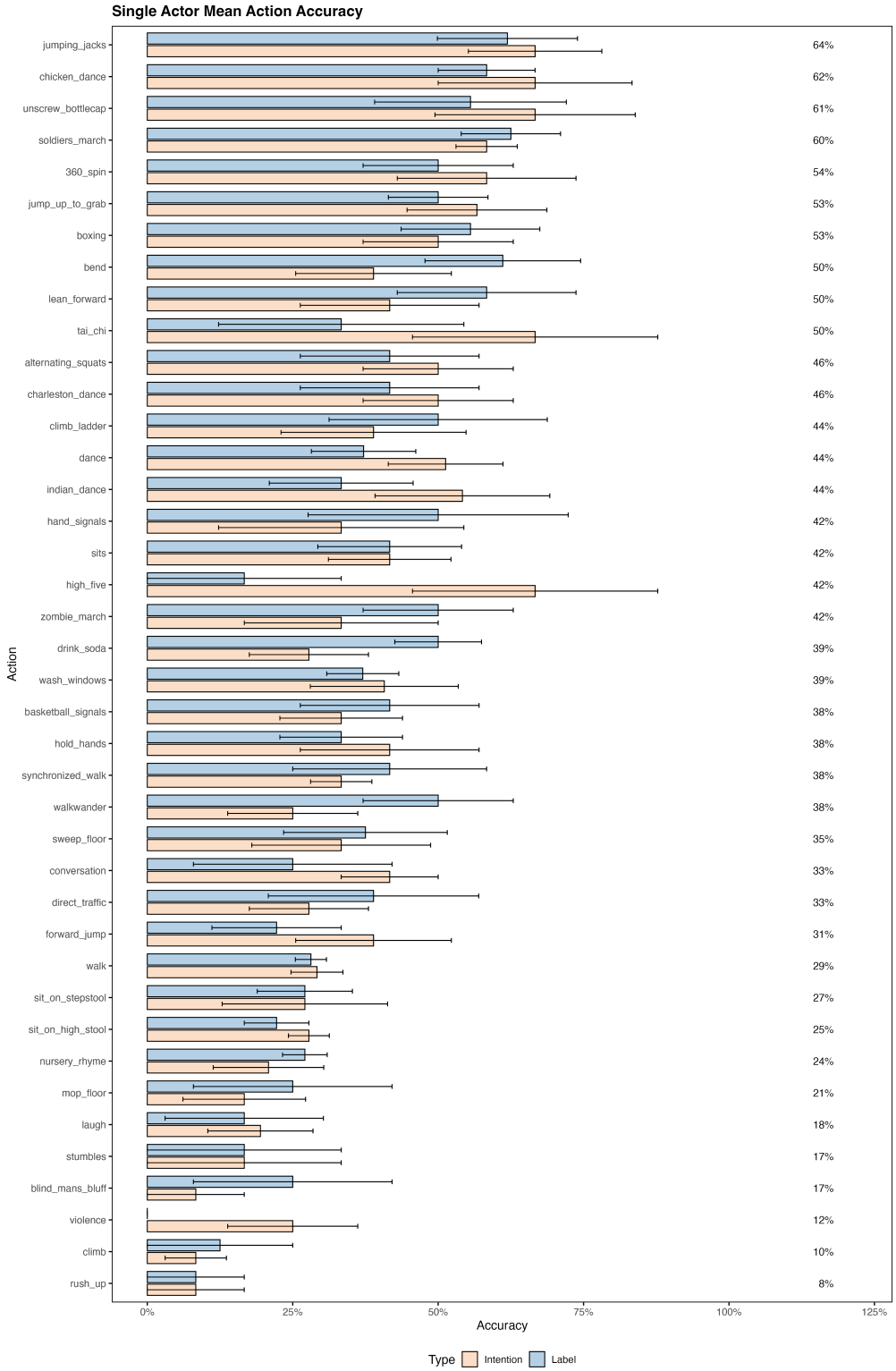

Figure 7: Single actor means for each action for action labeling and action intentions across all models. Mean percentages reflect average accuracy across label and intentions for each action. Actions are arranged from highest (top) to lowest (bottom) performing models.

