HUMANMOVEVQA: CAN VIDEO MLLMS REASON ABOUT HUMAN MOVEMENT IN VIDEOS?

Anonymous authors
Paper under double-blind review

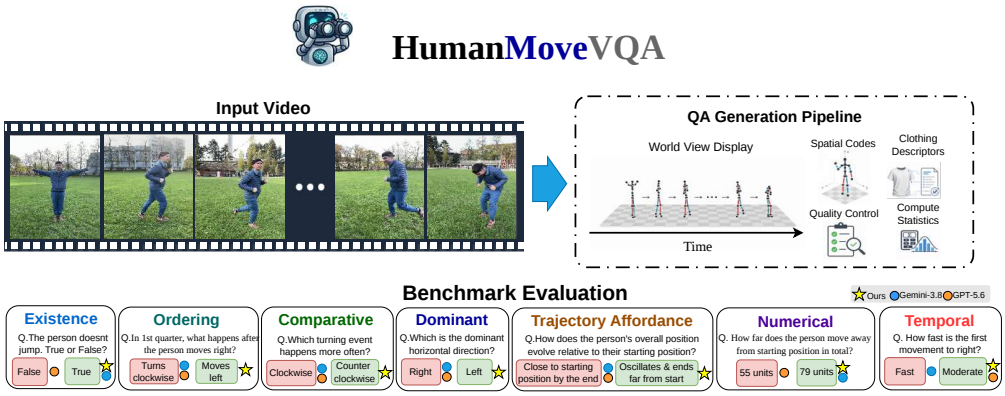


Figure 1: Overview of HumanMoveVQA. 3D human motion tracks, anchored to the person's first-frame position and heading, are used to generate multiple-choice questions in seven categories: Existence, Ordering, Comparative, Dominant, Trajectory Affordance, Numerical, and Temporal. Example predictions from Gemini-3.8-flash, GPT-5.6-Luna and our fine-tuned Qwen3-VL 4B are shown (green: correct, red: incorrect); aggregate results are in Table 4.

ABSTRACT

Multimodal Large Language Models (MLLMs) can describe what a person does in a video, but not where they go. Answering "does the person end up left of where they started?" or "how many times do they turn clockwise?" requires reasoning about global trajectory and orientation in 3D over time, beyond apparent motion in the image, which existing benchmarks, focused on scene-level events or local joint articulation, do not probe. We introduce HumanMoveVQA, a benchmark for reasoning about global human trajectory and orientation in exocentric video. We propose a scalable multi-stage pipeline that lifts 2D videos from five human-motion datasets into world-consistent 3D motion SMPL-X tracks anchored to the person's first-frame position and heading, and converts them into discrete motion events. From these, we generate 12,282 multiple-choice questions over 127 test videos (15 to 90 s) across seven categories spanning event detection, aggregation, sequencing and trajectory-level inference. The best of 12 Video MLLMs reaches a chance-normalised score of 14.0, against 78.7 for humans. Fine-tuning an open-source model on our world-consistent supervision roughly triples its score, showing the problem is learnable. Code and data are available here.

1 INTRODUCTION

A caption such as *"a person playing tennis"* collapses distinct physical sequences into one label: a rapid lateral sprint for the ball and a slow walk to the net become indistinguishable. Existing video-caption data largely omit information about **where a person moves, how their trajectory evolves, and how their orientation changes over time**, limiting MLLMs in sports analytics, robotics, navigation, and industrial safety. Recovering this information from video poses three coupled challenges: (i) *disentanglement*: a moving camera mixes subject motion with camera motion, while motion along the camera's viewing axis is difficult to observe in 2D, requiring motion to be recovered in world space rather than camera space; (ii) *a fixed reference*: "left" or "moved away" is only meaningful relative to a fixed origin; and (iii) *long-horizon aggregation*: net displacement,

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.

turn counts, and trajectory shape emerge only by integrating many short events over tens of seconds, beyond the 3 to 15 s clips of most existing motion benchmarks. To address all three, we recover motion in 3D and express it relative to the person's first-frame position and heading, making answers independent of viewpoint.

