SOMA: A BENCHMARK FOR FINE-GRAINED EGOCENTRIC BEHAVIOR AND BODY-MOTION UNDERSTANDING

Anonymous authors
Paper under double-blind review

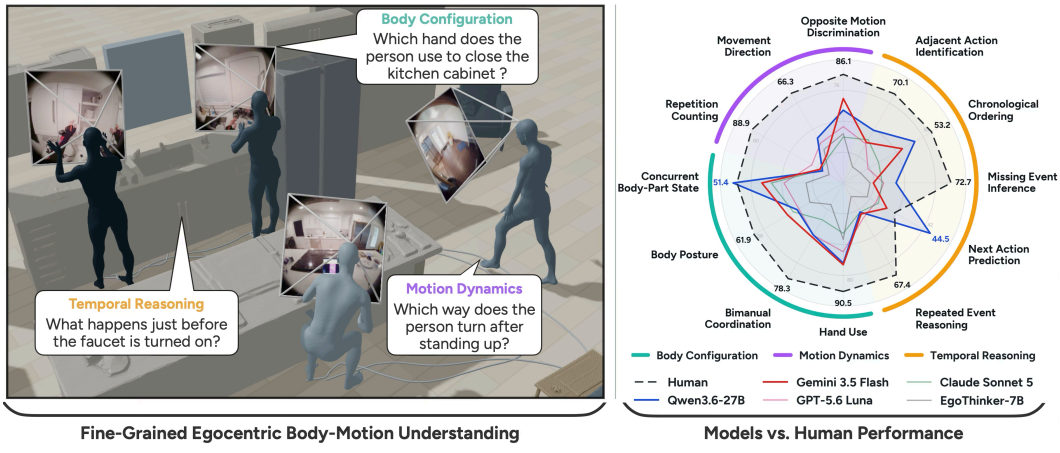


Figure 1: **Overview of SOMA.** Left: Representative questions from the three capability axes, **Body Configuration**, **Motion Dynamics**, and **Temporal Reasoning** grounded in egocentric video and body motion. Right: Performance of representative MLLMs across the twelve SOMA tasks compared with human performance (5 annotators), highlighting a human–model gap. Note: performance normalized for visualization purposes.

ABSTRACT

Perceiving human body motion is critical for multimodal systems that understand and assist people. Yet current evaluations of Multimodal Large Language Models (MLLMs) largely focus on objects, scenes, and coarse actions, leaving their ability to reason about the body itself poorly assessed. This is especially challenging from the egocentric perspective, in which the body is only partially visible and motion must be inferred from occluded and temporally distributed cues. We introduce SOMA, a benchmark for fine-grained body-motion understanding from egocentric video, comprising 1,632 human-verified questions across three capability axes: body configuration, motion dynamics, and temporal reasoning. We evaluate 37 leading closed-source, open-source, and specialized egocentric MLLMs, complemented by an extensive human evaluation. We find that humans also find this task difficult from video alone (68.5% accuracy), yet clearly outperform the strongest MLLM (Qwen3.6-27B), which reaches 50.0% zero-shot accuracy. Our error analysis shows that MLLMs often misread the wearer's own body, misjudge events in time, and are prone to biases such as favoring the right hand or undercounting repetitions. To narrow the gap to human performance and support further progress, we introduce a training dataset of aligned video, body motion, and language supervision, and a multimodal baseline that jointly learns from these modalities. By identifying common failure modes and missing capabilities in current MLLMs, we hope SOMA will inspire future research on body-motion understanding and multimodal reasoning. We will release the full benchmark, training data, and code.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: **Benchmark Comparison.** *#Cat.*: number of categories; *Input Modality*: RGB video, 3D body pose (MoVid provides either RGB or pose, hence $\veebar$); *Ego*: egocentric perspective; *Train*: provides a training split; *Body*: targets the body, body parts, or posture; *Mot.*: fine-grained motion dynamics beyond coarse activity recognition; *Temp.*: temporal reasoning.

	Benchmark	#Video Clips	#QA	#Cat.	Attributes			QA Scope		
					Modality	Ego	Train	Body	Mot.	Temp.
Human-centric	BABEL-QA (Endo et al., 2023)	1,109	2,577	3		✗	✓	✓	✓	✓
	MoVid (Chen et al., 2025)	1,350	1,350	5	$\veebar$	✗	✓	✓	✓	✓
	MotionBench (Hong et al., 2025)	5,385	8,052	6		✗	✗	✓	✓	✓
	ExAct (Yi et al., 2025)	3,521	3,521	2		✗	✗	✓	✗	✓
Egocentric Video QA	EgoTaskQA (Jia et al., 2022)	2,000	40,000	4		✓	✓	✗	✗	✓
	EgoSchema (Mangalam et al., 2023)	5,063	5,063	1		✓	✗	✗	✗	✓
	EgoThink (Cheng et al., 2024)	700	700	12		✓	✗	✗	✗	✗
	EgoGazeVQA (Peng et al., 2025)	913	1,757	11		✓	✗	✗	✗	✓
	EgoTempo (Plizzari et al., 2025)	365	500	10		✓	✗	✗	✗	✓
	LEMONADE (Bonnetto et al., 2025)	36,521	36,521	31		✓	✗	✗	✓	✓
	SOMA (Ours)	**1,611**	**1,632**	**12**	$\wedge$	✓	✓	✓	✓	✓

1 INTRODUCTION

Understanding human motion is fundamental for science, and for AI systems that interact with, assist, or learn from humans (Yin et al., 2025; Li et al., 2025a; Bonnetto et al., 2026). Recently, Multimodal Large Language Models (MLLMs) have demonstrated strong abilities to perceive and reason about human activity from video, yet largely from a third-person perspective (Li et al., 2024a; Qi et al., 2026). Egocentric video, captured from a first-person perspective using wearable devices, offers a fundamentally different viewpoint, central to augmented reality interfaces, wearable assistants, and embodied agents (Yang et al., 2025b; Li et al., 2025b; Bonnetto et al., 2026). However, this viewpoint introduces distinct challenges (Grauman et al., 2022; He et al., 2025). First, the actor's body and actions are often only partially visible to the camera, despite being essential for understanding the wearer's intention and activity. Second, fine-grained motion dynamics, such as repetition counts or movement direction, are often hard to perceive from the egocentric view. Finally, motion blur and rapid viewpoint changes are common in egocentric video, requiring MLLMs to integrate ego-motion cues across frames. Fine-grained body-motion understanding from an egocentric view is therefore a challenging task. Yet existing egocentric video benchmarks that drive progress in this field (Mangalam et al., 2023; Jia et al., 2022; Bonnetto et al., 2025) target other dimensions of video understanding, such as long-form video comprehension or domain-specific, goal-oriented task reasoning (Table 1). None directly measures whether a model can infer the moving body itself: its configuration at a moment and the way its actions unfold over time. As a result, it remains unclear whether progress on egocentric video reflects understanding of human motion or recognition of the objects and scenes in which motion takes place.

To fill this gap, we introduce SOMA (Greek for *body*), a multimodal benchmark for fine-grained understanding of human body motion and behavior from egocentric video. SOMA evaluates MLLMs on 12 tasks organized into three capability axes that mirror the aforementioned challenges: *Body Configuration*, *Motion Dynamics*, and *Temporal Reasoning*, which probe how the body is configured, how it moves, and how its actions unfold in time. As a source for egocentric videos with annotations, we selected two comprehensive corpora, Nymeria (Ma et al., 2024) and Ego-Exo4D (Grauman et al., 2024). Based on a tailored pipeline including human verification, we derived a final benchmark of 1,632 multiple-choice questions. We additionally release SOMA-Train, a multimodal training dataset of aligned video, body motion, and language supervision, a combination absent from existing egocentric video benchmarks (Table 1).

We evaluated 37 leading MLLMs on SOMA, spanning closed-source, open-source, and specialized egocentric systems. Furthermore, five human evaluators independently answered all 1,632 questions, requiring over 450 hours of manual labor. The best-performing model, Qwen3.6-27B, reaches only 50.0% zero-shot accuracy, well below the 68.5% mean accuracy achieved by humans. This gap persists even with denser temporal sampling and explicit test-time reasoning, which yield only limited, model-dependent gains. Error analysis shows that the most common failures stem from misreading the wearer's own body, along with difficulty localizing events in time. Motivated by this, we test whether an explicit body-motion signal helps: fine-tuning on SOMA-Train with aligned video and motion yields modest gains, suggesting that body motion can provide information complementary to video, paving the way toward MLLMs that better perceive human motion.

2 Related work

Human Motion Understanding. Understanding human motion is a long-standing problem (Aggarwal & Cai, 1999), studied through various approaches. One line of work focuses on perception: methods reconstruct 3D body pose from video (Goel et al., 2023; Wu et al., 2026; Qi et al., 2024) and operate on skeleton inputs to recognize and segment human actions (Duan et al., 2022; Jiang et al., 2023; Ren et al., 2024; Stoffl et al., 2024). A complementary line of work studies motion generation, modeling the distribution of plausible motions to synthesize new ones from text (Guo et al., 2024; Zeng et al., 2025). In contemporary work, MLLMs treat motion as a modality alongside text, unifying motion comprehension and generation within a single model (Wu et al., 2025a; Chen et al., 2025; Li et al., 2025c). Yet, models that explicitly combine video and body motion remain uncommon, with few publicly available implementations and model weights. Furthermore, these MLLMs focus on the exocentric setting, where occlusions are rare and motion-capture data is more straightforward. To address this, SOMA evaluates how well current MLLMs perceive body motion from first-person video, complementing the body-motion benchmarks that operate on third-person or motion-capture input alone (Table 1).

MLLMs for Video Understanding. Building on the success of large language models (Achiam et al., 2023; Yang et al., 2025a; Cai et al., 2024), MLLMs extend text-only systems to additional modalities (Han et al., 2024; Girdhar et al., 2022; Comanici et al., 2025). Among these, video MLLMs (Li et al., 2024a; Zhang et al., 2024; Wang et al., 2024; Li et al., 2024b; Qi et al., 2026) address the understanding and reasoning of video content. Motivated by applications in AR, robotics, and assistive systems, a separate line of work adapts MLLMs to the egocentric setting. For example, EgoThinker (Pei et al., 2025) enhances egocentric reasoning through chain-of-thought rationales, and EgoGPT (Yang et al., 2025b) fine-tunes a video MLLM for egocentric captioning and QA. However, existing egocentric MLLMs remain optimized for activity- and scene-level understanding, leaving the question of fine-grained body-motion perception largely unexplored. SOMA is designed to evaluate this capability explicitly, exposing the shortcomings of current models, while also providing a training split to further advance their development.

Egocentric VQA Benchmarks. Egocentric video question-answering is expanding rapidly across several directions. Early benchmarks established the format for first-person video QA, with EgoTaskQA (Jia et al., 2022) probing goal-oriented task understanding and EgoSchema (Mangalam et al., 2023) targeting long-form video. More recent benchmarks explore specific aspects of egocentric understanding, including first-person perspective thinking (Cheng et al., 2024), cross-domain and domain-specific understanding (Bonnetto et al., 2025; Li et al., 2026), and temporal reasoning (Plizzari et al., 2025). Despite this breadth, none of these benchmarks target the perception of the moving human body. Specifically, tasks such as activity recognition, task planning, and temporal grounding remain focused on objects and the surrounding environment, rather than the body itself. We fill this gap by introducing a body-centric question taxonomy grounded in fine-grained body-motion annotations, paired with video and 3D body pose, which probes competencies that existing benchmarks do not directly evaluate (Table 1).

3 The SOMA Benchmark

We introduce SOMA, a multimodal benchmark for fine-grained understanding of human body motion from egocentric videos. We describe the task suite (Sec. 3.1), the data curation pipeline (Sec. 3.2), dataset statistics (Sec. 3.3), and our training dataset SOMA-Train (Sec. 3.4).

3.1 Task Suite

We organize the benchmark around three axes, each targeting one of the aforementioned challenges: *Body Configuration*, *Motion Dynamics*, and *Temporal Reasoning*. Together, they comprise 12 tasks that span distinct competencies in understanding a moving body from the first-person view (Figure 2).

Body Configuration. These tasks probe perception of the body and its parts during an activity. *Hand Use* identifies which hand performs a given action. *Bimanual Coordination* characterizes how the two hands act together, including their role assignment and consistency across repeated instances. *Body Posture* identifies the person's overall posture at a specified moment or window. *Concurrent Body-Part State* identifies what one body part is doing given the activity of another.

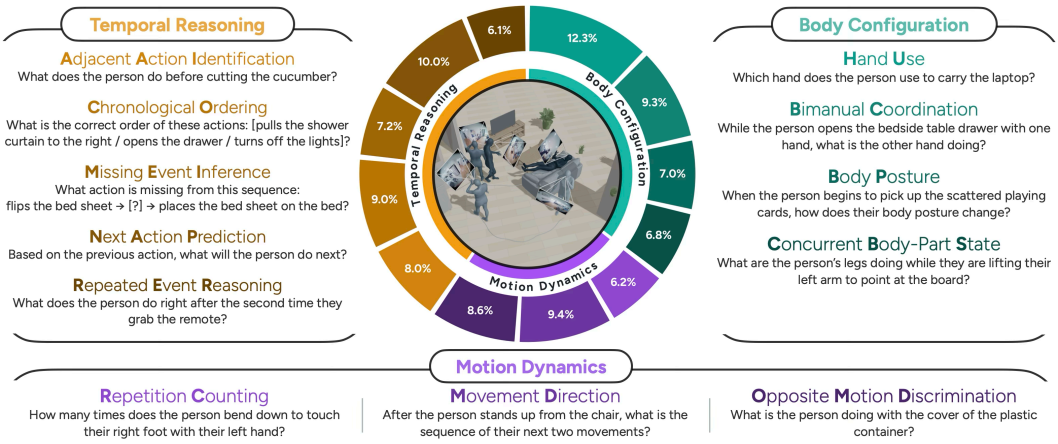


Figure 2: **SOMA Benchmark Taxonomy.** SOMA comprises 12 tasks organized into three capability axes: **Body Configuration**, **Temporal Reasoning** and **Motion Dynamics**. We show the distribution of questions across tasks, together with a representative question for each task. See Appendix A for more examples.

Motion Dynamics. This axis tests perception of fine-grained motion. *Repetition Counting* enumerates the occurrences of a specified repetitive motion. *Movement Direction* identifies the direction of motion, turning, or a short navigation sequence. *Opposite Motion Discrimination* distinguishes semantically opposed motions such as push and pull, open and close, that share visual context but differ in motion polarity.

Temporal Reasoning. This axis groups tasks that require reasoning about how motion-grounded events unfold over time, including their relationships and progression. *Adjacent Action Identification* identifies the action immediately preceding or following a reference event. *Chronological Ordering* places three or more actions in their correct temporal order. *Missing Event Inference* infers the missing intermediate action between two anchoring events. *Next Action Prediction* predicts the action following a trimmed clip, requiring inference of an unobserved action. *Repeated Event Reasoning* identifies what occurs after a specified occurrence of a repeated event.

3.2 Data Curation Pipeline

We construct SOMA through a multi-stage pipeline: sourcing data and task assignment, question generation, filtering, and human review (Figure 3). We describe each stage below.

Sourcing Data and Task Assignment. We sourced videos from Nymeria (Ma et al., 2024) and Ego-Exo4D (Grauman et al., 2024), two egocentric corpora that span a wide range of real-world domains with rich action and body motion annotations. Both corpora provide atomic action annotations $\mathcal{A} = \{(a_j, \tau_j)\}$, where a_j is a detailed natural-language action description and $\tau_j = [t_j^s, t_j^e]$ its temporal interval. Nymeria additionally provides motion narrations $\mathcal{M} = \{(p_j, h_j, \ell_j, \tau_j)\}$, fine-grained descriptions of body posture p_j, hand motion h_j, and leg motion ℓ_j over the same interval. These narrations are central to SOMA, as they describe body state and movement in natural language, which supports our body-centric question generation. Not every domain, however, supports every task equally; therefore, we discarded domain–task pairs that would produce a trivial question, such as asking about hand use during an activity where the hands are idle.

Question and Answer Generation. Our question generation procedure consists of two steps. First, an LLM *selector* receives the relevant textual annotation streams for a video (drawn from $\mathcal{A}$ and $\mathcal{M}$) and identifies a sub-sequence specifically relevant to the task. For example, for Bimanual Coordination, it identifies a segment in which both hands are simultaneously active. The annotation streams from the selected segment are then passed as input to an LLM *generator*, which produces a multiple-choice question with one correct answer and three task-specific distractors. This two-step design narrows the context before question construction, so each example is grounded in task-relevant annotations. We implemented both modules with Gemini 2.5 Flash and manually designed detailed prompts with step-by-step instructions and in-context examples for each task (Appendix B.1), producing 3,928 candidate questions. The only exception to this step is Repetition Counting, as exact counts of fine-grained motions cannot be reliably extracted from the narrations, and we found it more effective to create the questions and counts manually (Appendix B.3).

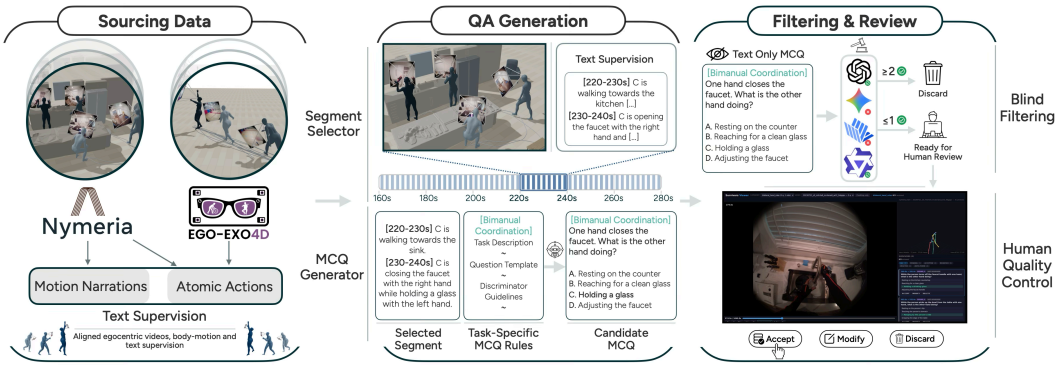


Figure 3: **SOMA Data Construction Pipeline.** We construct SOMA through task assignment, task-specific question generation, and quality control with blind-filtering and human validation.

Blind-Language Filtering. Modern LLMs carry substantial prior knowledge and can often answer questions correctly without "seeing" the video. A central goal of SOMA is to test whether models can perceive body motion in egocentric video, so questions answerable from language alone undermine the benchmark. To identify such questions, we apply a blind-language filter: four models (Gemini 2.5 Flash, GPT-4o, Qwen2.5-72B, and InternVL3-78B) answer each question without video, and we remove questions answered correctly by at least two solvers (Appendix B.4). This step removes 45.9% of the questions, leaving 2,127 for human review, and concentrates the benchmark on questions that cannot be reliably solved from language priors and require visual understanding.

Human Review. Finally, to verify the correctness and quality of every remaining question, a human reviewer inspects each question through a custom interface that streams the source video segment alongside a synchronized 3D body-pose rendering, used as an additional reference for validation. For each question, the reviewer chooses to accept, modify (its text, options, or segment bounds), or delete it, to minimize question ambiguity, so that the final benchmark contains only high-quality, verified questions (Appendix B.4).

3.3 Dataset Statistics

SOMA consists of 1,632 multiple-choice questions that span 12 tasks across three axes (Figure 2). The questions are drawn from 397 unique videos, producing 1,611 clips that span diverse everyday scenarios. The average length is 35.6 seconds with most concentrated between 10 and 60 seconds. Body Configuration and Motion Dynamics clips fall mainly within this range, reflecting the short-term nature of body-motion cues, while Temporal Reasoning clips extend further into the long tail. Answer distributions are balanced across options to mitigate position bias.

3.4 SOMA-Train

To facilitate further investigation and motivate future research in the domain, we introduce a training split, **SOMA-Train**, which comprises aligned egocentric video, body-motion, and language supervision. We adapted the benchmark generation pipeline (Sec. 3.2), without human verification, and constructed 28,596 question-answer pairs over 110.2 hours of aligned video and body motion, while keeping the training data disjoint from the SOMA evaluation set at the participant and scene level. Since existing Ego-Exo4D (Grauman et al., 2024) body-motion fits are provided at 10 fps and cover only part of the videos, we follow Patel et al. (2025) to reconstruct 3D body motion at 20 fps from the multi-view recordings. We detail the construction, debiasing method, and motion processing in Appendix E.

4 Experimental Setup

Human Performance. To contextualize model performance, we measure human accuracy without prior familiarity with the benchmark. Five compensated evaluators, none of whom took part in data annotation, independently answered all 1,632 questions, totaling over 450 hours of evaluation. We report their mean accuracy as an estimate of human-level performance on this task.