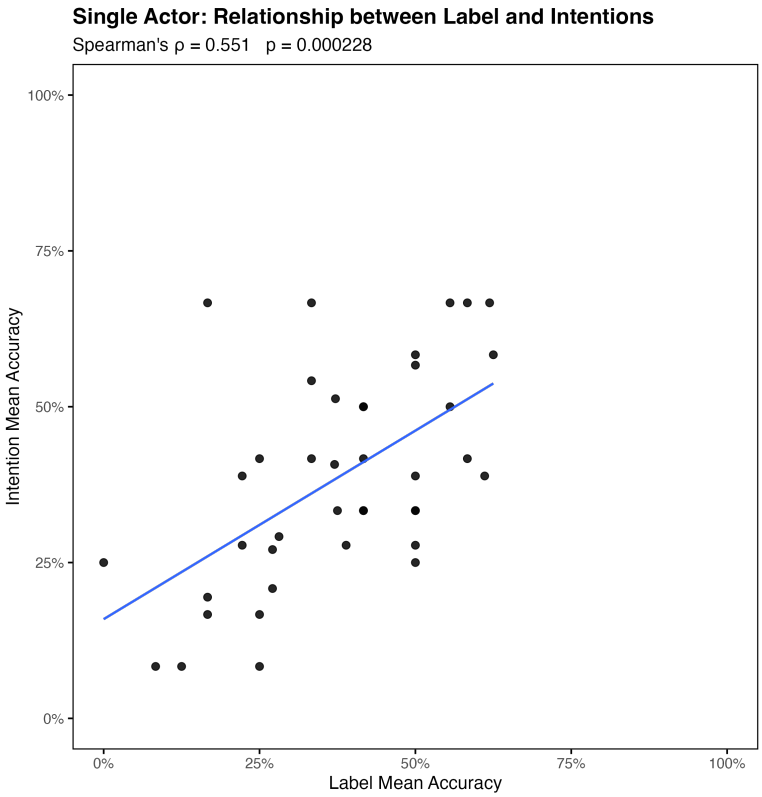

Figure 8: Plot depicts strong relationship between action labeling and action intention inference for individual single-actor actions across models (p<.0001)