Recent MLLMs (OpenAI, 2026; Google DeepMind, 2026; Bai et al., 2025; Wang et al., 2025a) perform strongly on high-level video tasks (Yu et al., 2019; Fu et al., 2025; Xiao et al., 2021; Wu & Yu, 2024; Xu et al., 2016; Wu et al., 2024; Wang et al., 2024; Ma et al., 2023), but are trained on video-text pairs (Zhu et al., 2023; Bain et al., 2021) whose captions describe events at a high level. Human motion benchmarks (Peng et al., 2025; Li et al., 2025; Chen et al., 2025; Hong et al., 2025) focus on canonical joint-level motion over clips of 3 to 15 s, rather than tracking global root trajectory and orientation relative to a fixed origin, while spatial reasoning work (Ma et al., 2025; Chen et al., 2024; Batra et al., 2025) studies object relations in static images. This leaves an open question: *can Video MLLMs reason about human trajectory and orientation in space over time?* By reasoning, we mean inferring quantities that no single frame shows, such as net displacement relative to the person's initial heading, rather than recognising a directly visible motion pattern.

We introduce **HumanMoveVQA**, a benchmark for global trajectory and orientation reasoning about people in video, with questions in seven categories grouped into event detection (Existence), aggregation (Comparative, Dominant, Numerical), sequencing (Ordering, Temporal) and trajectory-level inference (Trajectory Affordance). To build HumanMoveVQA at scale, we propose a novel pipeline that recovers world-space 3D SMPL-X tracks from video, converts them into discrete motion events, and generates question-answer pairs from these events deterministically. This approach enables us to capture continuous human movement in a consistent 3D space over extended sequences, a capability missing from existing benchmarks. We utilise five human-motion datasets, EMDB (Kaufmann et al., 2023), RICH (Huang et al., 2022), EgoBody (Zhang et al., 2022), 3DPW (von Marcard et al., 2018) and SLOPER4D (Dai et al., 2023), covering static and moving cameras, indoor and outdoor scenes, and single- and multi-person settings, which yield 86,671 training questions from 1,111 videos and 12,282 test questions over 127 held-out videos of 15 to 90 s.

Evaluating 12 Video MLLMs on HumanMoveVQA reveals a large gap (Table 4): the best zero-shot model, Qwen3-VL 8B, reaches a chance-normalised score of 14.0, and the proprietary Gemini-3.8-flash 10.2, against 78.7 for humans. Fine-tuning on our training split triples the score of Qwen3-VL (12.3 to 39.0 for the 4B model), yet event ordering remains near chance. Our contributions are:

- **Benchmark.** HumanMoveVQA, which probes global human trajectory and orientation in 3D over long exocentric videos with static and moving cameras, in a first-frame anchored reference frame.
- **Scalable, verifiable pipeline.** World-space reconstruction, discrete motion events and template-based generation, validated against motion-capture ground truth and giving 87K training questions.
- **Findings.** (i) Video MLLMs, including proprietary models, perform far below humans, and joint-level motion supervision (MotionLLM) falls below chance; (ii) world-consistent supervision roughly triples performance and transfers across datasets; (iii) temporal ordering of motion events remains an open challenge.

2 Related Work

Reasoning about how a person moves requires the three capabilities of Section 1: recovering world motion under a moving camera, expressing it relative to a fixed reference frame, and aggregating it over long sequences. We review prior work against these requirements for a person observed from an exocentric viewpoint.

2.1 Video MLLMs and Motion-Aware Models

Video MLLMs such as Video-LLaVA (Lin et al., 2024), VideoGPT+ (Maaz et al., 2024), LLaVA-NeXT-Video (Zhang et al., 2024b), InternVL3.5 (Wang et al., 2025a) and LLaVA-OV-2 (An et al., 2026) achieved significant results in general video understanding. Their image-text and video-text training data (Zhang et al., 2024a; Bain et al., 2021) carry little motion information: captions describe *what* happens (*e.g.*"a person walks across the room") but not *how far*, *in which direction* or *whether the person turned*, a gap we confirm in Section 4.