Evaluation Setup. We evaluate 37 leading MLLMs spanning closed-source, open-source, and specialized egocentric systems (Table 2 and Appendix C.2). We report per-task and overall accuracy (%), using `lmms-eval` (Zhang et al., 2025b), a standard multimodal evaluation framework, to ensure consistent evaluation and facilitate future comparisons. Unless otherwise specified, models are evaluated zero-shot with temperature set to 0 and uniformly sampled 16 frames per segment to balance computational cost and comparability across models with different context lengths. We also ablate key evaluation settings (Sec. 5.2). Further implementation details, including prompt templates, are provided in Appendix C.

5 EXPERIMENTS

We first compare MLLMs with human performance across the SOMA tasks (Sec. 5.1) and examine the effect of inference-time strategies (Sec. 5.2). We then analyze common failure modes (Sec. 5.3) and study whether combining video and body-motion signals can improve performance (Sec. 5.4).

5.1 MLLM AND HUMAN PERFORMANCE ON SOMA

All evaluated models struggle on SOMA, and no model reaches human performance. The best model, Qwen3.6-27B, reaches 50.0% average accuracy, and the best closed-source model, Gemini 3.5 Flash, reaches 47.0%, well below the human average of 68.5% (Table 2). The gap, however, is not uniform. Compared with the best model per task, it is largest on Repetition Counting (45.3 points), Repeated Event Reasoning (25.4), and Bimanual Coordination (25.3), whereas it closes or reverses on Concurrent Body-Part State (-1.1) and Next Action Prediction (-10.0). The difference in these tasks lies in whether the answer is present in the video or must be inferred. In Concurrent Body-Part State, the queried body part is often outside the field of view, and in Next Action Prediction, the action has not yet occurred. Since both humans and models must infer the answer, humans have little advantage on these tasks. In contrast, on the other tasks, the relevant cues are present in the video but are subtle or spread over time. While humans rely on these cues, we find that current models fail to use them reliably.

Performance is uneven across tasks. Models perform best on tasks where the evidence is concentrated in a few frames and can be directly observed, such as Hand Use and Opposite Motion Discrimination, where the best models reach 74.6% and 71.4%, respectively. In contrast, performance drops when the answer requires following the motion across the clip: on Chronological Ordering and Repeated Event Reasoning, most models score below 40%. This indicates that current MLLMs are better at recognizing frame-level cues than at integrating body motion over time.

Egocentric specialization alone is not sufficient. Despite being adapted to egocentric video, the three specialized models do not outperform general-purpose open-source models of similar size. This suggests a gap in current egocentric models when it comes to body-motion perception, consistent with their training data, which focuses primarily on actions, objects, and intentions, with no dedicated supervision of whole-body posture or motion. Addressing this limitation will likely require egocentric models that are more robust and more sensitive to fine-grained body dynamics.

Increasing model size improves performance, but larger models still fall short. Qwen3-VL gains 13.1 points from 2B to 32B, and Qwen3.5 gains 17.8 points from 0.8B to 27B, while InternVL3.5 improves by 8.1 points from 1B to 38B and is not monotonic. Larger backbones therefore help on SOMA, but scaling alone does not close the gap to human performance, and the largest models tested remain far from saturating the benchmark.

5.2 IMPACT OF TEMPORAL SAMPLING AND TEST-TIME REASONING

Effect of Temporal Sampling Density. To study how temporal sampling affects performance, we evaluate nine models on 1 to 64 uniformly sampled input frames (Figure 4A). Accuracy increases when moving from a single to multiple frames but saturates quickly. Averaged across models, it rises by 5.4 points from 1 to 16 frames, while adding frames beyond 16 contributes at most an additional 0.5. The magnitude of this gain is strongly model-dependent, ranging from 1.2 points (VideoLLaMA3-7B) to 10.3 points (Qwen3.5-9B) between 1 and 16 frames. Temporal sampling density is therefore not the principal limitation. Beyond access to visual evidence on the order of tens of frames, what models lack is the ability to integrate that evidence across time.

Table 2: **Comparison Results on SOMA.** Accuracy (%) across the three capability axes and their twelve tasks. HU: Hand Use; BC: Bimanual Coordination; BP: Body Posture; CBS: Concurrent Body-Part State; RC: Repetition Counting; MD: Movement Direction; OMD: Opposite Motion Discrimination; AAI: Adjacent Action Identification; CO: Chronological Ordering; MEI: Missing Event Inference; NAP: Next Action Prediction; RER: Repeated Event Reasoning. The best model results are marked in **bold**, and the second-best are underlined. Human performance is the mean over five annotators; standard deviations are reported in Table 4.

Models	Body Configuration				Motion Dynamics			Temporal Reasoning					Avg.
	HU	BC	BP	CBS	RC	MD	OMD	AAI	CO	MEI	NAP	RER	
Human Performance	*90.5*	*78.3*	*61.9*	*50.3*	*88.9*	*66.3*	*86.1*	*70.1*	*53.2*	*72.7*	*34.5*	*67.4*	*68.5*
Random Chance	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*
Closed-source MLLMs													
Claude Sonnet 5	52.7	42.4	43.5	40.5	36.6	33.3	47.9	44.3	33.3	39.0	29.9	35.0	40.3
Claude Opus 4.8	64.2	45.7	49.6	49.5	36.6	35.3	51.4	42.0	37.4	46.6	34.1	37.0	44.8
GPT-5.6 Luna	64.7	51.7	40.9	36.9	**43.6**	40.5	54.3	42.0	29.3	39.8	29.3	38.0	43.4
GPT-5.1	59.7	39.7	37.4	42.3	40.6	31.4	60.7	41.2	22.4	40.7	22.6	**42.0**	40.3
Gemini 3.5 Flash	73.1	**53.0**	46.1	43.2	37.6	34.6	**71.4**	42.7	42.2	34.7	32.3	36.0	47.0
Gemini 2.5 Pro	67.7	47.0	34.8	45.9	34.7	**46.4**	65.0	**49.6**	37.4	**55.1**	28.7	39.0	46.9
Gemini 2.5 Flash	43.3	41.1	38.3	36.9	36.6	36.6	39.3	38.2	27.2	26.3	20.1	32.0	34.8
Open-source MLLMs													
Qwen3.8-27B	72.1	47.7	46.1	44.1	36.6	40.5	58.6	44.3	34.7	48.3	39.0	37.0	47.0
Qwen3.6-27B	72.1	51.7	41.7	**51.4**	35.6	43.1	64.3	**49.6**	46.9	45.8	**44.5**	35.0	**50.0**
Qwen3.5-0.8B	40.3	39.1	25.2	35.1	24.8	28.1	33.6	29.8	32.7	26.3	31.7	24.0	31.7
Qwen3.5-2B	48.3	38.4	27.0	35.1	23.8	26.1	44.3	29.0	33.3	18.6	25.6	32.0	32.7
Qwen3.5-4B	**74.6**	42.4	34.8	40.5	39.6	32.0	51.4	45.0	38.8	46.6	25.0	32.0	43.1
Qwen3.5-9B	70.1	48.3	41.7	43.2	20.8	39.2	61.4	38.2	40.1	34.7	35.4	38.0	44.3
Qwen3.5-27B	73.6	47.0	44.3	47.7	31.7	42.5	67.9	45.8	**49.7**	45.8	39.0	**42.0**	49.5
Qwen3-VL-2B	33.8	33.8	33.0	34.2	34.7	33.3	32.1	32.8	26.5	27.1	19.5	36.0	31.1
Qwen3-VL-4B	46.8	29.1	33.9	32.4	28.7	32.0	52.9	34.4	32.0	22.9	22.6	36.0	34.1
Qwen3-VL-8B	56.2	33.1	32.2	28.8	32.7	32.7	57.9	38.2	26.5	39.0	19.5	**42.0**	37.1
Qwen3-VL-32B	70.1	41.7	35.7	40.5	36.6	34.6	67.1	43.5	37.4	44.9	25.6	40.0	44.2
InternVL3.5-1B	33.3	25.8	36.5	32.4	40.6	30.7	28.6	27.5	30.6	22.9	32.9	31.0	30.9
InternVL3.5-2B	33.3	31.8	45.2	34.2	36.6	33.3	35.0	35.9	29.9	25.4	28.0	36.0	33.4
InternVL3.5-4B	47.8	26.5	40.9	28.8	33.7	31.4	44.3	32.8	23.1	28.0	23.2	31.0	33.0
InternVL3.5-8B	35.8	31.1	41.7	37.8	31.7	32.0	46.4	34.4	17.7	37.3	29.3	27.0	33.4
InternVL3.5-14B	50.2	37.1	43.5	39.6	37.6	28.8	55.0	38.9	25.9	42.4	23.2	36.0	38.2
InternVL3.5-38B	52.2	44.4	37.4	32.4	36.6	30.1	59.3	42.0	29.9	45.8	17.7	38.0	39.0
Gemma-4-12B	45.3	41.7	36.5	46.8	32.7	22.9	30.7	28.2	47.6	22.0	30.5	26.0	34.8
Gemma-4-31B	67.2	41.1	39.1	46.8	37.6	34.6	50.7	38.2	38.8	30.5	40.2	38.0	43.1
GLM-4.6V-Flash	53.2	41.1	32.2	39.6	23.8	34.0	43.6	33.6	36.1	26.3	20.7	28.0	35.4
GLM-4.6V	54.7	36.4	36.5	36.9	35.6	34.6	57.1	42.0	40.8	26.3	20.7	37.0	38.8
Kimi-VL-A3B	38.8	30.5	33.9	34.2	32.7	34.6	45.0	35.1	25.2	24.6	23.2	30.0	32.5
MiniCPM-o-4.5-9B	55.2	35.8	40.9	46.8	32.7	29.4	42.9	35.1	30.6	26.3	23.2	30.0	36.3
Molmo2-4B	56.7	37.1	**53.0**	45.9	33.7	29.4	41.4	32.8	19.0	28.0	26.2	35.0	36.8
Molmo2-8B	54.2	42.4	51.3	41.4	33.7	36.6	39.3	39.7	38.1	37.3	36.0	29.0	40.6
LLaVA-OneVision-1.5-8B	44.3	32.5	34.8	28.8	27.7	28.1	35.7	26.7	30.6	27.1	21.3	29.0	31.1
VideoLLaMA3-7B	36.8	35.1	40.9	35.1	38.6	34.6	32.1	30.5	23.1	24.6	25.0	33.0	32.3
Egocentric MLLMs													
EgoGPT-7B	46.3	27.2	33.0	27.9	31.7	32.0	40.7	26.7	16.3	20.3	29.9	30.0	30.8
EgoVLM-3B	40.3	27.2	27.0	19.8	29.7	35.3	46.4	32.8	39.5	34.7	24.4	38.0	33.3
EgoThinker-7B	56.7	33.1	35.7	30.6	30.7	37.3	50.0	29.0	26.5	33.1	26.8	27.0	35.8

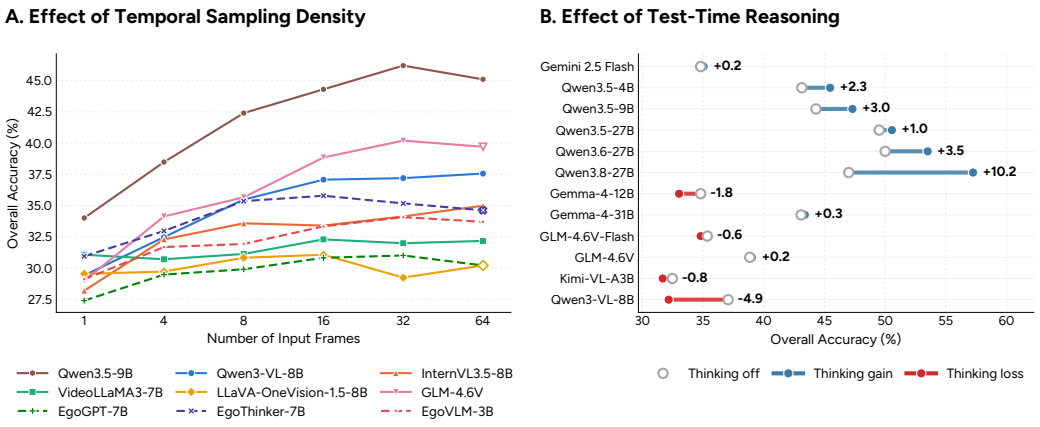


Figure 4: **Effect of Inference Strategies on SOMA. (A)** Increasing temporal sampling density. **(B)** Enabling test-time reasoning ("thinking").

Effect of Test-time Reasoning. We next ask whether explicit test-time reasoning, effective on many language reasoning tasks, also helps with visual reasoning, specifically body-motion perception in video. We compare the performance change after enabling thinking for instruction-tuned models, or after switching to their thinking variant (Qwen3-VL-8B and Kimi-VL-A3B), under each model's recommended sampling parameters. The effect is strongly model-dependent (Figure 4B). Five models improve by at least one point, all from the Qwen3.5–3.8 family, with gains of 1.0 to 10.2 points, while the remaining seven either show negligible change or decline. Test-time reasoning is therefore not a sufficient solution for closing the human–model gap: simply encouraging "thinking" in current reasoning mechanisms does not address their limitations in perceiving body motion.

5.3 Quantitative and Qualitative Error Analysis

The nature of some questions allows us to explicitly quantify model errors, which we complement with a qualitative analysis of reasoning traces.

Models show a strong right-hand bias on handedness questions. We split each handedness question by its correct answer and measure per-class accuracy. Averaged across the 37 models of Table 2, accuracy reaches 78.8% for *right hand* but drops to 44.3% for *left hand* (Appendix D.4.2), on a question that is, in principle, visually unambiguous: the five human annotators are equally accurate on both classes (mean accuracy of 92.7% left, 89.8% right). This asymmetry is consistent with a prior inherited from pretraining data: since roughly 90% of the human population is right-handed (Papadatou-Pastou et al., 2020), right-handed actions are over-represented in the video-text corpora used to train MLLMs, and models default to *right* when the visual evidence is weak.

Most models lean toward under-counting. Repetition counting shows the largest human–model gap in SOMA (88.9% for humans against 43.6% for the best model on this task), so we look more closely at how models fail, classifying each incorrect prediction as an under-count (below the true count) or an over-count (above the true count). In total, 24 of the 37 evaluated models lean toward under-counting, consistent with a frame-sampling explanation: repetitions falling between sampled frames are simply missed, biasing the count downward. The asymmetry is nonetheless strongly model-dependent, indicating that beyond the shared sampling bottleneck, individual models apply their own priors when they cannot directly enumerate repetitions (Appendix D.4.3).

Errors in Reasoning Traces. To better understand the limitations of current MLLMs, we analyze the errors made in their reasoning traces. We focus on Qwen3.8-27B and Gemma-4-31B, the latest models from the two leading open-source families on SOMA, and select 200 questions, stratified across tasks, that both models answered incorrectly. By inspecting the wrong predictions, we identify four primary error types: **1)** Visual Grounding Errors involve misidentifying the objects, scenes, or people involved; **2)** Motion and Body-State Perception Errors involve incorrectly perceiving the wearer's own movement, posture, or limb use; **3)** Temporal Localization Errors involve placing correctly identified events at the wrong point in time; and **4)** Prior-driven Inference Errors involve answering from expectation rather than visual evidence. Further details, including the annotation procedure and qualitative examples of each error type, are provided in Appendix D.4.

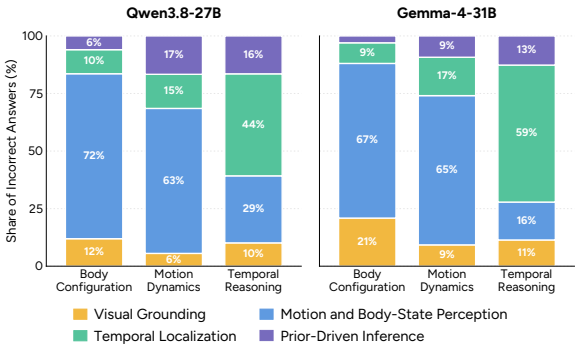


Figure 5: **Distribution of Error Types.** Root causes of incorrect answers across two representative MLLMs.

Motion and body-state perception is the dominant error for both models on Body Configuration and Motion Dynamics, indicating that most failures on these axes stem from misreading the wearer's own body, including misjudging their posture, the limb in use, or the direction of their movement (examples in Appendix D.4.1). In addition, models also struggle with temporal localization, showing difficulty in tracking the wearer's dynamics over time and in correct order. Visual grounding and prior-driven inference account for smaller shares of failures across all three axes, reflecting difficulties in identifying the relevant visual information and a tendency to favor plausible over ob-

served answers. By examining these failures, our analysis points to specific perceptual and temporal weaknesses that the next generation of MLLMs should address.

5.4 Training Body-Aware MLLMs

We found that MLLMs fail most when they need to perceive the wearer's body, which is hard to infer from egocentric vision. Humans do not rely on vision only: other exteroceptive and proprioceptive signals complement vision (vestibular, tactile and muscle spindles) and provide a continuous estimate of one's own posture and movement (Proske & Gandevia, 2012). This raises a natural question: if a model is given a comparable signal, a stream of body-motion information alongside the video, does it perceive and reason better?

We investigated this by extending two open-weight MLLM backbones, Qwen3-VL-8B (Bai et al., 2025) and InternVL3.5-8B (Wang et al., 2025), with a body-motion pathway following the standard LLaVA/MotionLLM design (Liu et al., 2023; Chen et al., 2025): a frozen VQ-VAE motion encoder (Zhang et al., 2023; Guo et al., 2022) was used to extract continuous latents, which were projected into the language-model embedding space and inserted as motion tokens in the input sequence (see Appendix E.3 for architecture, training curriculum, and hyperparameters). We fine-tuned each backbone on SOMA-Train (Appendix E) in three input settings, motion-only, video-only, and video+motion, and evaluated them on SOMA against the zero-shot video-only backbone (Table 3).

Table 3: **Effect of body-motion input.** Accuracy (%) per capability axis after fine-tuning with different input modalities. Per-task results in Table 11. †Zero-shot backbones are evaluated under the same video preprocessing as the fine-tuned models (16 frames, matched resolution).

Backbone	Setting	Body	Mot.	Temp.	Avg.
Qwen3-VL-8B	Zero-Shot†	35.88	40.89	31.67	35.39
	Motion	27.32	28.12	29.02	28.21
	Video	44.99	43.75	37.13	41.49
	Video+Mot.	**45.54**	**45.57**	**37.91**	**42.44**
InternVL3.5-8B	Zero-Shot†	38.43	40.10	29.49	35.20
	Motion	30.97	28.39	29.02	29.54
	Video	45.17	**44.53**	36.35	41.42
	Video+Mot.	**45.54**	43.75	**37.75**	**41.93**

Results. Motion-only fine-tuned MLLMs already perform above chance, suggesting that body motion is informative on its own, though performance remains low: motion alone is often insufficient to infer visual content, such as surrounding scenes or objects. For comparison, existing motion-only models evaluated zero-shot remain close to random chance on SOMA (Appendix D.5). Video fine-tuning improves both backbones well above zero-shot performance. Adding aligned motion provides a further but modest improvement for both backbones. However, the effect varies across capability axes: motion slightly improves Body Configuration and Temporal Reasoning for both backbones, while on Motion Dynamics it helps Qwen3-VL-8B but slightly hurts InternVL3.5-8B (see Appendix E.4 for task-level results). We also found results consistent with these findings when ablating video/motion modalities at test time (Table 12). Overall, these results suggest that body motion can provide information complementary to the visual signal, but that its benefits depend on the task and remain limited for the provided, vanilla baselines. This motivates future research on multimodal sensory integration.

6 Conclusion

We introduced SOMA, a benchmark for fine-grained understanding of human body motion in egocentric video, spanning 12 tasks across three capability axes. Across closed-source, open-source, and specialized egocentric models, we found a large gap to human performance, showing that perceiving the moving body from a first-person view remains an open problem.

We see a natural parallel to the human brain, where visual input from the eyes is complemented by multiple sensors (tactile, vestibular, proprioceptive) to form a coherent representation of the subject and its environment (Blanke, 2012). In egocentric video, the body is largely unseen, so a model must likewise infer posture, locomotion, and actions from indirect evidence such as hand pose and ego-motion. We found that current MLLMs underperform in this capacity. We hope that SOMA motivates the development of multimodal systems that better pair first-person video with motion signals to acquire the body awareness current models lack – and that systems meant to assist, instruct, or learn from humans will ultimately require. We note that beyond MLLMs, emerging world models (Matsuo et al., 2022) might be ideally suited for this challenge.

AI USE STATEMENT

In the preparation of this manuscript, we used generative AI tools for writing assistance. The primary purpose of their use was to aid in editing and polishing the writing of this paper to improve clarity and readability, such as improving grammar and refining sentence structure. All research ideas, methods, analyses, and conclusions in this paper were formulated and developed exclusively by the authors. We have manually reviewed and verified all AI-assisted edits for accuracy and take full responsibility for the final content of this work. In addition, LLMs were used as part of our benchmark construction pipeline to generate candidate question-answer pairs (Sec. 3.2), and all generated questions were human-verified.