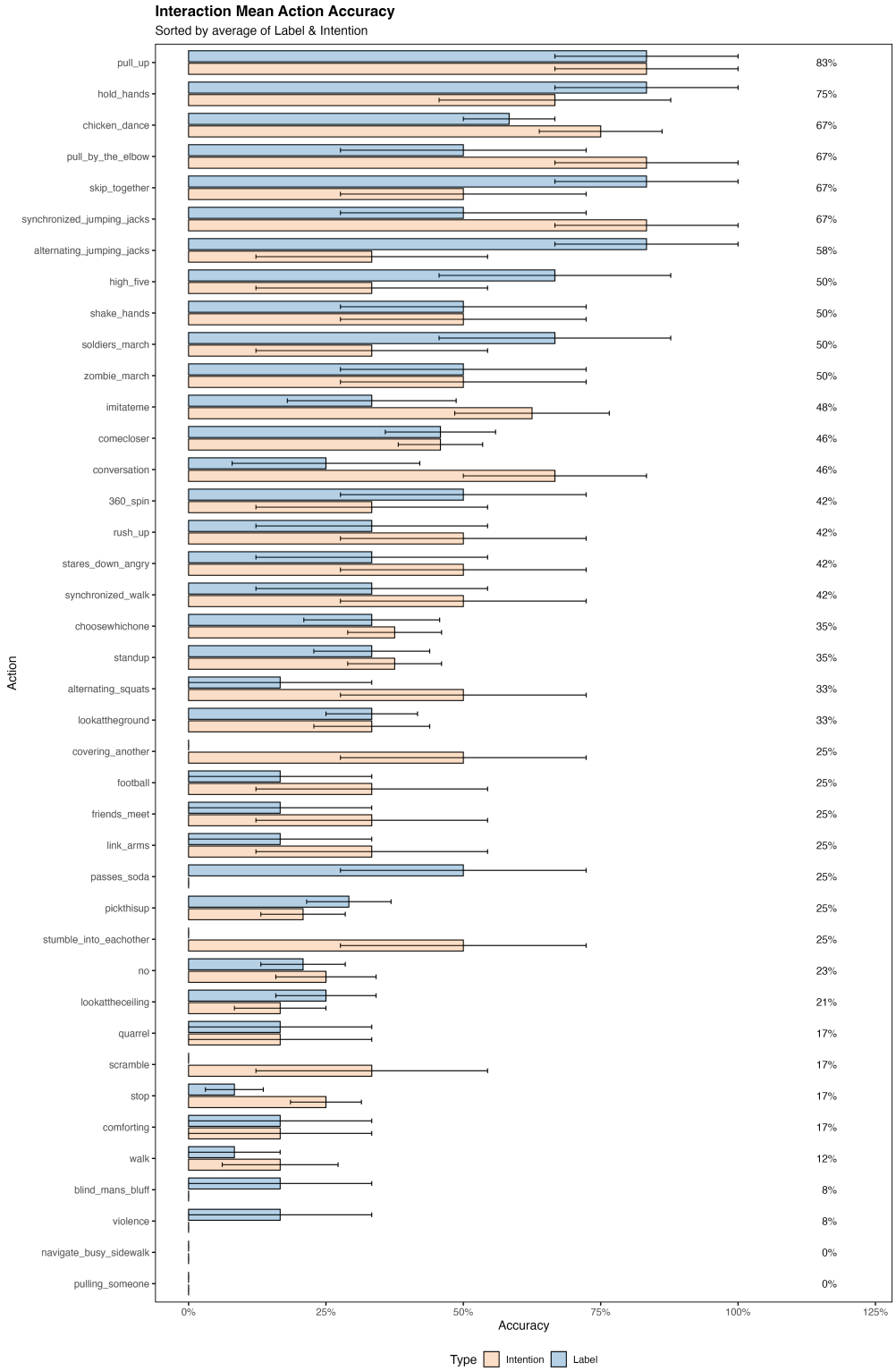

Figure 9: Social interaction means for each action for action labeling and action intentions across all models. Mean percentages reflect average accuracy across label and intentions for each action. Actions are arranged from highest (top) to lowest (bottom) performing models.

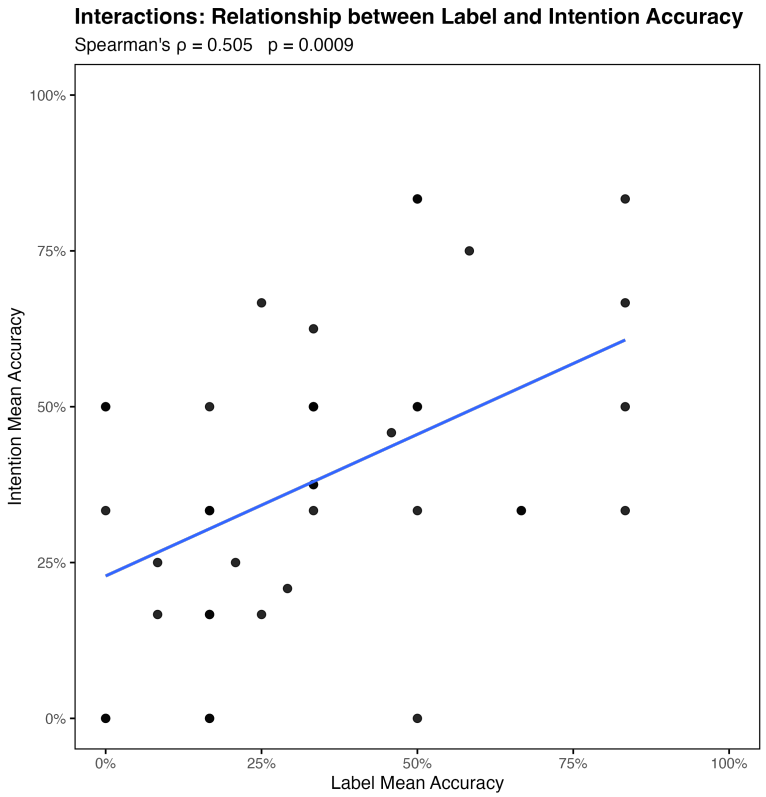

Figure 10: Plot depicts strong relationship for social interactions between action labeling and action intention inference for individual actions across models (p<.0001)

A.2 Human Data (N=2)

To measure ground truth comparisons, we also collected human data on a small participant sample on a subset of randomly selected actions. Below we include actual tabular results for each action; from left-right: action number, human written description, 3AFC choices, selected action by the participant, and accuracy (1 = correct, 0 = incorrect). Mean human accuracy ranged ∼93% for both social interactions and single actors.