MotionLLM (Chen et al., 2025) and subsequent work (Xu et al., 2024; Hong et al., 2025; Peng et al., 2025) add fine-grained motion supervision, and PoseScript (Delmas et al., 2022) and ChatPose (Feng et al., 2024) align SMPL poses with language. These methods operate in a root-normalised body

Table 1: Comparison with representative video-understanding benchmarks. *World Motion with Dynamic Camera* denotes world-frame motion under camera movement; *Reference-Frame Motion* denotes displacement or heading relative to a fixed reference; and *Long Horizon* denotes an average clip length of at least 30 s. HumanMoveVQA is the only benchmark that satisfies all three requirements for a person observed from an exocentric (third-person) viewpoint.

Benchmark	World Motion with Dynamic Camera	Reference-Frame Motion	Long Horizon	Human -Centric	Exo -centric	Trajectory /Path	Directional Reasoning	Event Ordering	Spatial Reasoning	#Videos	Duration (s)	#QA Pairs
General Video Understanding												
MVBench (Li et al., 2023)	✗	✗	✗	✗	✓	✗	✓	✓	✗	3,641	∼20	4,000
Spatial Reasoning												
MLLM-4D (Yin et al., 2026)	✓	✗	✗	✗	✓	✓	✓	✗	✓	3,600	∼4.28	6,000
4D-RGPT (Yang et al., 2025)	✗	✗	✗	✗	✓	✗	✗	✗	✓	–	∼4	1,500
VSI-Bench (Yang et al., 2024)	✗	✗	✓	✗	✗	✓	✓	✗	✓	288	∼120	5,000+
SAW-Bench (Li et al., 2026a)	✓	✓	✓	✓	✗	✓	✓	✗	✓	786	30–40	2,071
Human Motion Understanding												
ActionArt (Peng et al., 2025)	✗	✗	✗	✓	✓	✗	✗	✗	✗	675	∼8.6	2,678
MotionBench (Hong et al., 2025)	✗	✗	✗	✗	✓	✗	✗	✓	✗	5,385	<10	8,052
MotionLLM (Chen et al., 2025)	✗	✗	✗	✓	✓	✗	✗	✗	✗	1,350	∼10	1,400
MMHU (Li et al., 2025)	✓	✗	✗	✓	✓	✗	✗	✗	✗	840	∼3	∼21.8K
LLaVAction (Qi et al., 2026)	✗	✗	✗	✓	✗	✗	✗	✗	✗	∼138	∼3	9,668
Lemonade (Bonnetto et al., 2025)	✗	✗	✓	✓	✗	✗	✗	✓	✗	49	∼60	36,521
HumanMoveVQA (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	**127**	**15–90**	**12,282**

frame that removes global translation and heading, so they fail to capture a subject's position and orientation changes over time, that HumanMoveVQA evaluates.

Other works use 3D information into MLLMs through point clouds, depth or 3D-aware visual tokens (Zhu et al., 2025; Zheng et al., 2025; Wu et al., 2025), but reason about static scene geometry rather than the world-space motion of people. Our pipeline instead builds on world-grounded human mesh recovery (Wang et al., 2025b) and on MotionScript (Yazdian et al., 2025), which converts motion into discrete, language-aligned codes.

2.2 Video and Spatio-Temporal Benchmarks

Video benchmarks target high-level semantics: question answering (Yu et al., 2019), retrieval (Xu et al., 2016), causal and temporal reasoning (Xiao et al., 2021; Wu & Yu, 2024; Li et al., 2023) and long-form understanding (Fu et al., 2025; Wu et al., 2024). They treat video as a sequence of discrete events and do not probe how a person moves through space. Spatial reasoning benchmarks have moved from object relations in static images (Chen et al., 2024; Cheng et al., 2024; Ma et al., 2025) to spatio-temporal (4D: 3D space plus time) settings (Yin et al., 2026; Yang et al., 2024; 2025), but remain scene-centric: VSI-Bench (Yang et al., 2024) evaluates spatial understanding of 3D environments from egocentric video, while SAW-Bench (Li et al., 2026a) additionally reasons about the camera wearer's own trajectory, whereas HumanMoveVQA tracks the trajectory and orientation of an observed person from an exocentric viewpoint.