ETHICS STATEMENT

The ethical considerations and potential risks of this work are limited. SOMA is constructed from two existing public egocentric video datasets, Nymeria and Ego-Exo4D, collected under informed-consent protocols established by their original creators, which we enhance with human-verified novel question-answer annotations; this verification process is anonymized. We used these datasets for benchmark construction and evaluation, following their intended research use. We manually verified the generated QA to ensure quality. Our human evaluation study involved 5 paid participants, who gave informed consent and were compensated for their time. Our experiments evaluate publicly accessible MLLM systems, including released open-source checkpoints and public API models, and are limited to aggregate benchmark performance rather than deployment in real-world decision-making settings. Our analysis also revealed a handedness bias in LLMs, which future research could address. More broadly, while direct risks are limited, advances in egocentric body-motion understanding could support beneficial applications such as assistive technologies and robotics, but may also raise concerns around surveillance or behavioral profiling.

REPRODUCIBILITY STATEMENT

The details needed to reproduce our work are provided throughout the main paper and appendix. The SOMA construction pipeline is described in Section 3.2, with additional details, including the exact selector, generator, and blind-filtering prompts, given in Appendix B. The human evaluation protocol is detailed in Appendix C.1. The MLLM evaluation setup is explained in Section 4, with specific implementation details given in Appendix C.3, together with the complete list of evaluated models in Appendix C.2 and the exact prompts used for evaluation in Appendices C.4 and C.5. The construction of our training dataset is described in Appendix E, and our fine-tuning architecture and hyperparameters are given in Appendix E.3. Code, benchmark data, and training data will be made publicly available upon paper acceptance.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jake K Aggarwal and Quin Cai. Human motion analysis: A review. *Computer vision and image understanding*, 73(3):428–440, 1999.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Anthropic. Introducing Claude Opus 4.8, May 2026a. URL https://www.anthropic.com/news/claude-opus-4-8.

Anthropic. Introducing Claude Sonnet 5, June 2026b. URL https://www.anthropic.com/news/claude-sonnet-5.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Olaf Blanke. Multisensory brain mechanisms of bodily self-consciousness. *Nature reviews neuroscience*, 13(8):556–571, 2012.

Andy Bonnetto, Haozhe Qi, Franklin Leong, Matea Tashkovska, Mahdi Rad, Solaiman Shokur, Friedhelm C Hummel, Silvestro Micera, Marc Pollefeys, and Alexander Mathis. Epfl-smart-kitchen: An ego-exo multi-modal dataset for challenging action and motion understanding in video-language models. *Advances in Neural Information Processing Systems*, 38, 2025.

Andy Bonnetto, Sepideh Mamooler, Chengkun Li, and Alexander Mathis. The behavior biopsy: Interpreting animal behavior as embodied, situated, and hierarchical. *Current Opinion in Neurobiology*, 98:103201, 2026.

Edwin G Boring. A chart of the psychometric function. *The American Journal of Psychology*, 28 (4):465–470, 1917.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.

Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. Motionllm: Understanding human behaviors from human motions and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. EgoThink: Evaluating First-Person Perspective Thinking Capability of Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14291–14302, 2024. doi: 10.1109/cvpr52733.2024.01355.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28652–28668, 2026.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, Yuyang Wen, Huiping Liu, Rongkang Wang, Yaqi Zhang, Hongliang Wei, Chi Chen, You Li, Kechen Fang, Jie Zhou, Yuxuan Li, Guoyang Zeng, Chaojun Xiao, Yankai Lin, Xu Han, Maosong Sun, Zhiyuan Liu, and Yuan Yao. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. 2026. URL https://arxiv.org/abs/2604.27393.

Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2969–2978, 2022.

Mark Endo, Joy Hsu, Jiaman Li, and Jiajun Wu. Motion question answering via modular motion programs. In *International Conference on Machine Learning*, pp. 9312–9328. PMLR, 2023.

Aaron Ferguson, Ahmed A. A. Osman, Berta Bescos, Carsten Stoll, Chris Twigg, Christoph Lassner, et al. MHR: Momentum human rig, 2025. URL https://arxiv.org/abs/2511.15586.

Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens Van Der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16102–16112, 2022.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14783–14794, 2023.

Google DeepMind. Gemini 3.5 Flash model card, May 2026a. URL https://deepmind.google/models/model-cards/gemini-3-5-flash/.

Google DeepMind. Gemini 3.8 Flash model card, September 2026b. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J. Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, David Crandall, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024. doi: 10.1109/cvpr52733.2024.01834.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1900–1910, 2024.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26584–26595, 2024.

Yuping He, Yifei Huang, Guo Chen, Baoqi Pei, Jilan Xu, Tong Lu, and Jiangmiao Pang. EgoExoBench: A Benchmark for First- and Third-person View Video Understanding in MLLMs. In *Advances in Neural Information Processing Systems*, volume 38, pp. 129355–129398, 2025. doi: 10.52202/085713-3891.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie 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*, pp. 8450–8460, 2025. doi: 10.1109/cvpr52734.2025.00791.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. EgoTaskQA: Understanding Human Tasks in Egocentric Videos. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3343–3360, 2022. doi: 10.52202/068431-0242.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. In *International Conference on Machine Learning*, pp. 1078–1088. PMLR, 2020.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. Genmo: A generalist model for human motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11766–11776, 2025a.

Junlong Li, Huaiyuan Xu, Sijie Cheng, Kejun Wu, Kim-Hui Yap, Lap-Pui Chau, and Yi Wang. Building egocentric procedural ai assistant: Methods, benchmarks, and challenges. *arXiv preprint arXiv:2511.13261*, 2025b.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024b.

Lei Li, Sen Jia, Jianhao Wang, Zhongyu Jiang, Feng Zhou, Ju Dai, Tianfang Zhang, Zongkai Wu, and Jenq-Neng Hwang. Human motion instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17582–17591, June 2025c.

Yanjun Li, Yuqian Fu, Tianwen Qian, Qi'ao Xu, Silong Dai, Danda Pani Paudel, Luc Van Gool, and Xiaoling Wang. EgoCross: Benchmarking Multimodal Large Language Models for Cross-Domain Egocentric Video Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6592–6600, 2026. doi: 10.1609/aaai.v40i8.37589.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, Kevin Bailey, David Soriano Fosas, C. Karen Liu, Ziwei Liu, Jakob Engel, Renzo De Nardi, and Richard Newcombe. Nymeria: A Massive Collection of Multimodal Egocentric Daily Motion in the Wild. In *European Conference on Computer Vision*, pp. 445–465. Springer, 2024. doi: 10.1007/978-3-031-72691-0_25.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46212–46244, 2023. doi: 10.52202/075280-2004.

Yutaka Matsuo, Yann LeCun, Maneesh Sahani, Doina Precup, David Silver, Masashi Sugiyama, Eiji Uchibe, and Jun Morimoto. Deep learning, reinforcement learning, and world models. *Neural networks*, 152:267–275, 2022.

Arsha Nagrani, Sachit Menon, Ahmet Iscen, Shyamal Buch, Ramin Mehran, Nilpa Jha, Anja Hauth, Yukun Zhu, Carl Vondrick, Mikhail Sirotenko, et al. Minerva: Evaluating complex video reasoning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23968–23978. IEEE, 2025.

OpenAI. GPT-5.1: A smarter, more conversational ChatGPT. `https://openai.com/index/gpt-5-1/`, November 2025. Accessed: 2026-05-25.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition, July 2026. URL `https://openai.com/index/gpt-5-6/`.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Marietta Papadatou-Pastou, Eleni Ntolka, Judith Schmitz, Maryanne Martin, Marcus R Munafò, Sebastian Ocklenburg, and Silvia Paracchini. Human handedness: A meta-analysis. *Psychological bulletin*, 146(6):481, 2020.

Chaitanya Patel, Hiroki Nakamura, Yuta Kyuragi, Kazuki Kozuka, Juan Carlos Niebles, and Ehsan Adeli. UniEgoMotion: A unified model for egocentric motion reconstruction, forecasting, and generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10318–10329, 2025. doi: 10.1109/ICCV51701.2025.00961.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10975–10985, 2019.

Baoqi Pei, Yifei Huang, Jilan Xu, Yuping He, Guo Chen, Fei Wu, Jiangmiao Pang, and Yu Qiao. Egothinker: Unveiling egocentric reasoning with spatio-temporal cot. *Advances in Neural Information Processing Systems*, 38:44140–44168, 2025.

Taiying Peng, Jiacheng Hua, Miao Liu, and Feng Lu. In the Eye of MLLM: Benchmarking Egocentric Video Intent Understanding with Gaze-Guided Prompting. In *Advances in Neural Information Processing Systems*, volume 38, pp. 121404–121424, 2025. doi: 10.52202/085713-3659.

Chiara Plizzari, Alessio Tonioni, Yongqin Xian, Achin Kulshrestha, and Federico Tombari. Omnia de egotempo: Benchmarking temporal understanding of multi-modal llms in egocentric videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24129–24138, 2025.

Uwe Proske and Simon C Gandevia. The proprioceptive senses: their roles in signaling body shape, body position and movement, and muscle force. *Physiological reviews*, 2012.

Haozhe Qi, Chen Zhao, Mathieu Salzmann, and Alexander Mathis. Hoisdf: Constraining 3d hand-object pose estimation with global signed distance fields. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10392–10402. IEEE, 2024.

Haozhe Qi, Shaokai Ye, Alexander Mathis, and Mackenzie W Mathis. LLaVAction: evaluating and training multi-modal large language models for action understanding. In *The Fourteenth International Conference on Learning Representations*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL `https://qwen.ai/blog?id=qwen3.5`.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026b. URL `https://qwen.ai/blog?id=qwen3.6-27b`.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026c. URL `https://qwen.ai/blog?id=qwen3.8`.

Bin Ren, Mengyuan Liu, Runwei Ding, and Hong Liu. A survey on 3d skeleton-based action recognition using learning method. *Cyborg and Bionic Systems*, 5:0100, 2024.

Isaac Robinson, Peter Robicheaux, Matvei Popov, Deva Ramanan, and Neehar Peri. RF-DETR: Neural architecture search for real-time detection transformers, 2025. URL `https://arxiv.org/abs/2511.09554`. ICLR 2026.

Lucas Stoffl, Andy Bonnetto, Stéphane d'Ascoli, and Alexander Mathis. Elucidating the hierarchical nature of behavior with masked autoencoders. In *European conference on computer vision*, pp. 106–125. Springer, 2024.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haoning Wu, Haotian Yao, Haoyu Lu, Heng Wang, Hongcheng Gao, Huabin Zheng, Jiaming Li, Jianlin Su, Jianzhou Wang, Jiaqi Deng, Jiezhong Qiu, Jin Xie, Jinhong Wang, Jingyuan Liu, Junjie Yan, Kun Ouyang, Liang Chen, Lin Sui, Longhui Yu, Mengfan Dong, Mengnan Dong, Nuo Xu, Pengyu Cheng, Qizheng Gu, Runjie Zhou, Shaowei Liu, Sihan Cao, Tao Yu, Tianhui Song, Tongtong Bai, Wei Song, Weiran He, Weixiao Huang, Weixin Xu, Xiaokun Yuan, Xingcheng Yao, Xingzhe Wu, Xinxing Zu, Xinyu Zhou, Xinyuan Wang, Y. Charles, Yan Zhong, Yang Li, Yangyang Hu, Yanru Chen, Yejie Wang, Yibo Liu, Yibo Miao, Yidao Qin, Yimin Chen, Yiping Bao, Yiqin Wang, Yongsheng Kang, Yuanxin Liu, Yulun Du, Yuxin Wu, Yuzhi Wang, Yuzi Yan, Zaida Zhou, Zhaowei Li, Zhejun Jiang, Zheng Zhang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Zijia Zhao, and Ziwei Chen. Kimi-VL technical report, 2025a. URL https://arxiv.org/abs/2504.07491.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025b. URL https://arxiv.org/abs/2507.01006.

Ashwin Vinod, Shrey Pandit, Aditya Vavre, and Linshen Liu. Egovlm: Policy optimization for egocentric video understanding, 2025. URL https://arxiv.org/abs/2506.03097.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pp. 396–416. Springer, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Bizhu Wu, Jinheng Xie, Keming Shen, Zhe Kong, Jianfeng Ren, Ruibin Bai, Rong Qu, and Linlin Shen. Mg-motionllm: A unified framework for motion comprehension and generation across multiple granularities. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27849–27858, 2025a.

Jane Wu, Georgios Pavlakos, Georgia Gkioxari, and Jitendra Malik. Reconstructing hand-held objects in 3d from images and videos. In *International Conference on 3D Vision (3DV)*, 2026.

Qi Wu, Yubo Zhao, Yifan Wang, Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. Motion-agent: A conversational framework for human motion generation with llms. In *International Conference on Learning Representations*, pp. 20314–20334, 2025b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, Bo Li, and Ziwei Liu. EgoLife: Towards Egocentric Life Assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28885–28900, 2025b. doi: 10.1109/cvpr52734.2025.02690.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations*, volume 2026, pp. 157051–157088, 2026a.

Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, Matt Feiszli, Jitendra Malik, Piotr Dollár, and Kris Kitani. SAM 3D Body: Robust full-body human mesh recovery, 2026b. URL https://arxiv.org/abs/2602.15989.

Han Yi, Yulu Pan, Feihong He, Xinyu Liu, Benjamin Zhang, Oluwatumininu Oguntola, and Gedas Bertasius. ExAct: A Video-Language Benchmark for Expert Action Analysis. In *Advances in Neural Information Processing Systems*, volume 38, pp. 8031–8061, 2025. doi: 10.52202/085713-0246.

Shaofeng Yin, Yanjie Ze, Hong-Xing Yu, C Karen Liu, and Jiajun Wu. Visualmimic: Visual humanoid loco-manipulation via motion tracking and generation. *arXiv preprint arXiv:2509.20322*, 2025.

Ling-An Zeng, Guohong Huang, Gaojie Wu, and Wei-Shi Zheng. Light-t2m: A lightweight and fast model for text-to-motion generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9797–9805, 2025.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a. URL https://arxiv.org/abs/2501.13106.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. LMMs-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 881–916, 2025b. URL https://aclanthology.org/2025.findings-naacl.51/.

Yao Zhang, Zhuchenyang Liu, Thomas Ploetz, and Yu Xiao. Encoder-free human motion understanding via structured motion descriptions. *arXiv preprint arXiv:2604.21668*, 2026.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

Bingfan Zhu, Biao Jiang, Sunyi Wang, Shixiang Tang, Tao Chen, Linjie Luo, Youyi Zheng, and Xin Chen. Motiongpt3: Human motion as a second modality. *arXiv preprint arXiv:2506.24086*, 2025.

APPENDIX

TABLE OF CONTENTS

A ADDITIONAL SOMA EXAMPLES 18

B QUESTION & ANSWER GENERATION 29
- B.1 Selector and Generator Prompt Examples . . . 29
- B.2 Mitigating Generator Bias . . . 31
- B.3 Manual Annotation of Repetition Counting . . . 31
- B.4 Filtering and Review Statistics . . . 32

C EVALUATION DETAILS 32
- C.1 Details on Human evaluation . . . 32
- C.2 Evaluated Models . . . 33
- C.3 Implementation Details . . . 33
- C.4 Zero-Shot Evaluation Prompt Template . . . 33
- C.5 Chain-of-Thought Evaluation Prompt Template . . . 33

D ADDITIONAL RESULTS 34
- D.1 Effect of Chain-of-Thought Prompting . . . 34
- D.2 Inference-Time Analysis per Capability Axis . . . 35
- D.3 Performance by Question Difficulty . . . 36
- D.4 Error Analysis . . . 36
- D.5 Evaluation of Motion-Only Baselines . . . 39

E SOMA TRAIN 39
- E.1 QA Generation . . . 40
- E.2 Ego-Exo4D Refitting . . . 42
- E.3 Training Details . . . 44
- E.4 Additional Results . . . 45

F LICENSING INFORMATION 46

A Additional SOMA Examples

We provide examples for each of the twelve tasks (Figures 6 to 27), each showing eight frames sampled uniformly from the clip, together with the question, its four options, and the correct answer (marked with ✓).

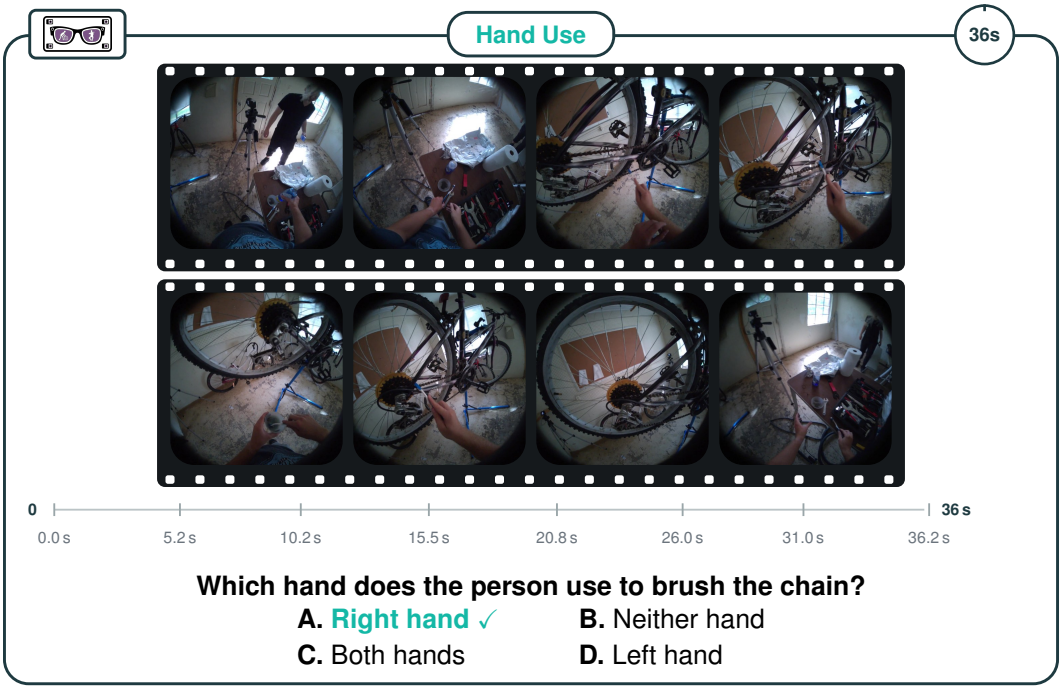


Figure 6: **Hand Use.** Ego-Exo4D clip, 36 s.

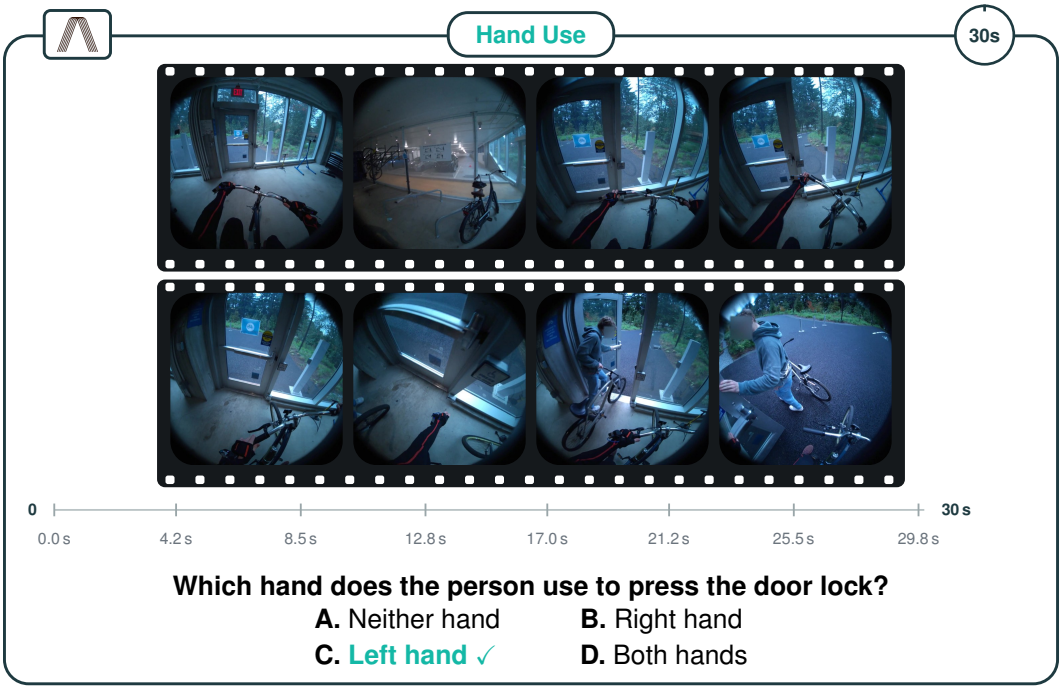


Figure 7: **Hand Use.** Nymeria clip, 30 s.