A.2.1 Social Interactions Human Data

Table 5: Social Interactions (Participant 1) - Mean Accuracy = 93.33%

Action	Description	Choices	Final Choice	Correct
1	two people greeting each other, waving hands	comecloser, laugh, rush_up	comecloser	1
2	one person pick things up from the ground and give them to the other person	lookattheground, choosewhichone, jumping_jacks	choosewhichone	1
3	two people fighting	sits, sit_on_stepstool, conversation	conversation	1
4	two people using their backs to carry something	covering_another, go_out_of_the_way, mop_floor	covering_another	1
5	two people running around cheering	blind_mans_bluff, walk, football	football	1

Continued on next page

Table 5: Social Interactions (Participant 1) - Mean Accuracy = 93.33% (continued)

Action	Description	Choices	Final Choice	Correct
6	two people meeting up for lunch	friends_meet, rush_up, nursery_rhyme	friends_meet	1
7	two people giving a highfive	high_five, conversation, stumbles	high_five	1
8	two people walking together and holding hands	swing, blind_mans_bluff, hold_hands	swing	0
9	one person squatting down and the other one mimicking the movements	turnover, bend, imitateme	imitateme	1
10	two people walking together and holding hands	sneeze, link_arms, movethisdown	link_arms	1
11	one person saying hi and the other turning around	alternating_squats, lookattheceiling, hold_hands	lookattheceiling	1
12	one person pointing at the ground and both squatting down to look	sit_on_high_stool, synchronized_walk, lookattheground	lookattheground	1
13	two people walking together and pushing each other	blind_mans_bluff, jumping_jacks, navigate_busy_sidewalk	navigate_busy_sidewalk	1
14	one person trying to approach and the other one rejecting	no, rush_up, conversation	no	1
15	two people sharing a bottle of water	360_spin, passes_soda, stretch	passes_soda	1
16	one person picking something up and the other one offering to help	lookattheceiling, sneeze, pickthisup	pickthisup	1
17	one person dragging the other one away	swing, blind_mans_bluff, pull_by_the_elbow	pull_by_the_elbow	1
18	two people fighting for something in their hands	stumbles, pulling_someone, sit_on_stepstool	pulling_someone	1
19	two people boxing	stand_up, imitate, quarrel	quarrel	1
20	two people fighting for a chair	scramble, nursery_rhyme, turnover	turnover	0
21	two people spinning by holding their arms	skip_together, climb, no	skip_together	1
22	one person asking the other one to stand up	standup, stop, basketball_signals	standup	1
23	one person trying to grab something from the other	360_spin, stares_down_angry, stretch	stares_down_angry	1
24	one person trying to approach and the other one rejecting	stop, blind_mans_bluff, wash_windows	stop	1
25	two people walking together	stumble_into_eachother, stretch, punch	stumble_into_eachother	1
26	two people dancing together	chicken_dance, sweep_floor, stand_up	chicken_dance	1
27	one person pulling the other person's ear	comforting, nursery_rhyme, sit_down	comforting	1
28	two people walking together and chatting	forward_jump, conversation, lateralsteps	conversation	1

Continued on next page

Table 5: Social Interactions (Participant 1) - Mean Accuracy = 93.33% (continued)

Action	Description	Choices	Final Choice	Correct
29	two people jumping and exercising	bend, alternating_jumping_jacks, no	alternating_jumping_jacks	1
30	two people playing the blind game by covering one person's eyes	hold_hands, jump, blind_mans_bluff	blind_mans_bluff	1

Table 6: Social Interactions (Participant 2) - Mean Accuracy = 100%

Action	Description	Choices	Final Choice	Correct
1	One person motioning the other to come towards them and then the other person walks towards them	comecloser, laugh, rush_up	comecloser	1
2	One person picking up a couple of things to look at them, and then they tell the other person to look and the other person grabs the items	lookattheground, choosewhichone, jumping_jacks	choosewhichone	1
3	One person stacking their hands on top of another person's hands	sits, sit_on_stepstool, conversation	conversation	1
4	Two people standing next to each other, maybe guarding something or preparing for a fight	covering_another, go_out_of_the_way, mop_floor	covering_another	1
5	Two people celebrating something exciting, one person shoots a basketball while the other person guards and then they both cheer	blind_mans_bluff, walk, football	football	1
6	One person comes over to sit next to someone else and daps them up before having a conversation	friends_meet, rush_up, nursery_rhyme	friends_meet	1
7	Two friends walking by each other and giving each other a high five	high_five, conversation, stumbles	high_five	1
8	Two friends or partners holding hands and swinging them while they walk	swing, blind_mans_bluff, hold_hands	hold_hands	1
9	One person showing the other person how to squat and then the other person shows what they learned and squats	turnover, bend, imitateme	imitateme	1
10	Two people holding hands	sneeze, link_arms, movethisdown	link_arms	1
11	One person waving and then the other person comes closer	alternating_squats, lookattheceiling, hold_hands	lookattheceiling	1