2.3 Human Motion Benchmarks

Human motion benchmarks focus on body articulation and action analysis. MotionBench (Hong et al., 2025) and MotionLLM (Chen et al., 2025) evaluate fine-grained body motion, and LLaVAction (Qi et al., 2026) evaluates recognition of egocentric object-manipulation actions in kitchens. ActionArt (Peng et al., 2025), the closest to our work, covers action categories, human-object interaction and body articulation, with only coarse global spatial questions. MMHU (Li et al., 2025) provides exocentric pedestrian trajectories but uses them only to classify behaviour (*e.g.*, crossing or waiting), whereas HumanMoveVQA asks how far and in which direction a person moves relative to their start. Lemonade (Bonnetto et al., 2025) derives detailed kinematics from calibrated multi-view capture, but asks only about hands in an egocentric setting. Most of these benchmarks and the underlying motion datasets (Lin et al., 2023; Hong et al., 2025) use clips of 3 to 15 s, whereas HumanMoveVQA is built from clips of 15 to 90 s. To our knowledge, HumanMoveVQA is the only benchmark that meets all three requirements (Table 1): it evaluates the root trajectory and orientation of a person observed from a third-person viewpoint, over clips of 15 to 90 s, with answers derived from world-space 3D tracks anchored to the first frame.

3 HumanMoveVQA

HumanMoveVQA evaluates whether MLLMs can reason about how a person moves and turns in 3D, relative to their first-frame position and heading. Unlike existing benchmarks, its answers depend on the 3D trajectory rather than on action labels or apparent motion in the image. Action labels are insufficient: two videos labelled "running" reveal neither the direction nor the magnitude of the motion. So is apparent image motion, since movement along the camera's viewing axis produces

Table 2: Source datasets. Camera: S = static, M = moving. Duration is the mean clip length.

Dataset	Camera	Views	People	Scenes	Dur. (s)	Train vids	Test vids
EMDB	M	1	1	in/out	37.0	57	11
RICH	S	multi	1	in/out	20.8	705	32
EgoBody	S	multi	2	in	38.7	300	54
3DPW	M	1	1–2+	in/out	28.3	40	15
SLOPER4D	M	1	1+	in/out	42.6	13	15

little displacement in the image. We therefore derive every answer deterministically from 3D world-space motion tracks, which makes answers independent of camera motion and viewpoint, generation scalable, and each answer traceable to geometric evidence. In the multi-view datasets RICH and EgoBody, the same motion seen from different viewpoints thus yields the same answers despite producing different 2D tracks.

This design distinguishes reasoning from perception, and the benchmark exposes the difference: the same motion viewed from different cameras yields zero-shot scores differing by up to 16 points (Section A.2). Recognising that a person walks is perception; determining where they end up relative to their initial heading requires inferences that no single frame provides: recovering world motion from monocular video under a moving camera, transforming it into the person's initial frame, and integrating events over 15 to 90 s. The seven question categories span this range and are grouped into four capabilities: event detection (Existence), which is primarily perceptual, and aggregation (Comparative, Dominant, Numerical), sequencing (Ordering, Temporal) and trajectory-level inference (Trajectory Affordance), which require these additional steps. HumanMoveVQA evaluates reasoning over observed motion; predicting unobserved motion, inferring intent and reasoning about interactions with the scene are outside the scope of this work.

To ensure that performance reflects true visual reasoning rather than linguistic bias, incorrect options are semantically plausible but geometrically inconsistent with the video, which prevents models from succeeding through language priors or artifacts of template-based generation. We quantify residual bias with a text-only baseline in Section 4. Section 3.1 describes the source datasets, Section 3.2 the pipeline that extracts world-consistent motion tracks, Section 3.3 the deterministic generation of question-answer (QA) pairs, and Section 3.4 the resulting benchmark statistics.