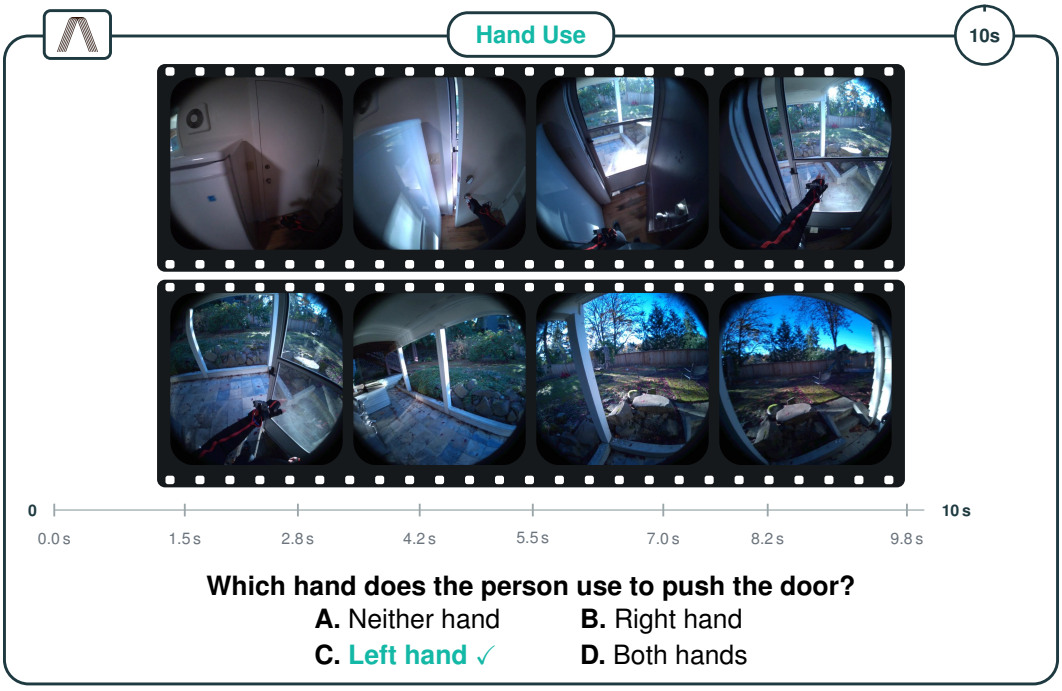


Figure 8: **Hand Use.** Nymeria clip, 10 s.

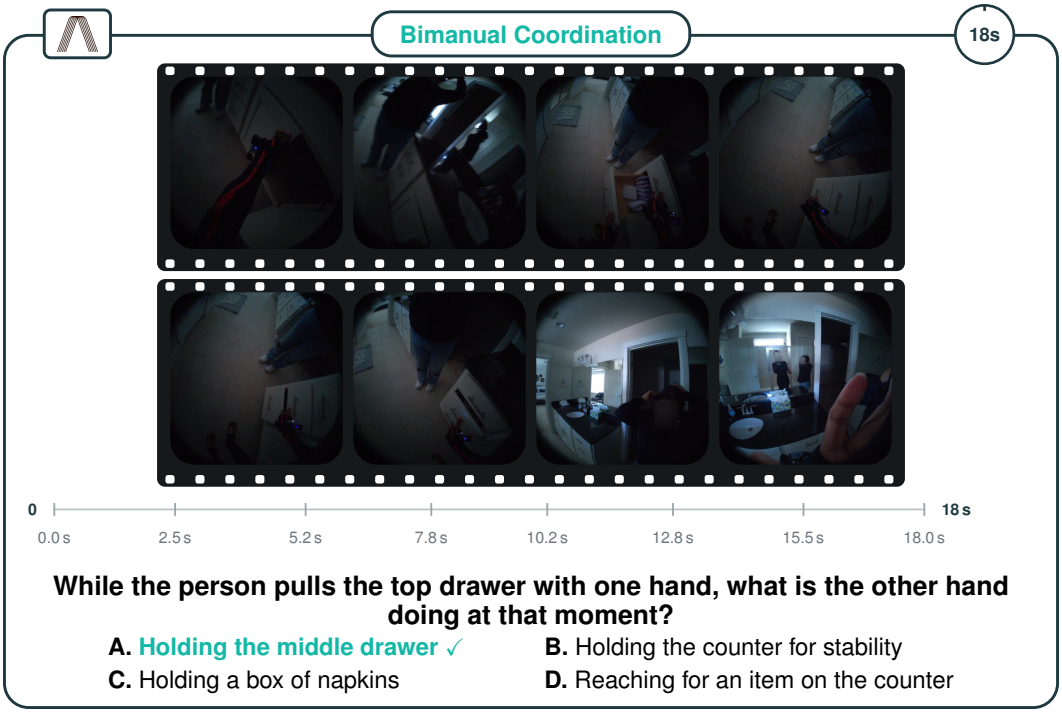


Figure 9: **Bimanual Coordination.** Nymeria clip, 18 s.

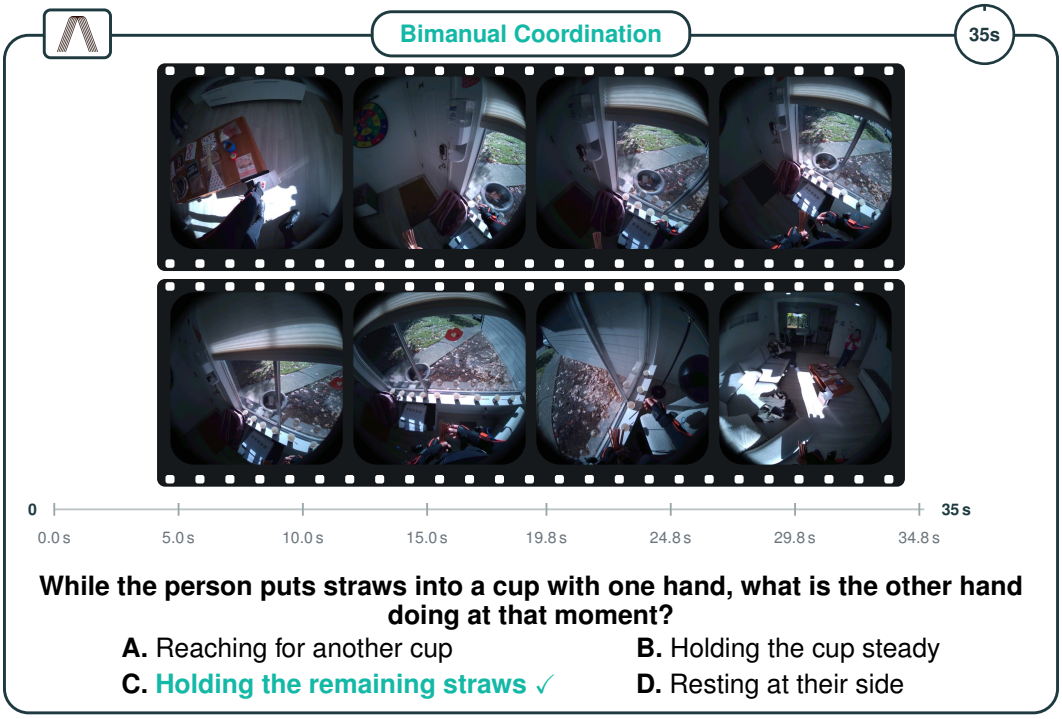


Figure 10: **Bimanual Coordination.** Nymeria clip, 35 s.

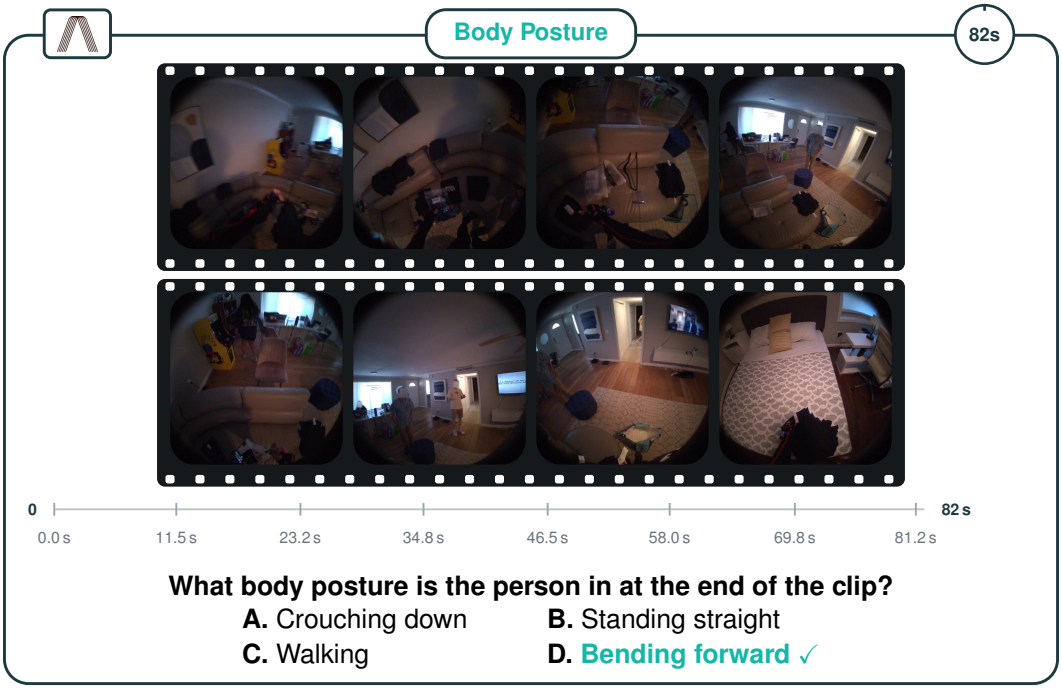


Figure 11: **Body Posture.** Nymeria clip, 82 s.

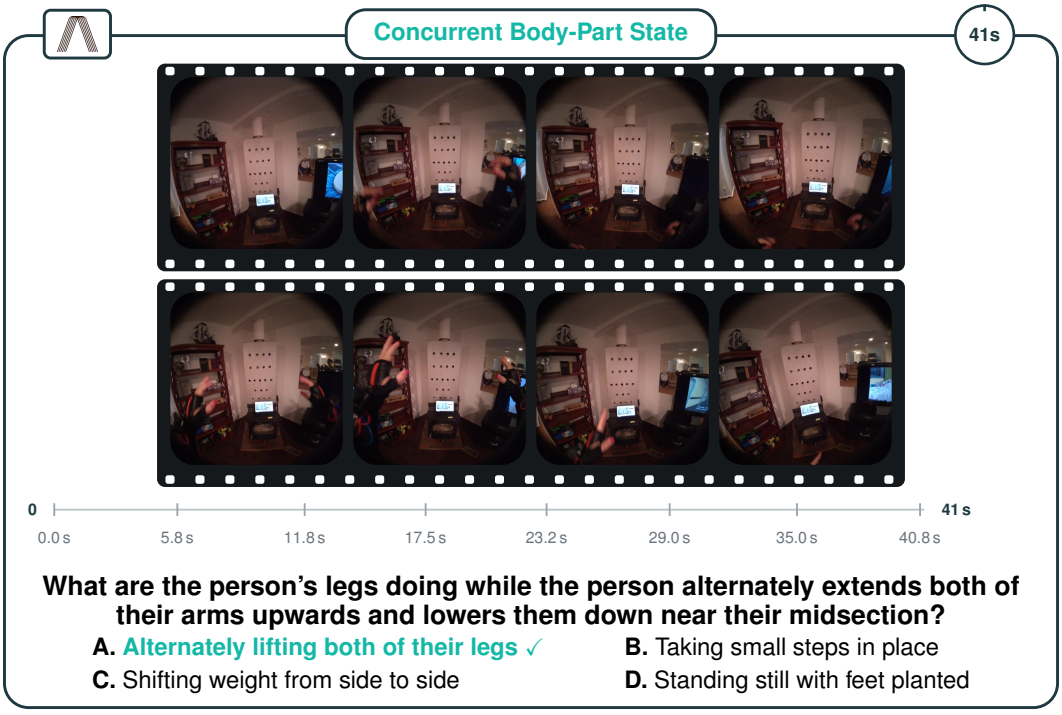


Figure 12: **Concurrent Body-Part State.** Nymeria clip, 41 s.

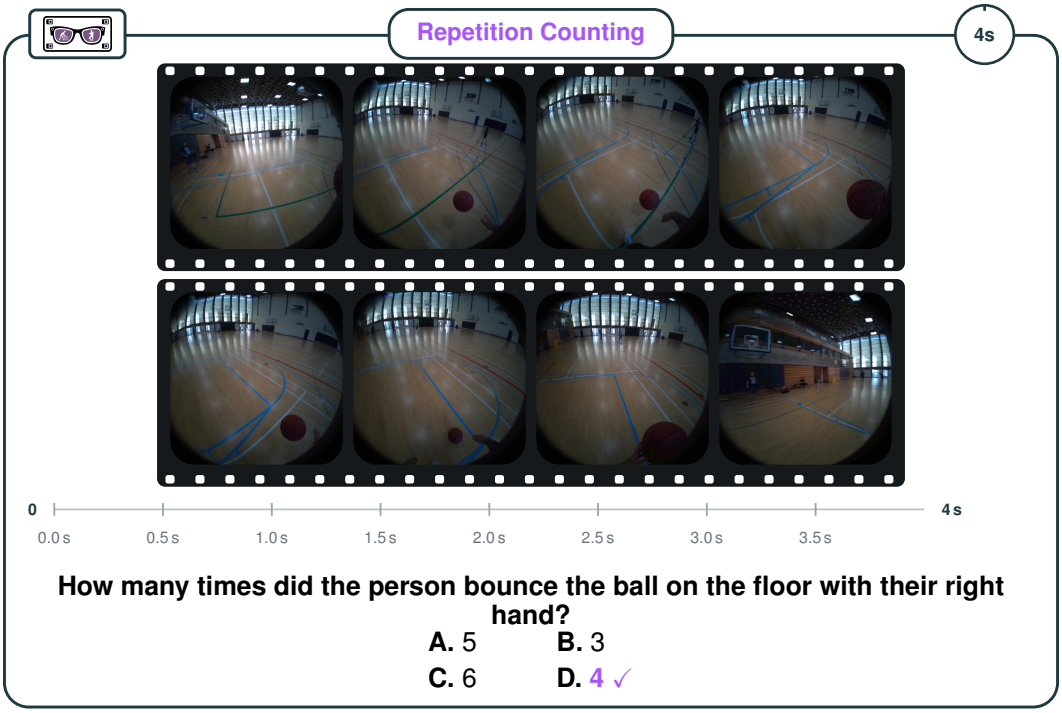


Figure 13: **Repetition Counting.** Ego-Exo4D clip, 4 s.

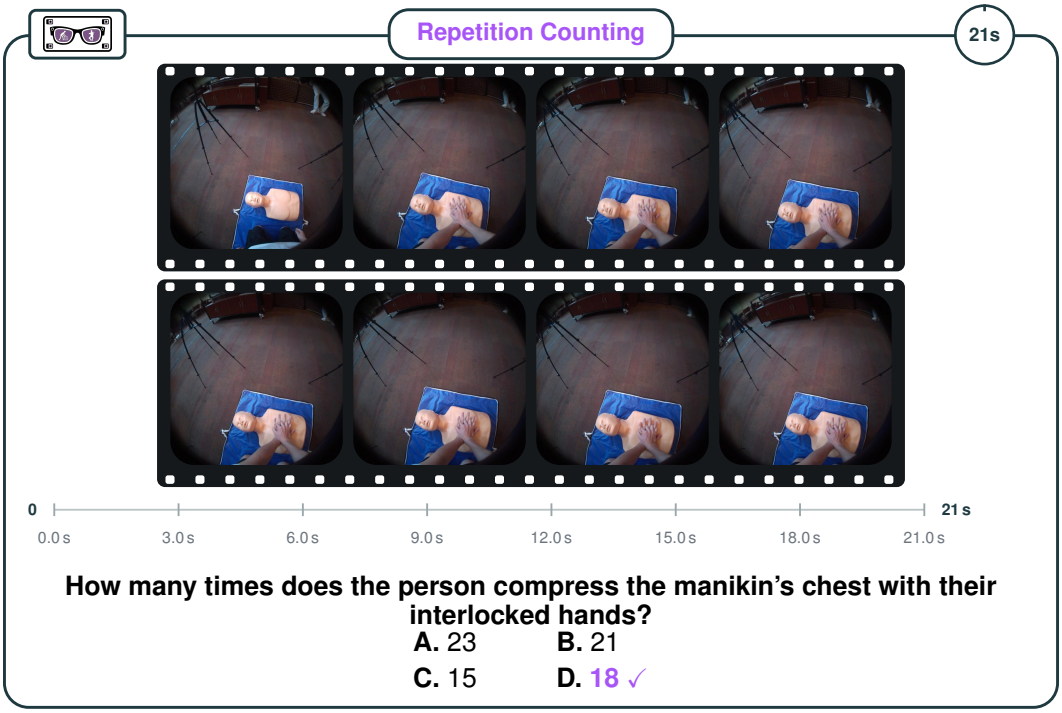


Figure 14: **Repetition Counting.** Ego-Exo4D clip, 21 s.

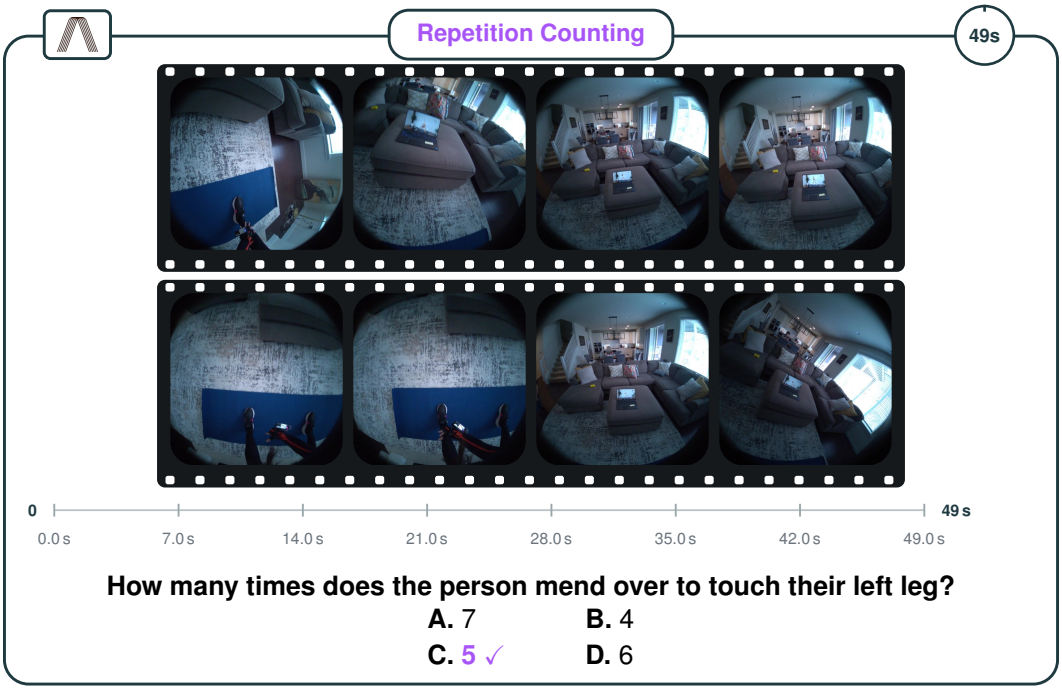


Figure 15: **Repetition Counting.** Nymeria clip, 49 s.



Figure 16: **Movement Direction.** Nymeria clip, 35 s.

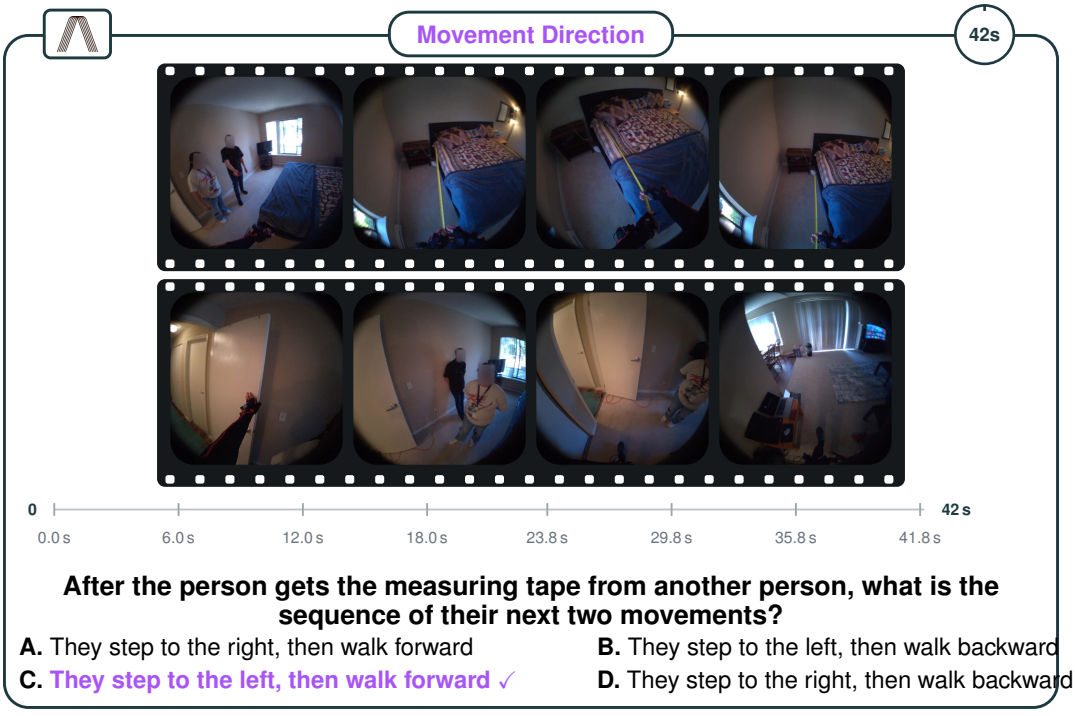


Figure 17: **Movement Direction.** Nymeria clip, 42 s.