Continued on next page

Table 6: Social Interactions (Participant 2) - Mean Accuracy = 100% (continued)

Action	Description	Choices	Final Choice	Correct
12	One person points at something on the ground and then both of them bend down to look at the ground	sit_on_high_stool, synchronized_walk, lookattheground	lookattheground	1
13	Two people holding hands and dancing	blind_mans_bluff, jumping_jacks, navigate_busy_sidewalk	navigate_busy_sidewalk	1
14	Someone bending over to pick something up and the other person tells them not to and they listen	no, rush_up, conversation	no	1
15	One person is drinking something from a can and then gives to the person next to them to drink from it	360_spin, passes_soda, stretch	passes_soda	1
16	A person points to something on the ground next to another person and that person picks up the item off the ground	lookattheceiling, sneeze, pickthisup	pickthisup	1
17	One person grabbing another person and trying to drag them away but they resist	swing, blind_mans_bluff, pull_by_the_elbow	pull_by_the_elbow	1
18	One person grabbing another person by the hand and dragging them away	stumbles, pulling_someone, sit_on_stepstool	pulling_someone	1
19	Two people are pumping each other up and listening to the crowd cheer their name as they celebrate	stand_up, imitate, quarrel	quarrel	1
20	One person sits down on a chair that someone else was trying to sit on	scramble, nursery_rhyme, turnover	scramble	1
21	Two people swinging around each other with their arms linked	skip_together, climb, no	skip_together	1
22	Someone waving at another person who is sitting down and then that person gets up	standup, stop, basketball_signals	standup	1
23	Two people have their hands in their laps and then one person gets up	360_spin, stares_down_angry, stretch	stares_down_angry	1
24	One person approaches another person but that person tells them to stop	stop, blind_mans_bluff, wash_windows	stop	1
25	One tall person and a shorter person walking together	stumble_into_eachother, stretch, punch	stumble_into_eachother	1
26	Two people doing the chicken dance	chicken_dance, sweep_floor, stand_up	chicken_dance	1
27	One person coming up to another person to rub their shoulders or back	comforting, nursery_rhyme, sit_down	comforting	1
28	Two people walking and having a conversation	forward_jump, conversation, lateralsteps	conversation	1

Continued on next page

Table 6: Social Interactions (Participant 2) - Mean Accuracy = 100% (continued)

Action	Description	Choices	Final Choice	Correct
29	Two people doing jumping jacks but then one person stops and looks at the other	bend, alternating_jumping_jacks, no	alternating_jumping_jacks	1
30	One person trying to touch another person probably with their eyes closed	hold_hands, jump, blind_mans_bluff	blind_mans_bluff	1

A.2.2 SINGLE ACTOR HUMAN DATA

Table 7: Single Actor (Participant 1) - Mean Accuracy = 90.00%

Action	Description	Choices	Selection	Correct
1	a person running forward and backward	charleston_dance, alternating_squats, run	run	0
2	a person boxing	forward_jump, sit_on_high_stool, boxing	boxing	1
3	a person moving their upper body around	walkaway, boxing, nursery_rhyme	nursery_rhyme	1
4	a person moving around like they are drunk or laughing very hard	laugh, 360_spin, synchronized_walk	laugh	1
5	a person doing jumping jacks	swing, jumping_jacks, basketball_signals	jumping_jacks	1
6	a person waving while walking by	high_five, pickthisup, direct_traffic	high_five	1
7	a person jumping forward	squat_down, drink_soda, forward_jump	forward_jump	1
8	a person drinking water	climb, turnover, drink_soda	drink_soda	1
9	a person waving and asking someone coming over	basketball_signals, zombie_march, direct_traffic	direct_traffic	1
10	a person dancing around	blind_mans_bluff, walk, chicken_dance	chicken_dance	1
11	a person using their hands to make dance moves	synchronized_walk, basketball_signals, stretch	basketball_signals	1
12	a person squat and do defense moves	walk, violence, drink	violence	1
13	a person doing taichi	sit_down, tai_chi, movethisdown	tai_chi	1
14	a person walking	wash_windows, synchronized_walk, hang	synchronized_walk	1
15	a person mopping or sweeping the floor	jog, stretch, sweep_floor	sweep_floor	1
16	a person walking with rhythm	sits, direct_traffic, soldiers_march	soldiers_march	1
17	a person standing up from a chair	go_out_of_the_way, sits, wash_windows	go_out_of_the_way	0
18	a person trying hard to move or take something	movethisdown, stop, rush_up	movethisdown	0