3.1 Data Source

We build HumanMoveVQA from five human-motion datasets with motion-capture ground truth: EMDB (Kaufmann et al., 2023), RICH (Huang et al., 2022), EgoBody (Zhang et al., 2022), 3DPW (von Marcard et al., 2018) and SLOPER4D (Dai et al., 2023), spanning static and moving cameras, indoor and outdoor scenes, one or more people, and clips of 15 to 90 s (Table 2). Labels are generated from video alone (Section 3.2); ground truth is used only to validate them (Section A.2), so the pipeline also applies to videos without motion capture.

Each source tests a different aspect of the task. In EMDB, 3DPW and SLOPER4D, a moving camera follows the subject, so camera motion must be separated from the subject's own; SLOPER4D also has the longest paths (median 44 m). RICH offers shorter static-camera clips with fewer events for short-range reasoning. EgoBody shows two interacting people from static exocentric views; we segment it into 30 to 60 s clips and ask about one person's trajectory at a time. In multi-person videos, an appearance description identifies the target person (Section 3.2).

In the multi-view datasets (RICH, EgoBody), each view is a separate sample, and corresponding questions share the same answers because they are defined relative to the person rather than the camera, directly testing viewpoint invariance. All views of a motion are assigned to the same split; the test set contains 7,534 motion events across 127 videos.

3.2 Data Generation Pipeline

Our pipeline converts monocular video into 3D, camera-independent motion events from which answers are computed (Figure 2), in four stages: (1) *world-space reconstruction*, recovering 3D human motion separately from camera motion; (2) *discretised spatial codes*, converting continuous root motion into discrete motion events; (3) *appearance-based identity grounding*, identifying the target person in multi-person scenes; and (4) *quality control*, removing unreliable reconstructions and events. QA pairs are then generated from the remaining events (Section 3.3).

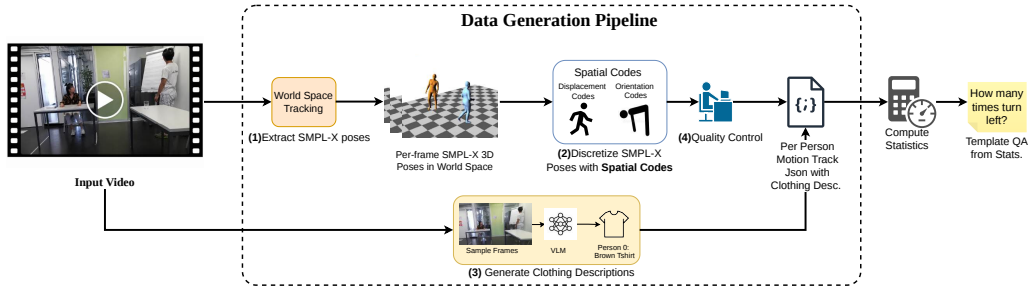


Figure 2: Data generation pipeline. **(1) World-space reconstruction:** PromptHMR (Wang et al., 2025b) recovers SMPL-X bodies and camera motion from monocular video, expressed relative to the person's first-frame position and heading (Equation (1)). **(2) Spatial codes:** root translation and rotation are segmented into binned motion events (Yazdian et al., 2025). **(3) Identity grounding:** Qwen3.5-9B (Qwen Team, 2026) describes each person's clothing from eight frames to identify the target. **(4) Quality control:** unreliable reconstructions and short or weak events are removed. Templates then deterministically generate QA pairs in seven categories.

World-Space Reconstruction. Under a moving camera, apparent motion mixes subject and camera motion, and motion along the viewing axis is hard to observe in 2D, so we recover motion in 3D. PromptHMR (Wang et al., 2025b) estimates per-frame SMPL-X (Pavlakos et al., 2019) bodies in a gravity-aligned world frame, jointly with camera motion; its per-person box prompts keep identities consistent in multi-person scenes. Let $\mathbf{p}_t \in \mathbb{R}^3$ and $\mathbf{R}_t \in SO(3)$ be the root position and orientation at frame t, and $\mathbf{R}_0^{\text{yaw}}$ the rotation about the vertical axis given by the person's heading at t_0. We express all motion in the person's initial frame:

$$\tilde{\mathbf{p}}_t = (\mathbf{p}_t - \mathbf{p}_0), \qquad \tilde{\mathbf{R}}_t = (\mathbf{R}_0^{\text{yaw}})^\top \mathbf{R}_t. \tag{1}$$

Here Y points up, $+Z$ is the person's initial facing direction and X their initial left-right axis; yaw is positive counterclockwise viewed from above. Because origin and axes are fixed by the person at t_0, answers do not depend on camera placement or motion.