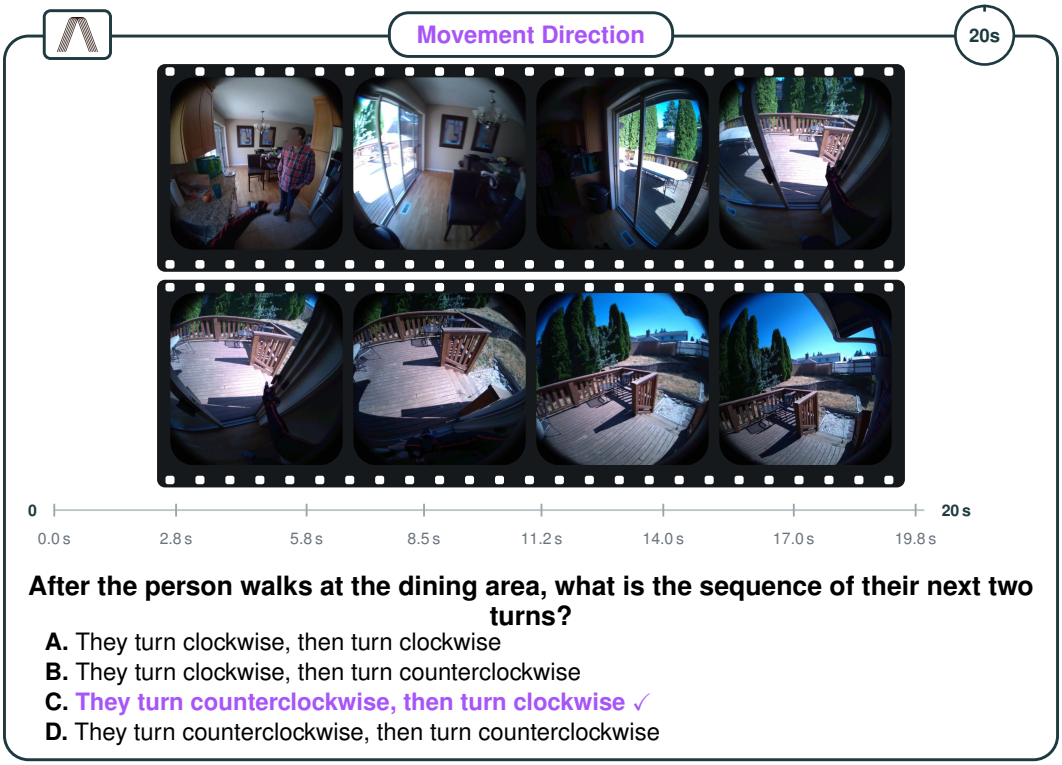


Figure 18: **Movement Direction.** Nymeria clip, 20 s.

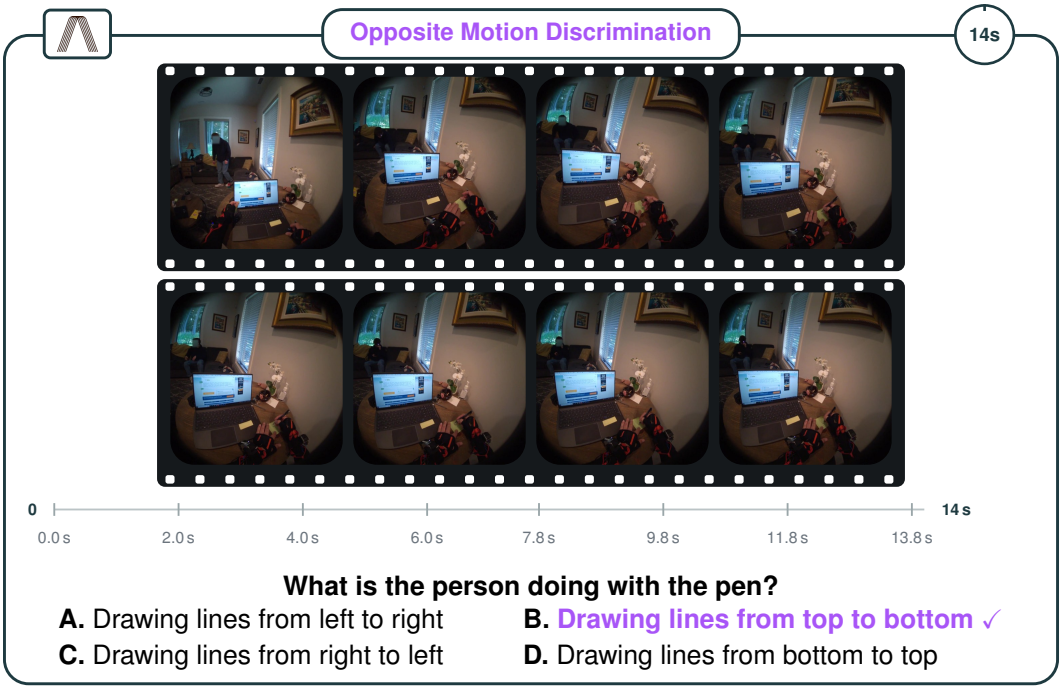


Figure 19: **Opposite Motion Discrimination.** Nymeria clip, 14 s.

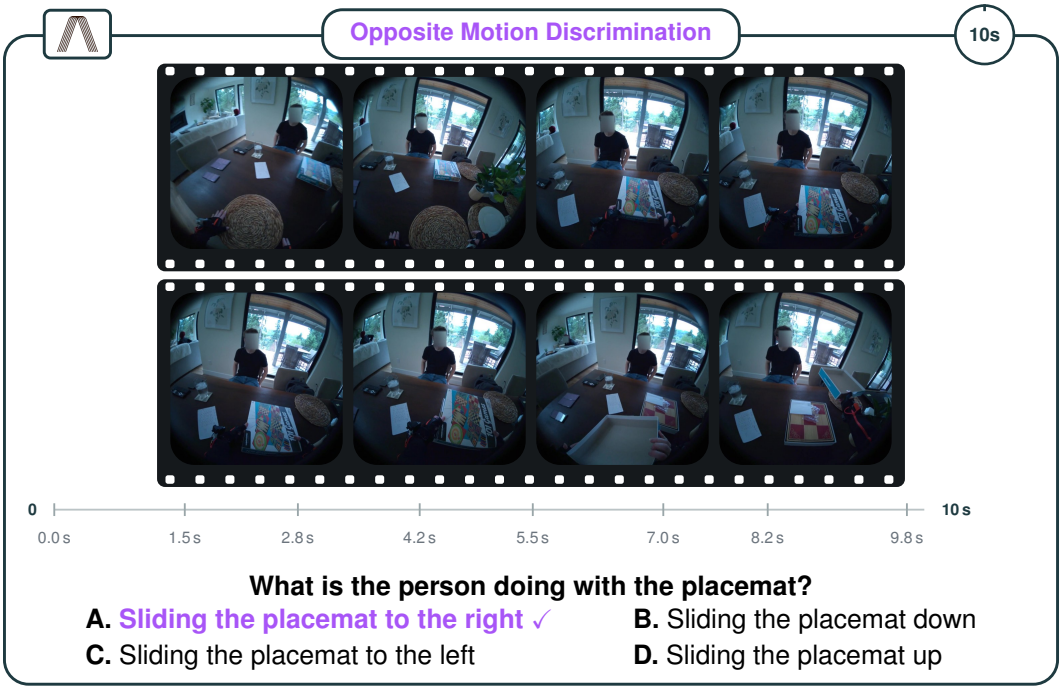


Figure 20: **Opposite Motion Discrimination.** Nymeria clip, 10 s.

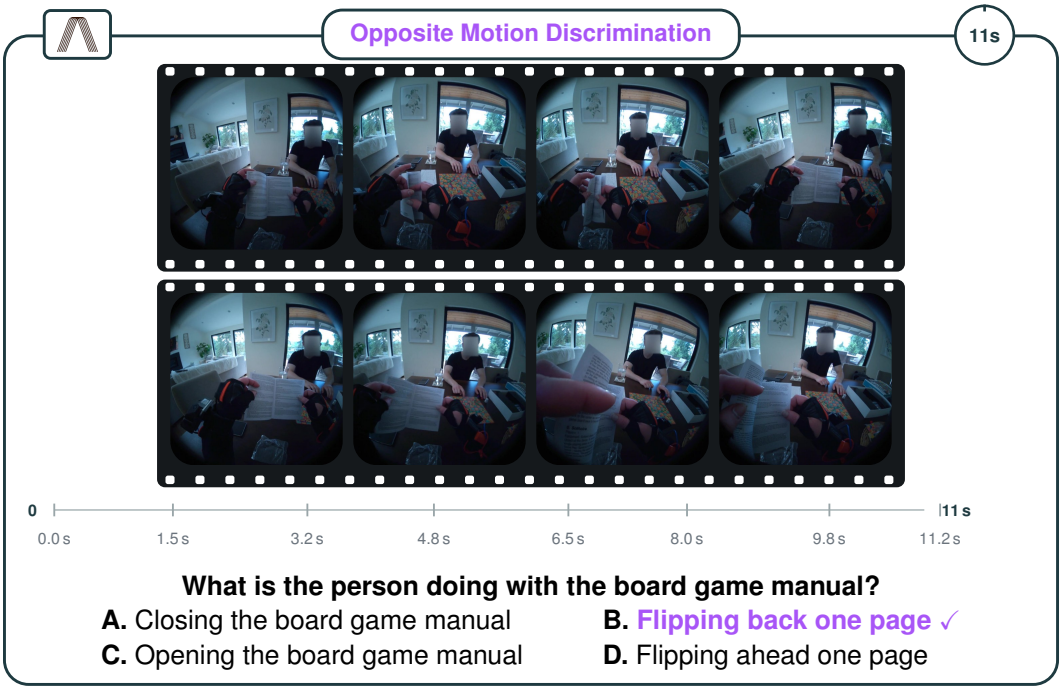


Figure 21: **Opposite Motion Discrimination.** Nymeria clip, 11 s.

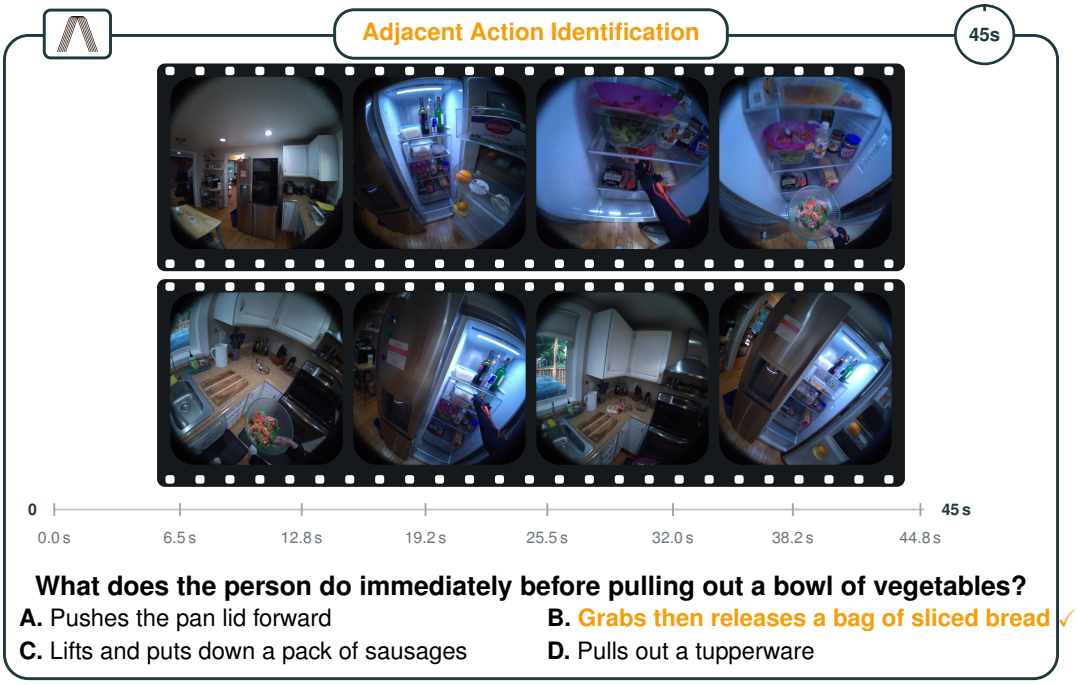


Figure 22: **Adjacent Action Identification.** Nymeria clip, 45 s.

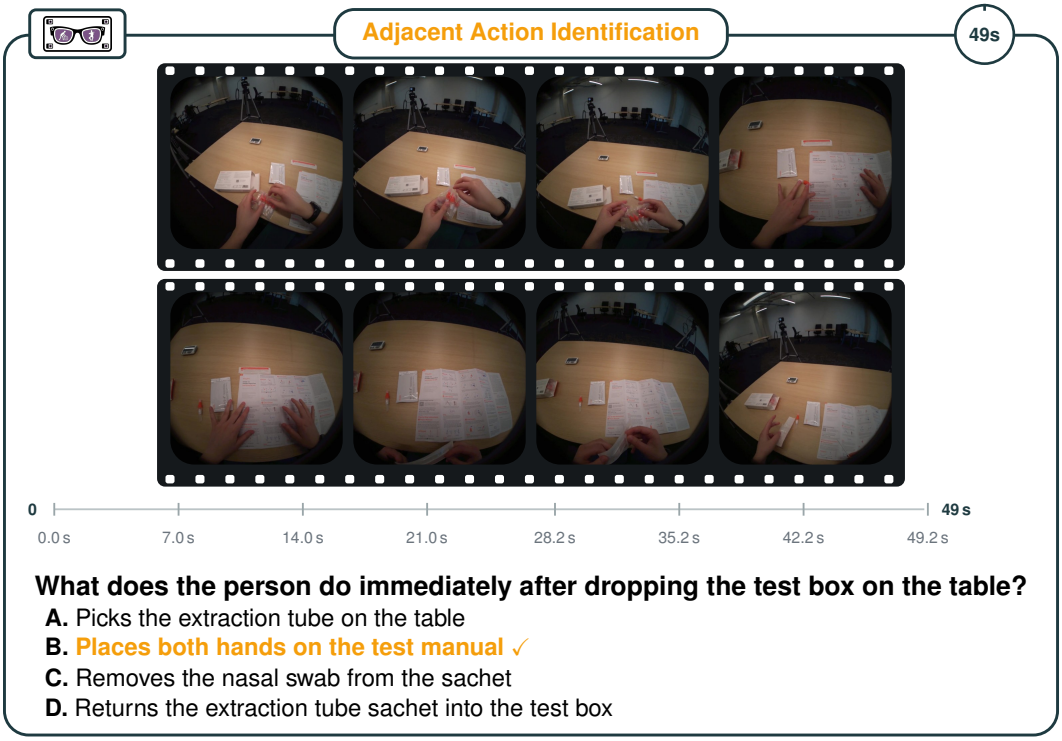


Figure 23: **Adjacent Action Identification.** Ego-Exo4D clip, 49 s.

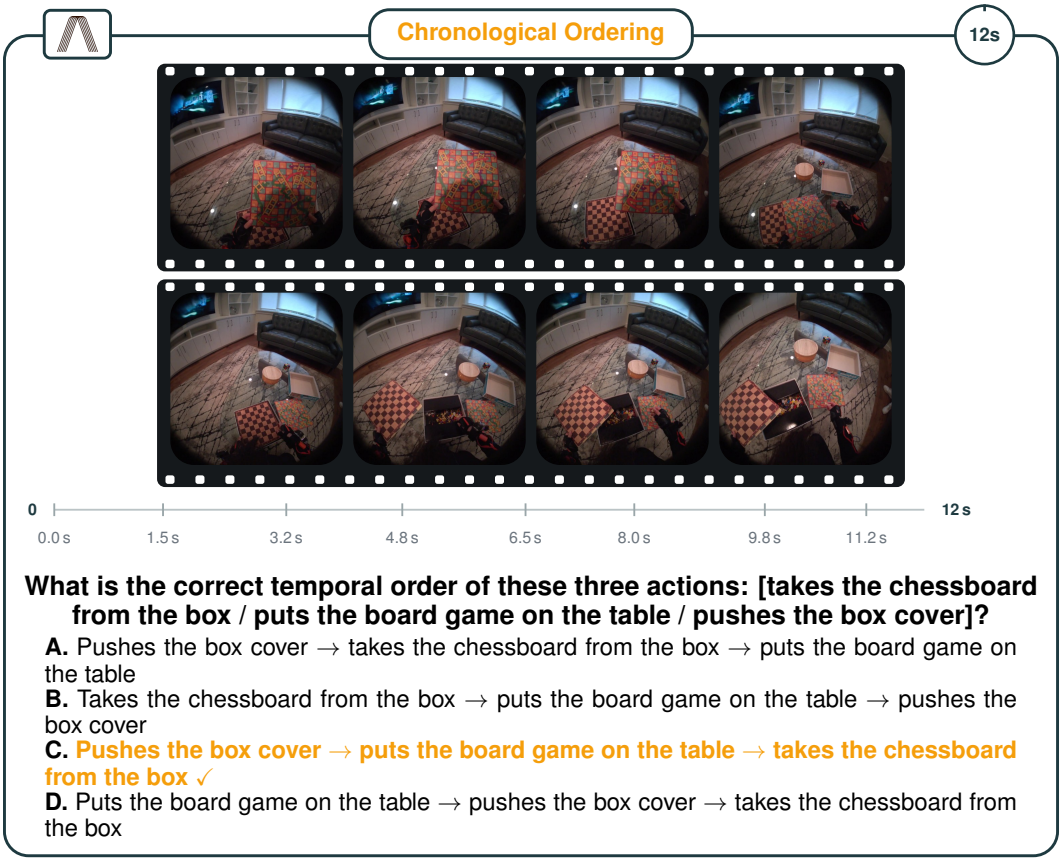


Figure 24: **Chronological Ordering.** Nymeria clip, 12 s.

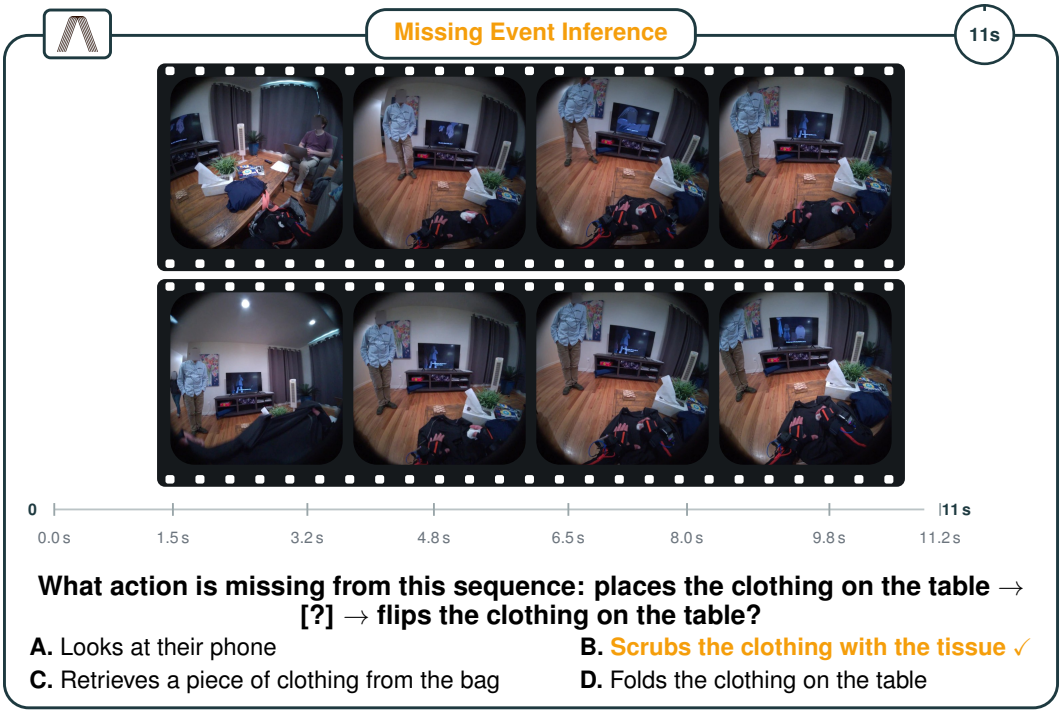


Figure 25: **Missing Event Inference.** Nymeria clip, 11 s.