Continued on next page

Table 7: Single Actor (Participant 1) - Mean Accuracy = 90.00% (continued)

Action	Description	Choices	Selection	Correct
19	a person sitting down and standing up and sitting down again	lateralsteps, stumbles, sit_on_high_stool	sit_on_high_stool	1
20	a person spinning with a pole	360_spin, pickthisup, direct_traffic	360_spin	1
21	a person walking around blind	blind_mans_bluff, stop, go_out_of_the_way	blind_mans_bluff	1
22	a person cheerleading	rush_up, wash_windows, lateralsteps	wash_windows	1
23	a person walking for balance tests	mop_floor, zombie_march, lateralsteps	zombie_march	1
24	a person walking around back and forth	walkwander, forward_jump, stumbles	walkwander	1
25	a person walking with rhythm	jump, stretch, walk	walk	1
26	a person doing defense moves	jumping_jacks, jump, violence	violence	1
27	a person walking up and down from ladders	imitate, climb_ladder, laugh	climb_ladder	1
28	a person picking something up from the ground	bend, wash_windows, sit_down	bend	1
29	a person sitting down and talking with their hands moving around	lookattheceiling, sit_down, conversation	conversation	1
30	a person walking up and down from stairs	stop, climb, stand_up	climb	1

Table 8: Single Actor (Participant 2) - Mean Accuracy = 96.67%

Action	Description	Choices	Final Choice	Correct
1	Dots in the shape of a person dancing	charleston_dance, alternating_squats, run	charleston_dance	1
2	Dots in the shape of a person doing different boxing moves, jumping up and down, punching, dodging and ducking and sometimes punching behind them	forward_jump, sit_on_high_stool, boxing	boxing	1
3	Dots in the shape of a person standing not doing much, but then it looks like he's carrying some broom or stick and cleaning something up high and then bends over	walkaway, boxing, nursery_rhyme	nursery_rhyme	1
4	Dots in the shape of a person pacing back and forth, sometimes with its hands in the air, sometimes hands on their knees, might be laughing or nervous	laugh, 360_spin, synchronized_walk	laugh	1
5	Dots in the shape of a person doing jumping jacks	swing, jumping_jacks, basketball_signals	jumping_jacks	1
6	Dots in the shape of a person waving their hands and moving their feet	high_five, pickthisup, direct_traffic	direct_traffic	0

Continued on next page

Table 8: Single Actor (Participant 2) - Mean Accuracy = 96.67% (continued)

Action	Description	Choices	Final Choice	Correct
7	Dots in the shape of a person putting their hands out and then jumping	squat_down, drink_soda, forward_jump	forward_jump	1
8	Dots in the shape of a person drinking a drink of some kind	climb, turnover, drink_soda	drink_soda	1
9	Dots in the shape of a person moving his hands and body to encourage people to move in certain directions, probably trying to direct traffic	basketball_signals, zombie_march, direct_traffic	direct_traffic	1
10	Dots in the shape of a person dancing: clapping, moving his hands behind his back, and then shaking his hips like the chicken dance	blind_mans_bluff, walk, chicken_dance	chicken_dance	1
11	Dots in the shape of a person doing different hand signals, like rolling his hands, pointing, and putting his hands in front of him	synchronized_walk, basketball_signals, stretch	basketball_signals	1
12	Dots in the shape of a person in a squat then gets up and has something in his hands that he might throw	walk, violence, drink	violence	1
13	Dots in the shape of a person slowly moving his hands and body back and forth to a rhythm, like tai chi	sit_down, tai_chi, movethisdown	tai_chi	1
14	Dots in the shape of a person walking	wash_windows, synchronized_walk, hang	synchronized_walk	1
15	Dots in the shape of a person sweeping the floor with a broom	jog, stretch, sweep_floor	sweep_floor	1
16	Dots in the shape of a person kicking his feet while he walks	sits, direct_traffic, soldiers_march	soldiers_march	1
17	Dots in the shape of a person sitting and then standing and pacing forward	go_out_of_the_way, sits, wash_windows	go_out_of_the_way	1
18	Dots in the shape of a person bending over to grab something maybe or put his hands on the floor	movethisdown, stop, rush_up	movethisdown	1
19	Dots in the shape of a person sitting and then standing and then sitting back down	lateralsteps, stumbles, sit_on_high_stool	sit_on_high_stool	1
20	Dots in the shape of a person swinging around a pole	360_spin, pickthisup, direct_traffic	360_spin	1
21	Dots in the shape of a person walking around with his arms in front of him like a zombie	blind_mans_bluff, stop, go_out_of_the_way	blind_mans_bluff	1
22	Dots in the shape of a person washing windows probably or doing wax on, wax off	rush_up, wash_windows, lateralsteps	wash_windows	1