Discretised Spatial Codes. Continuous root trajectories are noisy and cannot be mapped directly to language. Following MotionScript (Yazdian et al., 2025), we restrict its displacement and rotation codes to root-level dynamics, giving six *Spatial Codes* over $\tilde{\mathbf{p}}_t$ and $\tilde{\mathbf{R}}_t$: `displacement_x/y/z` and `rotation_roll/pitch/yaw`. Each code is segmented into *motion events*, contiguous spans in which the quantity changes in one direction beyond a no-action threshold. An event records its direction, start and end time, magnitude (change over the event) and mean speed, with magnitude and speed each discretised into five bins (Tables 24 and 25; units and frame rates in Section A.5). Discretisation absorbs small estimation errors and yields a structured vocabulary for question generation.

Appearance-Based Identity Grounding. In multi-person videos (EgoBody, 3DPW, SLOPER4D), Qwen3.5-9B (Qwen Team, 2026) describes each tracked person's clothing (*e.g.*, "blue jacket and black trousers") from eight frames with a valid bounding box, taking a majority vote over the parsed garment, colour and pattern attributes. The description is inserted into the question to identify the target person (prompt in Figure 14).

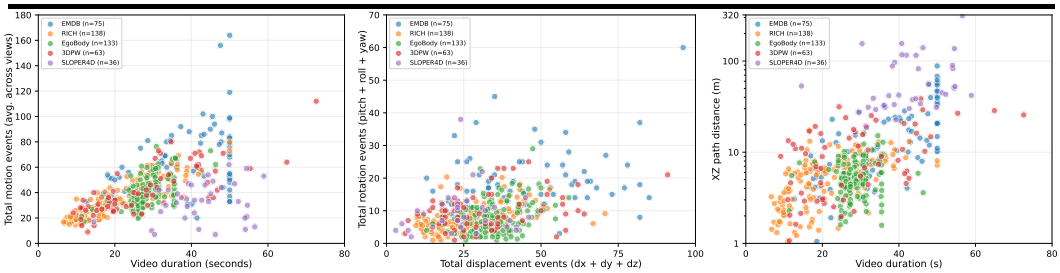
Quality Control. At the video level, we manually remove sequences whose reconstruction visibly diverges from the video, such as drifting ground planes or lost tracks. At the event level, we keep only events that exceed the no-action threshold and last at least 5 frames, discard the lowest magnitude bins (`very_short`, `slight`), and discard spans of eligible motion shorter than 1.0 s, so that questions target meaningful movements rather than reconstruction noise. Labels are validated against motion-capture ground truth in Section A.2.

3.3 Large-Scale Question-Answer Generation

To assess the reasoning capabilities of MLLMs, we generate QA pairs deterministically from structured representations, filling linguistic templates with statistics computed from the motion events and, in multi-person videos, the clothing descriptors (Section 3.2). Models receive only the video and the question; the structured events are used solely to compute answers. Following ActionArt, MVBench, MotionLLM and SAW-Bench (Peng et al., 2025; Li et al., 2023; Chen et al., 2025; Li et al., 2026a),

Table 3: Test set statistics for HumanMoveVQA.