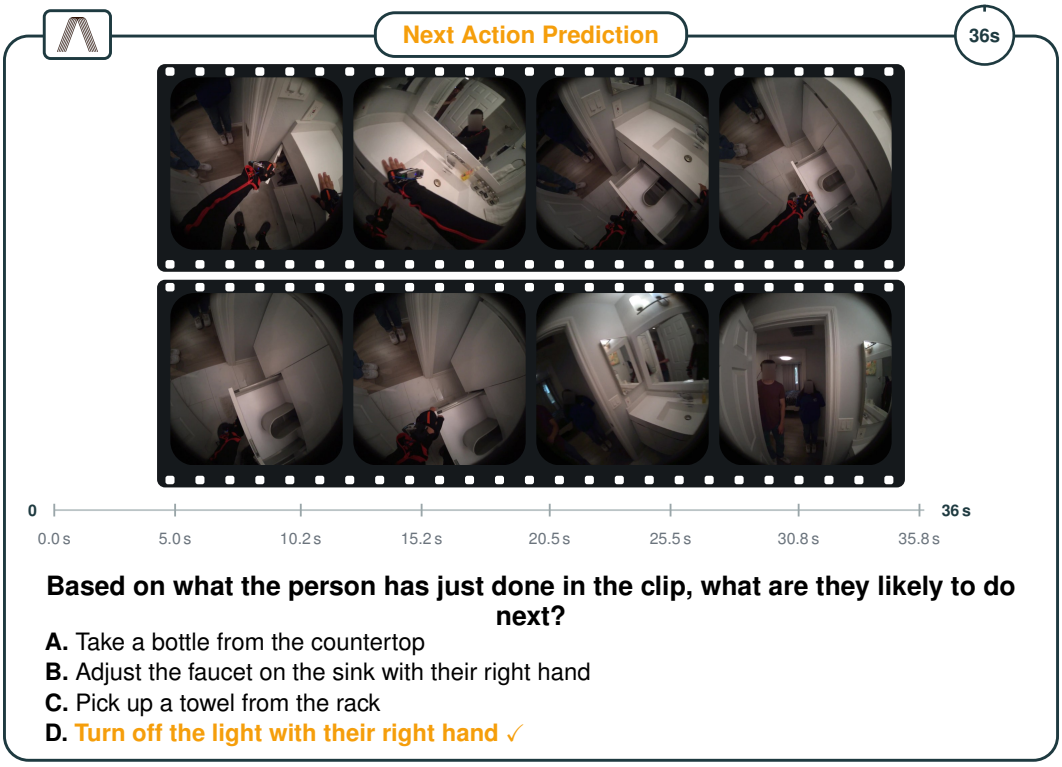


Figure 26: **Next Action Prediction.** Nymeria clip, 36 s.

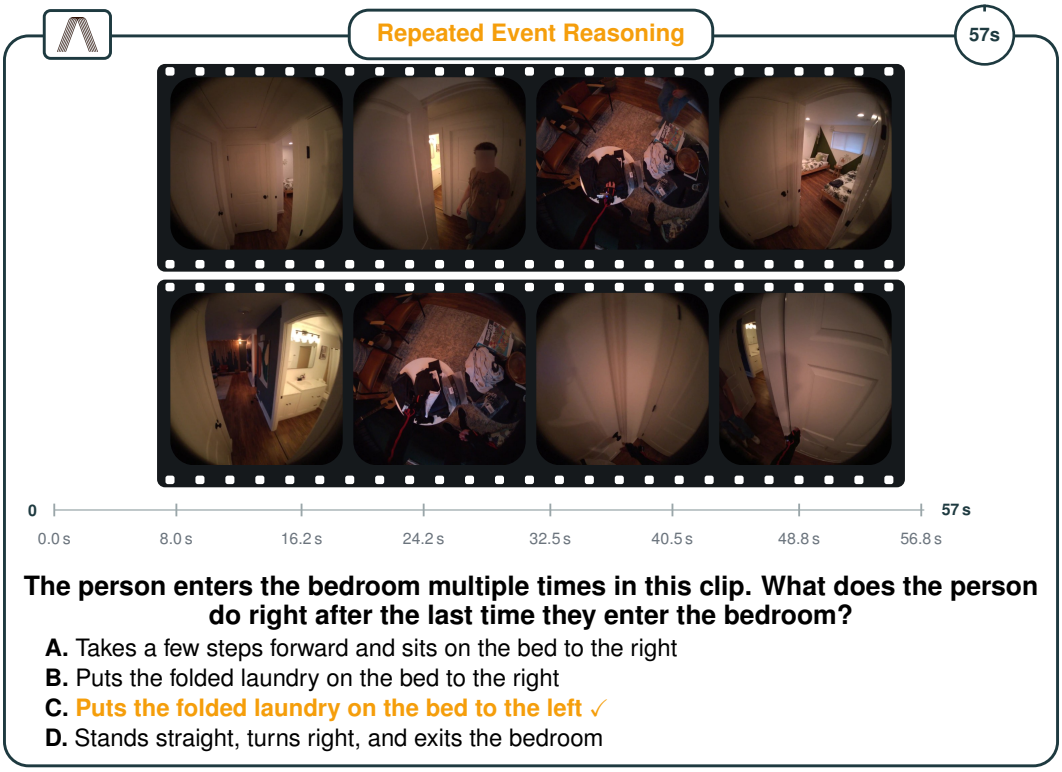


Figure 27: **Repeated Event Reasoning.** Nymeria clip, 57 s.

B Question & Answer Generation

B.1 Selector and Generator Prompt Examples

For each of the LLM-generated tasks we crafted a pair of task-specific prompts: one for the *selector*, which scans the textual annotation streams of a video and returns sub-sequences that support a valid question, and one for the *generator*, which drafts a single multiple-choice question from the selected sub-sequence. The strategy for constructing distractors is likewise task-specific. *Bimanual Coordination*, for instance, swaps the role assignment between the two hands, *Adjacent Action Identification* substitutes plausible alternative actions from the same activity, and analogous strategies apply to the remaining tasks. We show the two prompts for the *Repeated Event Reasoning* task below as a representative example; the complete prompts for all tasks will be released with the code. Long or repetitive sections (e.g., the full list of refusal conditions) are abbreviated as "[...]".

Selector Prompt: Repeated Event Reasoning

```
# Role
You are an expert at selecting sub-sequences from egocentric-video
atomic-action annotations for generating Repeated-Event Reasoning
multiple-choice questions.

# What this category tests
Whether a vision-language model can track a RECURRING event across a
long video, index into a specific occurrence (``first time'', ``second
time'', ``last time''), and reason about what happens right after that
particular occurrence.  It is not counting, not single-action
reasoning, and not generic before/after ordering.

Representative question types:
- ``The person picks up their phone multiple times during the session.
What happens right after the FIRST time they do this?''
- ``After the SECOND time the person opens the refrigerator, what do
they do next?''

# Input
Every atomic action annotation for one video, numbered [0], [1], [2],
..., in chronological order.  [...]

# Your job
Return sub-sequences (contiguous index ranges) that contain a
recurring event (≥2 occurrences) AND concrete successor content after
at least two of those occurrences.

## What counts as a recurring event
A specific, nameable action that occurs ≥2 times in the span and is
clearly the same action each time (same verb, same object, same
intent).  Examples:  picks up a phone, opens/closes the refrigerator,
transfers an object from one hand to the other.  The verb/object pair
must be stable across occurrences.

## What makes a good sub-sequence
A good span has ALL of:
1.  The same event appears 3+ times, with clearly different
occurrences (3 strongly preferred so the generator has enough material
for distractors).
2.  Concrete successor content after each occurrence (a specific
physical action, not just ``leans forward'' or ``turns around'').
3.  Distinct successors across occurrences (if all three are followed
by the same thing, there is no interesting distinction to ask about).
4.  Occurrences meaningfully separated in time (≥10s between
consecutive ones).

## What to AVOID
- Spans where no event clearly recurs.
- Recurring postures or generic motion (``walking'', ``standing'').
```

```
- Identical successor content after every occurrence.
- Back-to-back duplicate annotations (same moment re-described).
[...]

# Output format
Return ONLY a JSON array, no prose, no code fences:
{''start_idx'': <int>, ''end_idx'': <int>}
If the video has no recurring event with distinct successor content,
return [].
```

Generator Prompt: Repeated Event Reasoning

```
# Role
You are an expert in designing multiple-choice questions for the
Repeated-Event Reasoning category: questions that reference a
RECURRING event (multiple occurrences in the clip) and an INDEXING
word (first / second / last) to pick out a particular occurrence, then
ask what the person does right after that occurrence.

# What this benchmark evaluates
A good Repeated-Event question is one where:
- A human watching the clip can answer by identifying all occurrences
and looking at what follows the specified one.
- A model with weak temporal understanding gets it wrong.
- A blind reader (question only, no video) cannot guess the answer.

# Valid question patterns
Every question has the shape: ''After the {INDEX} time the person
{recurring event}, what does the person do next?'' Main subject:
''the person'', 3rd-person singular.
- Pattern A: First occurrence.
- Pattern B: Last occurrence.
- Pattern C: Specific middle occurrence (only when the span contains
3+ occurrences).

# Procedure
Step 1. Find the recurring event (same verb/object pair, with
multiple separated occurrences).
Step 2. Identify the successor of each occurrence.
Step 3. Pick the target occurrence whose successor is MOST DIFFERENT
from the others, and the pattern (A/B/C).
Step 4. Draft the question. Explicitly state that the event recurs
(''...multiple times in this clip''), include the INDEX word.
Step 5. Blind-reader check: reject if the answer can be guessed from
wording alone.
Step 6. Draft the correct answer: a 3--10-word phrase describing the
FIRST specific action right after the target occurrence, in 3rd-person
singular present tense.
Step 7. Draft 3 distractors. Each must (a) describe a concrete
physical action (no transitional verbs like ''moves'', ''leans'',
''turns'' without a named destination), and (b) be temporally possible
at the moment the answer happens.
At least TWO distractors must be immediate successors of OTHER
occurrences of the SAME recurring event. [...].

# Worked example for Pattern A (first occurrence)
Input span:
- C walks to the kitchen and opens the refrigerator.
- C takes a bottle of milk out of the refrigerator.
- C places the bottle on the counter and closes the refrigerator.
- C pours milk into a glass.
- C walks back to the refrigerator and opens it.
- C puts the bottle of milk back into the refrigerator.
- C closes the refrigerator.
```

```
- [... one more open/close cycle]
Recurring event: ``open the refrigerator'' occurs 3 times. After
1st: takes a bottle out. After 2nd: puts the bottle back. After
3rd: closes it and walks to the living room.
Expected output:
{``question'': ``The person opens the refrigerator multiple times in
this clip. What does the person do right after the FIRST time they
open it?'',
 ``answer'': ``Takes a bottle of milk out of the refrigerator'',
 ``distractors'': [``Puts the bottle of milk back into the
refrigerator'', ``Closes the refrigerator and walks to the living
room'', ``Pours milk into a glass'']}
[... one more worked example for Pattern B (last occurrence)]

# Output format
Return exactly:
{``question'': ``<your question>'', ``answer'': ``<correct option>'',
 ``distractors'': [<3 distractors>]}

# Refusal conditions
Return {``question'': null} if any hold:
R1. No event recurs multiple times with the same verb/object pair.
R2. All occurrences are followed by the same successor content.
R3. The answer is an obvious continuation of the event (blind-reader
leak).
R4. Cannot produce 3 distractors where at least one is a successor of
a DIFFERENT occurrence. [...]
```

B.2 Mitigating Generator Bias

Benchmark questions generated by an LLM may inherit regularities of the generating model in their phrasing, answer formulation, and distractor construction, which could advantage models from the same family at evaluation time, a concern analogous to self-preference in LLM-based evaluation (Panickssery et al., 2024). Since Gemini 2.5 Flash is both our generator and one of the evaluated models, we took two measures to mitigate such bias. First, the question generation stage is not an open-ended LLM writing process: the model fills fixed task-specific question templates, while both the correct answer and the distractors are derived from the ground-truth annotations under predefined rules (Appendix B.1). The generator therefore has limited freedom to introduce stylistic regularities of its own, and the resulting questions largely reflect the source annotations rather than the generating model. Second, Gemini 2.5 Flash is one of the four blind solvers in the filtering stage (Sec. 3.2). Questions that Gemini answers correctly without the video, together with at least one other blind solver, are therefore removed during construction, before any MLLM is evaluated.

B.3 Manual Annotation of Repetition Counting

For each selected take, we used the 3D body-pose data to identify segments containing repetitive motion; a human annotator then inspected each candidate segment, examining the video together with the synchronized 3D body-pose rendering, and created the question and recorded the count for each. One repetition is defined as a single completed instance of the action named in the question. Segments containing partial repetitions, occluded repetitions, or otherwise unclear repetition boundaries are discarded rather than annotated with an uncertain count, so a repetition is counted only when it is clearly visible. Examples are available in Appendix A. This adds 101 Repetition Counting questions to the candidate set.

B.4 Filtering and Review Statistics

Across the eleven automatically generated tasks, our pipeline produced 3,928 candidate questions. The blind-language filter removed questions answered correctly by at least two of four solvers, using the prompt shown below.

Blind Solver Prompt

```
You are answering a multiple-choice question about a first-person
(``egocentric'') video that you have NOT been shown.

You will see only the question text and four options labeled A, B, C,
D. Use your world knowledge, common sense, and any clues in the
wording of the question itself to pick the option you think is most
likely correct.  Commit to exactly one letter even if uncertain ---
guess if you have to.

Respond with ONLY a JSON object with the single key ``answer'', whose
value is one of ``A'', ``B'', ``C'', ``D'':
{``answer'':  ``X''}

Do not include any other text, explanation, markdown, or extra keys.
```

This process removed 1,801 questions (45.9%), leaving the rest 2,127 for human review. Removal rates were highest for Body Posture (62.4%) and Hand Use (53.7%), where the answer space is small and natural language priors were strongest, and lowest for Repeated Event Reasoning (30.8%) and Chronological Ordering (39.8%), where the correct answer depended on multi-event temporal structure that is difficult to recover from text alone.

During human review, reviewers accepted 945 questions as-is, modified 586, and deleted 596. Combined with the 101 manually annotated Repetition Counting questions (Appendix B.3), the final benchmark contains 1,632 multiple-choice questions.

C Evaluation Details

C.1 Details on Human evaluation

Human performance was measured as the average accuracy of five human subjects who were not involved in the construction or annotation of the benchmark. Each subject independently evaluated the full set of 1,632 samples. For every sample, annotators were presented with the video, question, and four answer options, and are allowed unlimited time to select the single best answer based on the visual information in the video. When the camera wearer's body is not directly visible, annotators were instructed to rely on indirect visual cues, such as camera motion, viewpoint changes, and hand-object interactions. Completing the full evaluation required 89, 90, 94, 97, and 95 hours for the five annotators, respectively, corresponding to an average of 93 hours per annotator and 465 hours of human evaluation in total. Annotators were compensated for their time spent completing the evaluation.

We report the standard deviation in accuracy across the five human annotators for each task (Table 4). Human performance is most consistent on Hand Use and varies most on Chronological Ordering and Missing Event Inference.

Table 4: **Variability of Human Performance.** Accuracy (%) on each SOMA task, reported as mean $\pm$ standard deviation across the five annotators. Task abbreviations follow Table 2.

Body Configuration				Motion Dynamics			Temporal Reasoning					Avg.
HU	BC	BP	CBS	RC	MD	OMD	AAI	CO	MEI	NAP	RER	
90.5 ± 1.8	78.3 ± 3.7	61.9 ± 8.0	50.3 ± 9.2	88.9 ± 4.3	66.3 ± 9.1	86.1 ± 8.5	70.1 ± 7.3	53.2 ± 14.4	72.7 ± 15.6	34.5 ± 4.9	67.4 ± 5.9	68.5 ± 5.5

C.2 Evaluated Models

We conducted a comprehensive evaluation spanning closed-source, open-source, and specialized egocentric MLLMs, covering diverse model scales and recent architectural advances. For closed-source models, we included Gemini 2.5 Flash and Gemini 2.5 Pro (Comanici et al., 2025), Gemini 3.5 Flash (Google DeepMind, 2026a), GPT-5.1 (OpenAI, 2025) and GPT-5.6 Luna (OpenAI, 2026), and Claude Sonnet 5 (Anthropic, 2026b) and Claude Opus 4.8 (Anthropic, 2026a). For open-source models, we evaluated Qwen3-VL (2B–32B) (Bai et al., 2025), Qwen3.5 (0.8B–27B) (Qwen Team, 2026a), Qwen3.6-27B (Qwen Team, 2026b), and Qwen3.8-27B (Qwen Team, 2026c), InternVL3.5 (1B–38B) (Wang et al., 2025), Gemma-4 (12B, 31B) (Team et al., 2026), GLM-4.6V and GLM-4.6V-Flash (Team et al., 2025b), Molmo2 (4B, 8B) (Clark et al., 2026), Kimi-VL-A3B (Team et al., 2025a), MiniCPM-o-4.5-9B (Cui et al., 2026), LLaVA-OneVision-1.5-8B (An et al., 2025), and VideoLLaMA3-7B (Zhang et al., 2025a). For specialized egocentric models, we included EgoGPT-7B (Yang et al., 2025b), EgoVLM-3B (Vinod et al., 2025), and EgoThinker-7B (Pei et al., 2025).

C.3 Implementation Details

We uniformly sampled 16 frames per video segment across all models to balance computational cost and comparability across models with different context lengths; Section 5.2 studies the effect of denser sampling on performance. We used `lmms-eval` (Zhang et al., 2025b), a standardized framework for evaluating multimodal models across a wide range of tasks. We instructed all models to respond with a single letter corresponding to their chosen option; when a response requires further parsing, we used the common method that other multiple-choice tasks in `lmms-eval` apply, which checks the response against a set of common answer-formatting patterns to identify the selected option. All open-source model experiments were run on either 4 H100 or 4 A100 GPUs.

C.4 Zero-Shot Evaluation Prompt Template

Every model received the same user prompt, shown below, instantiated per question by inserting the question text and the four answer options. Options were presented in the order stored in the benchmark, which is shuffled once at benchmark-creation time.

Evaluation prompt template

```
Answer the following multiple-choice question using the given video.
Question: {question}
Choices:
A. {option A}
B. {option B}
C. {option C}
D. {option D}
Respond only with the letter of the correct answer.
```

The model's free-form response is then mapped to one of {A, B, C, D} by the standard `lmms-eval` multiple-choice parser, and counted as correct if it matches the correct answer.

C.5 Chain-of-Thought Evaluation Prompt Template

The chain-of-thought (CoT) ablation keeps the original question and answer options unchanged from the zero-shot prompt (Appendix C.4), and replaces the final instruction with the reasoning instruction shown below.

Chain-of-thought prompt template

```
Answer the following multiple-choice question using the given video.
Question: {question}
Choices:
A. {option A}
B. {option B}
C. {option C}
D. {option D}
Before answering, work through the question step by step. First,
identify what the question is asking. Then carefully observe the
video and briefly describe the observations relevant to the question.
Combine these observations to evaluate each option, explaining why it
is or is not supported. Finally, choose the option best supported by
your analysis and end your response with a final line in the format
"Answer: X", where X is the letter of the correct answer.
```

D Additional Results

D.1 Effect of Chain-of-Thought Prompting

Test-time reasoning can also be invoked without a dedicated thinking mode, by instructing a model to reason step by step before answering (Wei et al., 2022). We evaluated five open-source models by extending the original prompt with a CoT instruction that asks the model to observe, reason, then answer (Appendix C.5).

Reasoning step-by-step did not systematically help on SOMA (Table 5). Only Qwen3.8-27B benefited, gaining 3.9 points and improving on 11 of 12 tasks, with the largest gains in temporal reasoning and motion dynamics. For the remaining models, CoT had a neutral effect or reduces performance. This mirrors the thinking-mode result (Figure 4) and suggests that current reasoning mechanisms do not address the difficulty of perceiving body motion, and additional reasoning tokens do little to help.

Table 5: **Effect of chain-of-thought prompting on SOMA.** Accuracy (%) across the three capability axes. For each model, the top row is direct answering (*w/o CoT*) and the bottom row asks the model to reason step by step before answering (*w/ CoT*). The better of the two is marked in **bold**. Task abbreviations follow Table 2.