Continued on next page

Table 8: Single Actor (Participant 2) - Mean Accuracy = 96.67% (continued)

Action	Description	Choices	Final Choice	Correct
23	Dots in the shape of a person walking around lazily like a zombie	mop_floor, zombie_march, lateralsteps	zombie_march	1
24	Dots in the shape of a person walking then turning around, but looks like they're still walking forward	walkwander, forward_jump, stumbles	walkwander	1
25	Dots in the shape of a person walking with a pep in their step	jump, stretch, walk	walk	1
26	Dots in the shape of a person sitting and then getting up to fight with his hands raised	jumping_jacks, jump, violence	violence	1
27	Dots in the shape of a person climbing up and down something like a ladder	imitate, climb_ladder, laugh	climb_ladder	1
28	Dots in the shape of a person bending down to pick something up	bend, wash_windows, sit_down	bend	1
29	Dots in the shape of a person sitting down and explaining something with their hands	lookattheceiling, sit_down, conversation	conversation	1
30	Dots in the shape of a person stepping up probably up some stairs and then going back down those stairs	stop, climb, stand_up	climb	1

A.3 Actual Description Matching Scorer

For each row in the CSV, decide whether the action in column 2 (`true_label`) is the same as the action that the response in column 6 ultimately describes for the corresponding `video_name`.

General principles Read the entire response for meaning, not just keywords. The `true_label` must be the action the model's overall description supports, not necessarily the last one it mentions.

Negation-aware If the true label is explicitly negated (e.g., "not X", "unlikely to be X", "rather than X"), this counts against it. However, if other parts of the description strongly support the true label, it can still match.

Positive cue matching (slightly expanded) Treat the following as positive cues when they appear near the label (either order): *is, shows, depicts, represents, looks like, best described as, most consistent with, clearly, demonstrates, doing*. Also count: *best match(es), matches, consistent with, resembles, most likely*. "Near" means the cue appears in the same sentence or within ∼80 characters of the label name (including across line breaks/bullets).

Descriptive fallback (unchanged in spirit, with broader coverage) If the true label name is not mentioned positively, check whether the description includes defining features or keywords characteristic of that action (e.g., "flapping/wing-like arm motions" for *chicken dance*). Use label-specific descriptors when available (examples you may see in responses):

- **360_spin:** rotate, spin, full rotation, turn in place, "360"
- **jump_up_to_grab:** jump up, reach up, grab overhead, extend arms upward
- **mop_floor:** mop/mopping, push-pull with handle, cleaning the floor
- **lean_forward:** bend/lean forward, torso incline

- **sit_on_high_stool / sit_on_step_stool:** high/elevated stool, step stool, perch on stool
- **nursery_rhyme:** nursery rhyme, sing-song, rhythmic hand/clap
- **laugh:** laugh/laughing/chuckle

This fallback only applies if competing labels are not positively supported or are explicitly negated.

Match decision Mark `label_match = 1` if any of these hold:

1. The true label has more positive than negative evidence.
2. The true label has at least one positive cue, and all other labels are either not positively supported or are negated.
3. The descriptive fallback applies (the text clearly matches the true label's characteristic motion and other labels are unsupported).

Otherwise, mark `label_match = 0`.

Error rows If the response contains "ERROR:", ignore that row when calculating the percentage match.

Synonym handling Normalize close synonyms and wording changes (e.g., *pick up = pickthisup; embrace = hug; stand_up = standup; sit_down/sits = sitdown; walking = walk; run/running = jog*).