Dataset	Videos	Total	Existence	Comparative	Dominant	Numerical	Ordering	Temporal	Traj. Afford.
EMDB	11	768	143	110	110	110	110	110	75
RICH	32	2,218	416	317	316	320	288	318	243
EgoBody	54	6,994	1,378	1,011	996	1,060	792	1,007	750
3DPW	15	1,294	247	180	181	190	170	190	136
SLOPER4D	15	1,008	195	143	145	150	134	148	93
Total	127	12,282	2,379	1,761	1,748	1,830	1,494	1,773	1,297



(a) Motion Events vs Video Length (b) Displacement vs Rotation (c) Path Distance vs Video Duration

Figure 3: Motion characteristics of the five sources before splitting (one point per video; multi-view datasets averaged over views). (a) Events vs. video length: EMDB has the most events, RICH and SLOPER4D the fewest. (b) Displacement vs. rotation events: EMDB varies most; EgoBody is mostly translational. (c) Path length (log scale) vs. video length: SLOPER4D travels furthest.

we define seven categories in the four capabilities of Section 3, several of which are asked both over the full video and within intervals given by timestamps.

Event detection. *Existence* asks whether a given movement or rotation occurs, as a binary question with positive and negative forms balanced.
Aggregation. *Comparative* asks which of two opposing directions on an axis has more events, greater total magnitude or a faster first event, with "roughly equal" as an option. *Dominant* asks for the direction with the larger summed magnitude, with ties decided by a stricter threshold. *Numerical* asks for the number of events of a given type or the total displacement mapped to world-space distance.
Sequencing. *Ordering* asks which event immediately follows an anchor event within an interval. *Temporal* asks for the first event within an interval, or the speed of a specific movement.
Trajectory-level inference. *Trajectory Affordance* aggregates displacement over the whole sequence: net displacement, total path length, the interval with the largest or smallest movement, and trajectory shape (*e.g.*, returning to the start versus moving away).

For each video, we generate up to ten questions per category, each traceable to the underlying 3D track. Incorrect options come from per-category templates (Table 23), drawn from the same answer space as the correct option (*e.g.*, other directions, intervals or event labels), so they are plausible under the question's wording but inconsistent with the observed motion. Option positions are shuffled and the correct-option positions are balanced within each category, to limit shortcuts from language priors or template regularities. Table 27 gives example questions and templates for every sub-type.

3.4 Benchmark Composition and Statistics

We select sequences with high-fidelity tracking and diverse motion, and split them into training and test sets. The test set comprises 12,282 QA pairs over 127 videos (Table 3), chosen for reliable reconstruction and a balanced mix of displacement and rotation; the remaining videos form the training split (Table 2). To prevent leakage, all views of a motion in RICH and EgoBody are assigned to the same split.

Figure 3 and Table 22 summarise the motion of the five sources before splitting, which differ in how much people move, turn and travel. EMDB has the most motion events (63.8 per video) and the widest spread in displacement and rotation (Figures 3a and 3b). 3DPW (42.3) and EgoBody (41.2) are intermediate; EgoBody is mostly translational, as its two people largely face each other. RICH (34.2) and SLOPER4D (32.8) have the fewest events, RICH because its clips are shortest and SLOPER4D because its long clips contain sustained movement rather than frequent turns. Path length differs (Figure 3c): RICH and EgoBody have the shortest paths (median 5.1 m), 3DPW and EMDB intermediate ones (9.0 and 13.3 m), and SLOPER4D the longest (44.0 m). The sources thus

span short, turn-heavy motion to long, sustained trajectories, so no single motion profile dominates. Overall, answering HumanMoveVQA questions requires tracking trajectories and orientation changes rather than superficial cues: answers are computed from 3D tracks, and the videos vary widely in motion and trajectory length. We evaluate reliance on language priors with a text-only baseline in Section 4.

4 EXPERIMENTS

4.1 RESULTS

We evaluate proprietary, open-source and motion-specialised Video MLLMs zero-shot, alongside models fine-tuned on our training split. Each model uses its recommended frame budget (Table 4); questions give intervals as timestamps, and a response that selects no valid option counts as incorrect. Implementation and training details are in Section A.1, and qualitative videos in the supplementary.

Metrics. Chance accuracy differs by category: 50% for Existence (2 options), 33.3% for Comparative and Dominant (3) and 25% for the rest (4). We therefore report per-category accuracy and an aggregate **Score**, the mean chance-normalised accuracy over the $C = 7$ categories:
Score $= \frac{1}{C}\sum_{i=1}^{C} \frac{\text{acc}_i - \text{chance}_i}{100 - \text{chance}_i} \times 100$, where 0 is chance, 100 perfect accuracy and negative values below chance. Scores are computed per dataset and averaged over the five. Table 4 reports the averages; per-dataset results and qualitative examples (Figure 4) are in Section A.1.1.