Models	Setting	Body Configuration				Motion Dynamics			Temporal Reasoning					Avg.
		HU	BC	BP	CBS	RC	MD	OMD	AAI	CO	MEI	NAP	RER	
Qwen3.8-27B	w/o CoT	72.1	47.7	**46.1**	44.1	36.6	40.5	58.6	44.3	34.7	48.3	39.0	37.0	47.0
	w/ CoT	**73.1**	**49.0**	43.5	**49.5**	**41.6**	**49.0**	**59.3**	**51.1**	**46.3**	**52.5**	**40.9**	**40.0**	**50.9**
Gemma-4-31B	w/o CoT	**67.2**	41.1	39.1	**46.8**	**37.6**	34.6	**50.7**	38.2	38.8	30.5	**40.2**	**38.0**	**43.1**
	w/ CoT	58.2	**42.4**	**42.6**	44.1	25.7	**39.2**	47.9	**43.5**	**41.5**	**39.8**	33.5	36.0	42.2
InternVL3.5-38B	w/o CoT	**52.2**	**44.4**	37.4	32.4	**36.6**	30.1	**59.3**	**42.0**	29.9	**45.8**	17.7	**38.0**	**39.0**
	w/ CoT	45.8	36.4	**42.6**	**36.0**	35.6	30.1	48.6	40.5	**32.0**	36.4	**21.3**	32.0	36.5
MiniCPM-o-4.5-9B	w/o CoT	**55.2**	35.8	40.9	**46.8**	**32.7**	29.4	42.9	35.1	**30.6**	26.3	**23.2**	30.0	**36.3**
	w/ CoT	47.8	**37.7**	**47.0**	42.3	31.7	**35.3**	**43.6**	**35.9**	29.9	**27.1**	18.9	**36.0**	36.2
Molmo2-8B	w/o CoT	**54.2**	**42.4**	**51.3**	**41.4**	**33.7**	**36.6**	**39.3**	**39.7**	**38.1**	**37.3**	**36.0**	29.0	**40.6**
	w/ CoT	37.3	32.5	37.4	34.2	30.7	20.9	37.1	30.5	27.2	20.3	26.8	**34.0**	30.8

D.2 Inference-Time Analysis per Capability Axis

D.2.1 Effect of Temporal Sampling Density

Denser frame sampling does not help the three capability axes equally. For Body Configuration tasks, where questions ask about the body's state at a specific moment, accuracy saturates early. A possible explanation is that the relevant evidence for these questions is concentrated in a small number of frames, so additional frames beyond that point add little new information, even though correctly extracting that evidence remains difficult. For Motion Dynamics tasks, which ask how the body moves and are hard to answer from only a few frames, accuracy continues to improve for most models, though only marginally beyond 16 frames. Temporal Reasoning remains the weakest axis at every sampling density, suggesting that regardless of how many frames the model is given, it struggles to integrate that information over time.

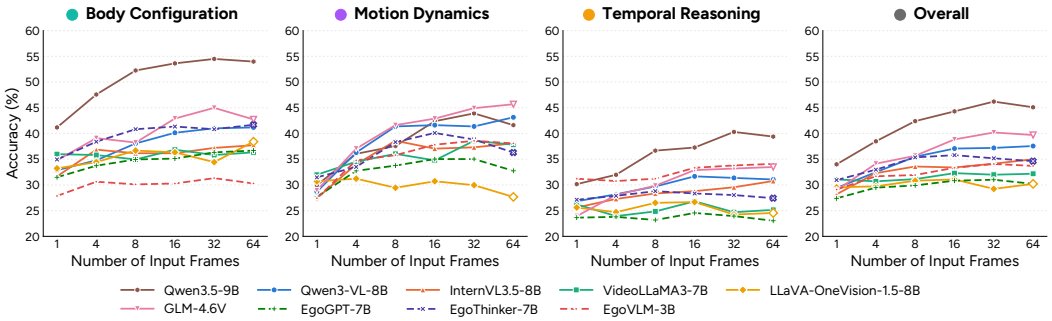


Figure 28: Effect of temporal sampling density on accuracy, per capability axis.

D.2.2 Effect of Test-time Reasoning

The effect of test-time reasoning also varies by axis. Gains are largest on Motion Dynamics, where most models improve; these tasks often require counting or distinguishing observed motion, which can benefit from explicit reasoning once the motion itself has been perceived. On Body Configuration, which is the most perceptual of the three axes, most models show negligible change or a decline. Changes on Temporal Reasoning are modest overall. This suggests that reasoning helps once the relevant evidence has been observed, but does not compensate for a failure to perceive it in the first place.

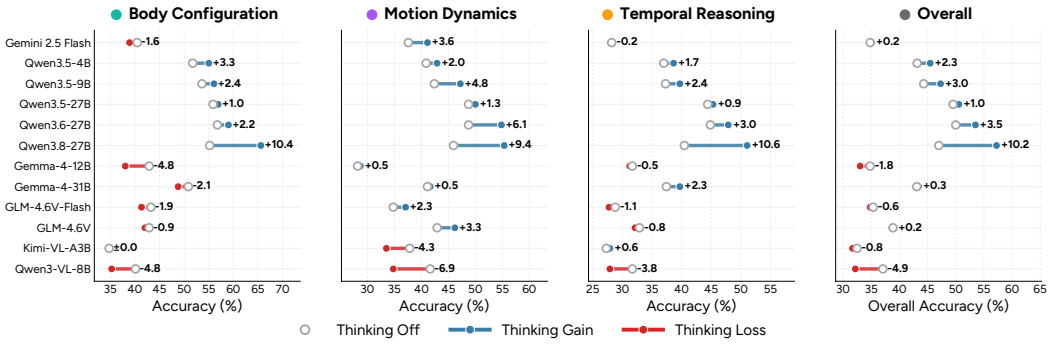


Figure 29: Effect of test-time reasoning on accuracy, per capability axis.

D.3 Performance by Question Difficulty

Questions in SOMA differ in how demanding they are. To understand how models and humans perform across different difficulty levels, we adopted the concept of psychometric curves (Boring, 1917), following Qi et al. (2026). We used Gemini 2.5 Pro to rate each question from 1 (very easy) to 4 (hard) according to how difficult it is to distinguish between the options, and report the accuracy of every evaluated model and our human annotators at each level (Figure 30). Accuracy falls with difficulty for both humans and all models, with humans remaining far ahead at every level. The gap is largest on the easiest questions, showing that models fall short even on simple cases and across the full difficulty range of the benchmark.

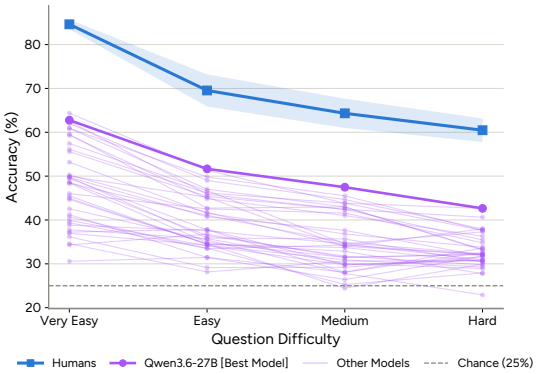


Figure 30: Human and model performance across question difficulty levels. The human curve is the mean over five annotators with the band showing ± 1 SE.

D.4 Error Analysis

Here we provide the full definitions of the four error types used in Section 5.3 and describe the annotation procedure. Each incorrect answer is assigned the single error type that most directly leads to the chosen option:

- **Visual Grounding Error.** The model misidentifies what or who is involved. For example, it confuses or overlooks an object or scene element, or attributes the action to another visible person instead of the camera wearer.
- **Motion and Body-State Perception Error.** The model misperceives what the wearer's own body does, such as their locomotion, movement direction, posture, the limb used, or the action performed, as well as body actions it fails to notice at all.
- **Temporal Localization Error.** The model recognizes the relevant events but places them wrongly in time. For instance, it anchors on the wrong moment or reverses their order.
- **Prior-Driven Inference.** The answer follows from expectation rather than evidence, such as procedural common sense, default assumptions, reinterpreted question or option wording, or assumptions about the dataset, and may rely on observations invented to support that expectation.

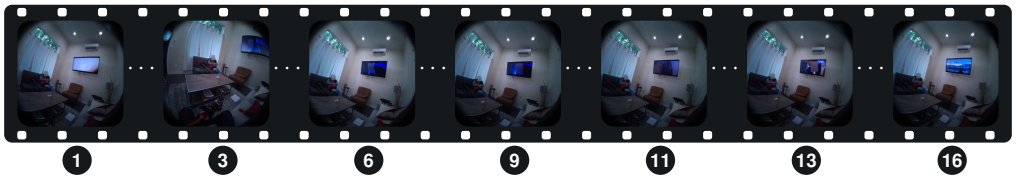
Since manually reading all 400 long reasoning traces is extremely labor intensive, we combined automated labeling with manual verification, following similar approaches in the literature (Yang et al., 2026a; Nagrani et al., 2025). Moreover, every clip in SOMA comes with ground-truth atomic action and motion narration annotations, so each claim in a reasoning trace can be compared with the annotated ground truth. We therefore used LLM judges: given the question, the judge compares the model's reasoning and chosen answer against the ground-truth annotations, together with the correct answer, and assigns one of the four error types, quoting the excerpt of the reasoning that shows the error.

To ensure that the labels do not depend on the biases of a single judge, we used two judges from different model families, GPT-5.1 and Gemini 2.5 Pro, each labeling all 400 traces independently. The agreement between the two judges is high, as they assign the same error type to 80.0% of the traces (Cohen's $\kappa = 0.69$, 95% bootstrap CI 0.63-0.75). We additionally verified a sample of the labels manually against the video clip and ground-truth annotations and found the assigned types and the quoted evidence consistent with the definitions above.

D.4.1 Qualitative Examples

For each error type, we show a representative failure case in Figures 31-34.

Visual Grounding Error



Hand Use · Gemma-4-31B

Which hand does the person use to hold the remote control?

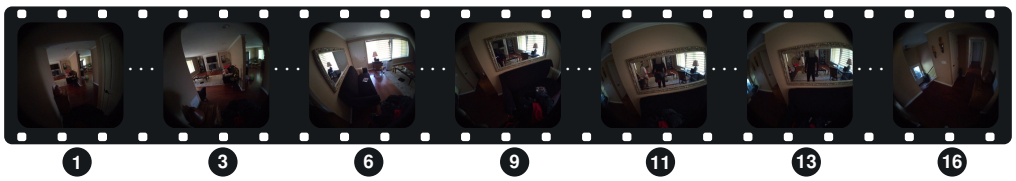
A. Left hand ✓ **B.** Both hands
C. Neither hand **D.** Right hand

Reasoning trace

[. . .] video uses to hold the remote control.
Observation: I am reviewing the provided sequence of images, which are frames from a video captured through a fisheye lens.
Person Analysis: A person is sitting on a couch on the left side of the room.
Remote Control Search: I am scanning the person's hands and the surrounding area (the coffee table, the couch) for a remote control.
Details: Throughout all the frames, the person is sitting still, leaning back on the couch. Their hands are resting in their lap or near their body, but they are not holding any device.
Weighing each option in turn, the model concluded that the person is not holding a remote control in any frame.

Figure 31: **Visual Grounding Error.** Gemma-4-31B never locates the remote control and concludes that no device is held. The failure is in identifying the object, not in reading the body.

Motion and Body-State Perception Error



Opposite Motion Discrimination · Gemma-4-31B

What is the person doing with the clothes?

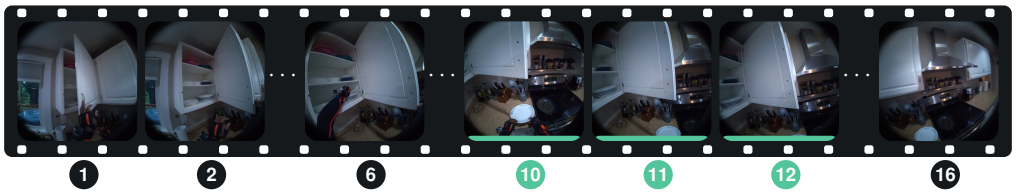
A. Picking up the clothes from the sofa **B.** Folding the clothes
C. Unfolding the clothes **D.** Throwing the clothes on the sofa ✓

Reasoning trace

The question asks what the person is doing with the clothes.
Observation: Looking at the sequence of images, particularly images 6 through 11, a person is seen in a mirror reflection.
Detailed Action: In images 8, 9, and 10, the person (wearing a black shirt) is standing in front of a sofa. They are reaching down towards a pile of clothes that are on the sofa.
Action Analysis: The person's posture and arm movement indicate they are grasping the clothes and lifting them upward from the sofa's surface.
Weighing each option in turn, the model concluded that the person is picking up clothes from the sofa.

Figure 32: **Motion and Body-State Perception Error.** Gemma-4-31B reads the arm movement as lifting the clothes off the sofa when the wearer is throwing them onto it. The direction of the movement is inverted.

Temporal Localization Error



Adjacent Action Identification · Gemma-4-31B

What does the person do immediately before closing the wall cabinet?

A. Puts the plate down on the countertop ✓
B. Puts the oil bottle back on the countertop
C. Grabs a plate from the wall cabinet
D. Picks up the pan on the cooking range

REASONING TRACE

[. . .] is seen in the kitchen performing various tasks.

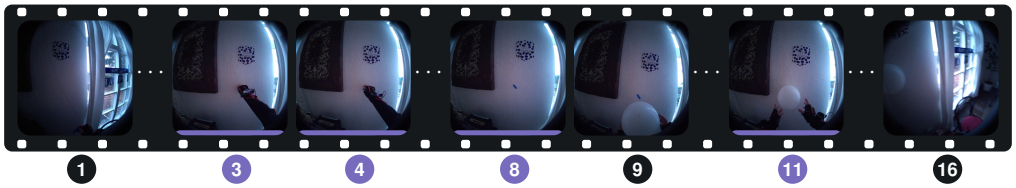
Locate the closing of the wall cabinet: In the latter part of the video, the person opens a white wall cabinet, reaches inside, and then closes it.

Analyze the action immediately preceding the closing: • The person opens the cabinet (frames around image 10). • They reach in and grab a white plate from one of the shelves (image 11). • Immediately after retrieving the plate and bringing it out of the cabinet, they use their other hand to swing the cabinet door shut (image 12).

Weighing each option in turn, the model concluded that the person grabs a plate from the cabinet and then closes the door.

Figure 33: **Temporal Localization Error.** Gemma-4-31B recognizes both the plate and the cabinet but places the grab immediately before the door closes. The events are correct and their order is wrong.

Prior-Driven Inference



Hand Use · Qwen3.8-27B

Which hand does the person use to stick the balloon to the wall?

A. Left hand
B. Both hands ✓
C. Right hand
D. Neither hand

REASONING TRACE

[. . .] the right hand places the tape. • Let's look at the options again. • A. Left hand: Holds the balloon. • B. Both hands: Both are involved. Left holds, Right tapes. • C. Right hand: Applies the tape. • Let's look at the very end of the sticking process (around 0:10-0:11). The left hand is holding the balloon against the wall. The right hand is adjusting the tape. • However, often in these datasets, the "action" is attributed to the hand performing the primary manipulation. The right hand is doing the fine motor task of taping. • The model then re-checked the same frames a further nine times, each pass restating that the right hand applies the tape, and answered *right hand*.

Figure 34: **Prior-driven Inference Error.** Qwen3.8-27B observes both hands acting, then overrides that observation with the expectation that the manipulating hand is the answer.

D.4.2 Right-hand Prior in Laterality Questions

To analyze whether the right-hand bias is shared across models, we analyzed per-class accuracy for every model of Table 2. The bias is widespread, with all but two models, Gemini 2.5 Pro and EgoThinker-7B, above the diagonal and more accurate on *right hand* than on *left hand* (Figure 35). Accuracy on *right hand* questions is high for most models, so most of the spread comes from *left hand* accuracy, which ranges from 9.2% to 80.6%. The asymmetry tends to be smaller for the strongest models, though it does not disappear entirely.

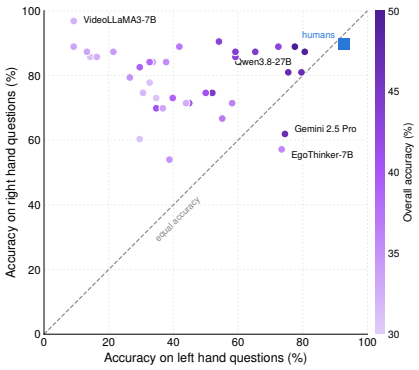


Figure 35: Accuracy on *left hand* versus *right hand* questions, per model.

D.4.3 Counting Error Asymmetry per Model

For each model, we partitioned incorrect predictions on the Repetition Counting questions into *under-counts* (the predicted integer is below the true count) and *over-counts* (the predicted integer is above the true count). We report the resulting shares for all 37 models from Table 2, computed relative to each model's total number of incorrect predictions on this task and ordered from the largest under-count share to the smallest (Figure 36). In total, 24 of the 37 models lean toward under-counting and 12 toward over-counting. The two extremes are Qwen3-VL-8B, which under-counts on 89.7% of its incorrect predictions, and Gemini 2.5 Flash, which over-counts on 84.4%. Beyond the shared sampling bottleneck discussed in Section 5.2, the strength and direction of this bias is highly model-dependent.

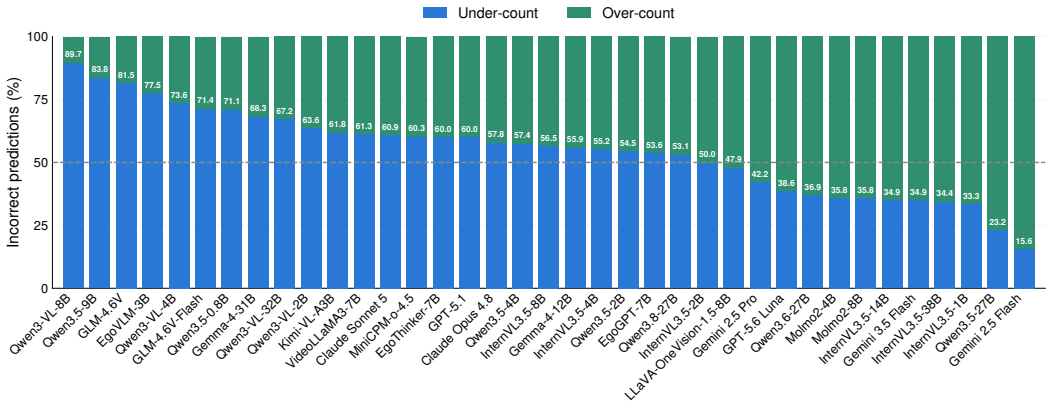


Figure 36: Direction of error on Repetition Counting.

D.5 Evaluation of Motion-Only Baselines

Most motion-only baselines are not designed for multiple-choice question answering and do not reliably output an answer option. We therefore scored the four likelihood-scored models by ranking the four options under the model's own likelihood, and evaluated Motion-SMD through autoregressive decoding. Additionally, the models' motion inputs are out of distribution relative to SOMA, introducing a further limitation. Given both factors, most models scored close to random chance (25%), with average accuracy ranging from 24.9% to 26.8% across the five baselines (Table 6).

E SOMA Train

We first detail the construction of SOMA-Train, covering question generation (Appendix E.1), debiasing (Appendix E.1.2), and the reconstruction of Ego-Exo4D body motion (Appendix E.2). We then describe the architecture and training procedure (Appendix E.3) and additional results (Appendix E.4) of our multimodal baselines.

Table 6: **Performance of Motion-Only Baselines on SOMA.** Accuracy (%) across the three capability axes and their twelve tasks, for models that receive body motion and no video. Scored on the 1,549 questions whose samples carry ground-truth motion. The four likelihood-scored models rank the options directly while Motion-SMD answers autoregressively through its own extractor Task abbreviations follow Table 2.

Models	Body Configuration				Motion Dynamics			Temporal Reasoning					Avg.
	HU	BC	BP	CBS	RC	MD	OMD	AAI	CO	MEI	NAP	RER	
Random Chance	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*
Likelihood-Scored													
MG-MotionLLM (Wu et al., 2025a)	16.3	40.4	33.6	27.3	22.1	20.8	29.4	27.4	20.6	22.8	32.9	24.5	26.4
MotionGPT (Jiang et al., 2023)	23.0	26.7	33.6	30.0	26.3	19.5	31.0	25.8	20.6	28.9	27.2	22.4	26.0
MotionGPT3 (Zhu et al., 2025)	27.5	32.9	20.0	27.3	21.1	27.5	25.4	21.0	11.3	22.8	30.4	27.6	24.9
Motion-Agent (Wu et al., 2025b)	29.8	29.5	25.5	28.2	17.9	35.6	29.4	21.0	7.8	21.9	30.4	19.4	25.2
Generative													
Motion-SMD (Zhang et al., 2026)	37.1	27.4	25.5	17.3	36.8	34.2	18.3	28.2	34.8	19.3	15.8	22.4	26.8

E.1 QA Generation

Our training dataset, SOMA-Train, comprises aligned egocentric video and body motion, designed to investigate whether body-motion supervision helps close the gaps exposed by SOMA. We followed the taxonomy of the benchmark and drew on the same two sources, Nymeria and Ego-Exo4D. In total, SOMA-Train contains 28,596 questions from 1,779 takes (570 from Nymeria and 1,209 from Ego-Exo4D), covering 110.2 hours of aligned egocentric video and body motion.

Unlike the evaluation setting, where we removed questions answerable from language priors alone, the training setting introduces an additional failure mode: shortcuts that cannot be exploited zero-shot may become learnable during fine-tuning. We therefore adapted the generation pipeline for SOMA-Train by controlling distractor construction per task, so that the correct option cannot be identified from the question and answer choices alone. When this cannot be guaranteed by construction, we additionally applied adversarial filtering. Together, these steps target training-time bias, i.e., shortcuts learnable from the training distribution itself.

We present the pipeline in two parts. We first describe the construction of SOMA-Train in four steps: question templating, task-specific extraction, distractor construction and balancing, and a final grammar and factuality verification (Appendix E.1.1). We then test the resulting dataset for training-time bias by measuring whether a model trained without video or motion input transfers to SOMA (Appendix E.1.2).

E.1.1 Construction

Question Templates. For each question, we randomly sampled one of the task-specific templates (Table 7), which follow the question formats of the corresponding SOMA tasks, and fill it with fields extracted from the clip annotation. In the templates, {a} denotes an anchor clause from the annotation (for Missing Event Inference, {a} and {c} denote the surrounding events), {b} the referenced event, {o} the parsed object, {l} the limb pair, {e} the repeated event, and {e1} to {e4} the listed actions. When a template has several grammatical variants, we selected the one that matches the form of the extracted anchor.

Extraction. The benchmark pipeline uses an LLM selector to identify task-relevant windows in the source annotations. For SOMA-Train, we replaced this step with deterministic task-specific regex-based extractors, which make the extraction fully reproducible and reduce potential sampling biases introduced by LLM-based selection. Each extractor follows phrasing conventions observed in the source annotations and includes task-specific rejection rules for failure cases that we identified throughout the process.

Table 7: **SOMA-Train Question Templates.** The full inventory, grouped by capability axis. One template is drawn at random per question and filled from the clip's own annotation.

Task	Templates
Body Configuration	
Hand Use	Which hand does the person use in this clip? In this clip, which hand performs the action? Which hand is the person using here? During this segment, which hand does the person use? Which of the following describes the person's hand use in this clip? Which hand carries out the action in this clip?
Bimanual Coordination	Which arm does the distinctive movement in this clip? Which of the person's arms is more actively moving? In this clip, which arm carries the notable movement? Which arm stands out as the moving one during this segment? Over this clip, which arm does the distinctive action? Which of the following describes the arms' roles in this clip? During this segment, which arm is the active one? Which arm is doing the distinctive movement here?
Concurrent Body-Part State	While the person is {a}, what are their {l} doing? What are the person's {l} doing while they are {a}? As the person is {a}, what are their {l} doing? When the person {a}, what are their {l} doing? What are the person's {l} doing while the person {a}? While the person {a}, what are their {l} doing?
Body Posture	How many distinct body-posture states does the person move through in this clip, and what are they? What is the sequence of body postures the person moves through in this clip? What body posture is the person in at the end of the clip? Over the course of this clip, which of the following posture changes occurs?
Motion Dynamics	
Opposite Motion Discrimination	What is the person doing with the {o}? What does the person do with the {o} in this clip? In this clip, what is the person doing with the {o}? Which of the following is the person doing with the {o}?
Movement Direction	After the person {a}, what is the sequence of their next two movements? After the person {a}, what are their next two movements? What is the sequence of the person's next two movements after they {a}? After {a}, what is the sequence of the person's next two movements? After {a}, what are the person's next two movements? After {a}, which direction does the person move?
Temporal Reasoning	
Adjacent Action Identification	What does the person do immediately before {b}? Just before {b}, what does the person do? Which action comes immediately before {b}? What happens right before {b}?
Missing Event Inference	What action is missing from this sequence: {a} → [?] → {c}? Which action belongs in the gap: {a} → [?] → {c}? The middle step is missing: {a} → [?] → {c}. What was it? Fill the gap in this sequence: {a} → [?] → {c}.
Chronological Ordering	What is the correct temporal order of these three actions: [{e1} / {e2} / {e3}]? We also use a four-action variant.
Repeated Event Reasoning	The person {e} multiple times in this clip. What does the person do right after the FIRST/SECOND/LAST time they {e}?

Distractors Sampling. As the benchmark follows a multiple-choice format, the construction of distractor options is crucial. We designed the distractors separately for each task: we aimed to balance the options by construction, making sure that no candidate is more likely to be correct than another based on the textual information and task-specific category structure alone. We denote the ground-truth answer by g, an option string by s, a conditioning context by c, and the number of candidates within that context by k, and use the following strategies. A closed option set presents the same four answer options for every question, such that $P(g \mid s) = 1/4$ (Hand Use). Product construction forms a fixed set of candidate combinations while keeping the shared movement clause identical across options (Movement Direction). Rotation within a conditioning context fixes k candidate actions and rotates the correct answer such that $P(g \mid \text{offered}, c) = 1/k$ within each context (Opposite Motion Discrimination). Exchangeable option sets are formed from four answers from the same take, with the correct answer rotated so that each appears once as the ground truth and three times as a distractor (Bimanual Coordination, Concurrent Body-Part State, Missing Event Inference). When balance cannot be guaranteed by construction, we applied adversarial filtering following Le Bras et al. (2020) (Chronological Ordering).

Grammar and Factuality Corrections. Template- and regex-based question generation can introduce systematic grammatical and wording artifacts, e.g., incorrect verb forms, subject-verb agreement errors, missing subjects, truncated options, or spelling errors inherited from the source annotations. These artifacts can also introduce shortcuts when they correlate with the correct option or alter the meaning of an answer.

We used Gemini 3.8 Flash (Google DeepMind, 2026b) to correct these grammatical and wording artifacts while preserving the original semantic content and avoiding the introduction of new answer-specific biases. We additionally used the same model to verify the generated questions against the original human annotations, correcting or removing examples with remaining inconsistencies.

E.1.2 Debiasing Validation

To validate our approach, we measured whether a model trained without modality input (motion and/or video) could still exploit a learnable signal from the training dataset. Ideally, its performance should remain close to the random baseline, indicating that successful fine-tuning requires relying on task-relevant perceptual information rather than on shortcuts in the training distribution. To this end, we trained the model on the full SOMA-Train set with text input only and evaluated it "blind" on SOMA (Table 8). We found that blind performance remained close to the random baseline, while the gap between the video+motion variant and the blind model reached +16.65 points. This provides evidence that the observed fine-tuning gains rely on perceptual input rather than on language understanding and distributional shortcuts alone.

Table 8: **Blind Transfer to SOMA** (Qwen3-VL-8B, accuracy %). Each row fine-tunes the same LoRA recipe on SOMA-Train and is evaluated on SOMA with and without perceptual input (motion, video). $\Delta_{\text{I}-\text{B}}$ is the modality–blind gap.

Training Input	Input	Blind	$\Delta_{\text{I}-\text{B}}$
None (blind)	–	28.14	–
Motion	28.21	25.60	*+2.60*
Video	41.49	27.38	*+14.10*
Video+Motion	42.44	25.79	***+16.65***
Chance (uniform)		*25.00*	

E.2 Ego-Exo4D Refitting

SOMA-Train pairs each question and its video clip with an aligned body-motion sequence. This requires Ego-Exo4D (Grauman et al., 2024) motions to follow the same body representation and frame rate as Nymeria (Ma et al., 2024). The original Ego-Exo4D human annotations are sparse, providing 17-joint keypoints rather than the SMPL-X representation (Pavlakos et al., 2019) used by our motion pipeline, without joint-angle information. EE4D-Motion (Patel et al., 2025) provides SMPL-X fits, but at 10 fps rather than our target 20 fps. Its filtering pipeline additionally covers

only 2,592 of the 3,634 non-climbing takes, leaving 12.9% of our Ego-Exo4D benchmark questions without body motion. We compared against the released EE4D-Motion dataset (Patel et al., 2025).

We therefore reprocessed Ego-Exo4D to obtain denser SMPL-X fits. We followed the multi-view fitting pipeline of EE4D-Motion (Patel et al., 2025), which detects the camera wearer in the exocentric views, estimates 2D keypoints, initializes SMPL-X from a single-view reconstruction, and refines the sequence through multi-view reprojection. We kept the same multi-view approach but replaced the single-view pose estimation and reconstruction stages with SAM 3D Body (Yang et al., 2026b). We provide further details below:

- **Detection and Subject Selection.** We detected people with RF-DETR (Robinson et al., 2025) and identified the camera wearer by projecting the Aria trajectory into each exocentric view and selecting the nearest detection. We re-detected when the tracked box moved too far from the projected Aria position. We set this threshold independently for each camera to 2.5 times its median head-to-anchor distance, clamped to $[60, 250]$ px. This trajectory-based association is important in scenes containing multiple people, where tracking based only on the detected person can switch to a bystander.
- **Per-view Pose Estimation.** We applied SAM 3D Body independently to each selected view. The model jointly predicts 2D and 3D keypoints, body parameters, and camera parameters, allowing both the keypoints and body initialization to come from the same model. We ran this stage at 20 fps.
- **Triangulation.** We undistorted the predicted keypoints using each camera's Kannala–Brandt fisheye model and triangulated them using the Ego-Exo4D GoPro calibration. We required at least two valid views and removed joints with a reprojection residual above 60 px. We used the calibration stored in the take trajectory.
- **Multi-view Fitting.** We jointly optimized one SMPL-X sequence across all available views, with body shape shared across the take. Triangulated joints initialize the global orientation and translation. The objective combines a robust reprojection loss over the 34 matched joints, agreement with the per-view SAM 3D Body predictions, temporal smoothness, and pose and shape regularization. We additionally removed cameras whose median distance to the selected subject exceeds three times that of the closest camera.

Evaluation. We evaluated our fits and EE4D-Motion against the human-verified Ego-Exo4D annotations, which provide 3D keypoints triangulated from manually annotated exocentric views. We regressed both SMPL-X sequences to the same joint set before evaluation. As SAM 3D Body derives its output keypoints from the MHR model (Ferguson et al., 2025), while Ego-Exo4D follow the 17-keypoint COCO convention (Grauman et al., 2024), the error may also include a systematic convention mismatch. We corrected this by estimating one constant offset per joint on the training takes and applying the same correction to validation for both methods. To validate our pipeline, we report the reconstruction error on the intersection of takes covered by the annotations, EE4D-Motion and our pipeline, averaging first within each take and then across takes. We evaluated with PA-MPJPE because EE4D-Motion refits are defined independently for each window and do not preserve a common world coordinate system, which prevents a direct comparison of global trajectories.

Results. Our fits reached 36.7 mm PA-MPJPE on validation, compared with 40.3 mm for EE4D-Motion (Table 9). The joint correction reduced the errors by 5.9 and 3.9 mm, moving our result from 42.6 to 36.7 mm and EE4D-Motion from 44.2 to 40.3 mm, respectively, without changing their ordering. We report the training results for completeness, although the offsets are fitted on the training takes.

Coverage. Following EE4D-Motion, we excluded Rock Climbing because our motion representation expresses height relative to the floor. Although this removes 1,401 of the 5,035 takes, these account for only 17.1 of 221.3 video hours. Of the remaining 3,634 takes, our pipeline fitted 3,475, corresponding to 190.7 of 204.1 hours. The remaining 159 takes have missing source data.

We measured coverage over the original 30 fps video timeline while accounting for the native sampling rate of each motion source. Our fits covered 93.1% of the training video and 96.2% of validation, compared with 70.5% and 73.1% for EE4D-Motion (Table 10).

Table 9: **Ego-Exo4D Refitting Quality.** PA-MPJPE (mm, take-averaged) against Ego-Exo4D's human annotations, over the non-climbing takes common to those annotations, EE4D-Motion and our fits: 827 train and 212 val, of which the 792 and 194 scored here pass the per-frame filters. *Corrected* adds the per-joint convention offset, fitted on train.

	Validation (194 takes)		**Train** (792 takes)	
	Uncorrected	Corrected	Uncorrected	Corrected
EE4D-Motion	44.2	40.3	45.9	41.8
Ours	**42.6**	**36.7**	**45.7**	**39.6**
Δ	-1.6	-3.6	-0.2	-2.2

Table 10: **Ego-Exo4D Coverage.** Share of non-climbing takes and video time carrying body motion, framewise on the 30 fps grid. †At 10 fps, equal video time carries half the samples.

			EE4D-Motion†		**Ours**	
Split	**Takes**	**Hours**	Takes	Video time	Takes	Video time
Train	2,232	119.6	91.0	70.5	**95.9**	**93.1**
Val	597	35.4	93.8	73.1	**97.2**	**96.2**

At the target 20 fps rate, our fits covered on average 96.6% of each SOMA segment, compared with 37.9% for EE4D-Motion, and 96.3% of each SOMA-Train segment, compared with 48.1% (Figure 37). The lower frame-level coverage of EE4D-Motion partly comes from its 10 fps sampling rate. After interpolating its fits to 20 fps, our pipeline still improved mean SOMA coverage by 20.8 points. On SOMA-Train, the coverage was similar because all Ego-Exo4D questions were generated from the 1,209 takes already covered by EE4D-Motion. Finally, we resampled all fitted sequences to 20 fps and converted them to the 263-dimensional body-motion representation (Guo et al., 2022) used throughout our experiments.

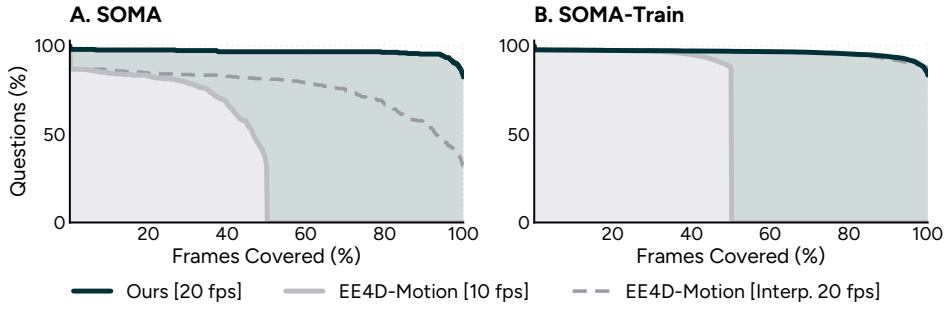


Figure 37: **Per-Question Body-Motion Coverage.** Percentage of Ego-Exo4D questions with at least $x\%$ of their frames covered by fitted body motion at the target 20 fps rate. The dashed curve interpolates the 10 fps EE4D-Motion fits to 20 fps.

E.3 Training Details

Architecture. We extended two open-weight backbones, Qwen3-VL-8B (Bai et al., 2025) and InternVL3.5-8B (Wang et al., 2025), with a body-motion pathway following the standard LLaVA/-MotionLLM design (Liu et al., 2023; Chen et al., 2025; Qi et al., 2026). A motion encoder maps each motion sequence to a sequence of continuous features, which a trainable MLP projects into the language-model embedding space. The projected features are inserted into the input sequence as motion tokens.

Motion encoder. We used the convolutional encoder of a VQ-VAE (Zhang et al., 2023) that operates on 263-dimensional body-motion features (Guo et al., 2022). The VQ-VAE was trained for reconstruction on 156,659 motion clips from the SOMA-Train takes of Nymeria and Ego-Exo4D, sampled as 64-frame windows at 20 fps. We used its continuous pre-quantization features, one per 4 frames. The encoder was kept frozen throughout all subsequent training steps.

Training overview. We compared video-only, motion-only, and video+motion variants. Motion-based variants were trained in two stages: we first aligned the motion projector on motion–narration pairs, then fine-tuned the language model on SOMA-Train with LoRA (Hu et al., 2022). We skipped Stage 1 for the video-only variant since the backbones already include a pretrained vision encoder and projector, and fine-tuned each backbone on SOMA-Train using video input only.

Stage 1: Motion–language alignment. We aligned the motion modality using the motion-narration annotations of the Nymeria takes in SOMA-Train (Ma et al., 2024). We rewrote each narration into three body-part descriptions (legs and feet, arms and hands, overall posture), each treated as a separate target, resulting in 24,549 samples from 8,183 clips. The model received motion as the only input, and only the projector was trained while the language model remained frozen. We trained for one epoch with a learning rate of 10^{-3}, a cosine schedule with a warmup ratio of 0.03, and a global batch size of 16.

Stage 2: Question answering. All variants were fine-tuned on SOMA-Train with LoRA (rank 128, $\alpha = 256$, dropout 0.05) applied to the language model. For motion-based variants, we initialized from the Stage 1 projector and kept it frozen. We trained for two epochs with a LoRA learning rate of 5×10^{-5}, using the same schedule and warmup as Stage 1, gradient clipping at 1.0, and bf16 precision with gradient checkpointing. Training ran on 8 H100 GPUs with a per-device batch size of 6 (global batch size 48).

E.4 Additional Results

To analyze how the contribution of motion varies across task types, we examined per-task results (Table 11). Adding motion to video improves both backbones on Body Posture, Opposite Motion Discrimination, Missing Event Inference, and Next Action Prediction, while hurting both on Hand Use and Adjacent Action Identification. This suggests that the effect of motion depends largely on what each task requires: motion helps when body state or movement carries information that is hard to recover from video alone. When the evidence is already visible in the video, however, motion tokens add little new information and may instead introduce redundant or noisy signals that distract the model from the relevant visual evidence.

Table 11: **Fine-tuning with Body-Motion Input, Per Task.** Accuracy (%) on SOMA after fine-tuning on SOMA-Train, by input modality. Task abbreviations follow Table 2. †Zero-shot rows use the same video preprocessing as the fine-tuned models. Best per backbone in **bold**.

Setting	Body Configuration				Motion Dynamics			Temporal Reasoning					Avg.
	HU	BC	BP	CBS	RC	MD	OMD	AAI	CO	MEI	NAP	RER	
Random Chance	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*
Qwen3-VL-8B													
Zero-Shot†	47.0	28.1	38.2	25.5	**37.0**	34.2	51.1	35.7	33.6	32.2	23.3	36.7	35.4
+ Motion	19.7	24.0	38.2	33.6	23.0	25.5	34.8	29.4	33.6	25.2	26.4	30.6	28.2
+ Video	**57.9**	**34.2**	46.4	36.4	31.0	**38.9**	58.5	**38.9**	**39.2**	47.0	25.2	39.8	41.5
+ Video+Mot.	53.0	**34.2**	**51.8**	**41.8**	32.0	**38.9**	**63.0**	35.7	**39.2**	**47.8**	**28.9**	**41.8**	**42.4**
InternVL3.5-8B													
Zero-Shot†	37.2	30.1	**51.8**	38.2	**39.0**	36.2	45.2	38.1	21.7	34.8	**26.4**	28.6	35.2
+ Motion	27.3	21.9	41.8	38.2	26.0	28.9	29.6	28.6	33.6	29.6	24.5	29.6	29.5
+ Video	**58.5**	28.8	47.3	**42.7**	30.0	**44.3**	55.6	**40.5**	29.4	45.2	24.5	**50.0**	41.4
+ Video+Mot.	52.5	**35.6**	50.0	**42.7**	26.0	**44.3**	**56.3**	34.9	**40.6**	**47.0**	25.8	45.9	**41.9**

To further study how each modality contributes at inference, we evaluated the same Video+Mot. checkpoint with either video or motion removed at test time (Table 12). Unlike Table 11, where each setting corresponds to a separately fine-tuned model, all rows here share the same weights. Removing the video decreases the average accuracy by 13.3 points for Qwen3-VL-8B and 14.0 points for InternVL3.5-8B, showing that both backbones rely primarily on the visual input. On the other hand, removing body motion has a much smaller effect: Qwen3-VL-8B performance decreases by 1.4 points while InternVL3.5-8B improves by 0.8 points. As in Sec. 5.4, the effect varies across tasks: removing motion decreases Body Posture accuracy from 51.8 to 40.9 for Qwen3-VL-8B

and from 50.0 to 33.6 for InternVL3.5-8B, but increases Hand Use accuracy by 3.3 and 9.8 points, respectively. Motion therefore helps when the body state is hard to recover from the visual signal alone, but adds little or can even hurt when the relevant evidence is already visible.

Table 12: **Test-Time Modality Ablation.** Accuracy (%) of the same jointly fine-tuned Video+Mot. checkpoint for each backbone when evaluated with different input modalities at test time. Task abbreviations follow Table 2; Δ denotes the change in Avg. relative to Video+Mot.

Input at inference	Body Configuration				Motion Dynamics			Temporal Reasoning					Avg.	Δ
	HU	BC	BP	CBS	RC	MD	OMD	AAI	CO	MEI	NAP	RER		
Random Chance	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	
Qwen3-VL-8B, fine-tuned on video + motion														
Video + Motion	53.0	34.2	51.8	41.8	32.0	38.9	63.0	35.7	39.2	47.8	28.9	41.8	42.4	—
Video	56.3	34.9	40.9	33.6	36.0	32.9	60.0	40.5	42.7	47.0	26.4	35.7	41.0	−1.4
Motion	25.1	27.4	40.9	30.9	27.0	31.5	33.3	32.5	29.4	29.6	20.1	25.5	29.1	−13.3
InternVL3.5-8B, fine-tuned on video + motion														
Video + Motion	52.5	35.6	50.0	42.7	26.0	44.3	56.3	34.9	40.6	47.0	25.8	45.9	41.9	—
Video	62.3	39.0	33.6	34.5	29.0	46.3	55.6	41.3	42.7	45.2	26.4	46.9	42.7	+0.8
Motion	27.3	18.5	37.3	36.4	19.0	22.8	25.9	27.0	34.3	32.2	25.8	32.7	27.9	−14.0

F LICENSING INFORMATION

All artifacts are used within the terms of their respective licenses and for research purposes consistent with their intended use (Table 13). Our benchmark questions will be released for non-commercial research use, consistent with the source data licenses.

Table 13: Third-party artifacts used in this work and their licenses.

Artifact	License
Datasets	
Nymeria (Ma et al., 2024)	CC BY-NC 4.0
Ego-Exo4D (Grauman et al., 2024)	Ego-Exo4D Model License
Evaluation framework	
lmms-eval (Zhang et al., 2025b)	MIT and Apache 2.0