Table 4: Evaluation of MLLM on HumanMoveVQA. Each cell is the mean of the corresponding per-split value. Results for seven reasoning categories and an overall normalized score. Green, orange, and blue indicate the best, second-best, and third-best results per column, respectively.

Model	Frames	Existence	Comparative	Dominant	Numerical	Ordering	Temporal	Traj. Afford.	Score
Random Chance	–	50.00	33.33	33.33	25.00	25.00	25.00	25.00	0.00
Human (subset)†	–	100.00	80.56	84.68	87.34	75.77	90.00	74.38	78.74
Proprietary Models									
Gemini-3.8-flash	2 FPS	63.00	36.97	35.80	40.35	26.85	31.49	28.60	10.22
GPT-5.6-Luna	32	64.54	39.29	46.59	29.38	21.43	30.12	26.51	9.69
Gemini-3.8-flash text	–	59.03	34.86	39.24	26.31	31.47	25.12	22.82	5.26
Open-Source MLLM									
Qwen3-VL 8B	32	66.10	37.30	46.49	34.77	31.21	31.52	32.63	14.01
LLaVA-OV-2 8B	32	65.39	34.24	46.93	42.86	27.69	31.28	28.64	13.31
Qwen3-VL 4B	32	61.07	39.44	45.80	34.69	33.21	28.01	30.90	12.25
VideoChat3 4B	32	62.34	38.47	48.73	29.68	31.09	26.66	28.08	10.88
InternVL3_5-8B	8	62.91	37.59	46.66	31.93	27.48	29.61	27.13	10.53
Molmo2 4B	32	60.83	38.34	48.91	28.69	30.71	24.29	20.96	8.39
LLaVA-NeXT-Video-7B	64	48.68	35.87	45.28	28.01	26.87	23.83	26.52	3.72
VideoGPT+ 4B	16	50.72	36.49	46.91	15.94	26.77	22.14	23.70	1.61
Video-LLaVA 7B	8	48.29	35.39	48.28	20.31	24.51	25.51	16.02	0.55
Motion-Specialized Model									
MotionLLM 7B	8	44.65	27.01	37.18	21.25	21.41	18.56	11.16	-7.32
Qwen3-VL 8B SFT	32	82.46	46.71	59.08	78.06	34.54	46.41	49.43	38.32
Qwen3-VL 4B SFT (Ours)	32	81.69	46.31	59.86	79.82	35.80	49.15	47.77	38.96
InternVL3.5-4B SFT	8	77.80	44.17	53.71	79.69	34.99	43.00	43.62	33.93

Proprietary models. Gemini-3.8-flash is the strongest proprietary model (Score 10.22), best on Existence (63.00) and Numerical (40.35) but near chance on Ordering (26.85). GPT-5.6-Luna is near chance on Trajectory Affordance (26.51) and below chance on Ordering (21.43), indicating difficulty tracking motion over long horizons. Without video, Gemini-3.8-flash reaches about half its Score (5.26 vs. 10.22), with its residual advantage concentrated in Dominant (39.24) (wording only).

Open-source models. Qwen3-VL 8B is the strongest zero-shot model (14.01), followed by LLaVA-OV-2 8B (13.31) and Qwen3-VL 4B (12.25); Video-LLaVA, VideoGPT+ and LLaVA-NeXT-Video are barely above chance. MotionLLM, trained with joint-level motion supervision, is below chance in every category (-7.32), showing that such supervision does not work for global trajectory reasoning.

Fine-tuned models. Fine-tuning roughly triples the Score of Qwen3-VL (4B: 12.25 to 38.96; 8B: 14.01 to 38.32). For the 4B model, gains are largest in Numerical ($+45.1$ pp), Temporal ($+21.1$ pp) and Trajectory Affordance ($+16.9$ pp), while Ordering improves by only $+2.6$ pp.

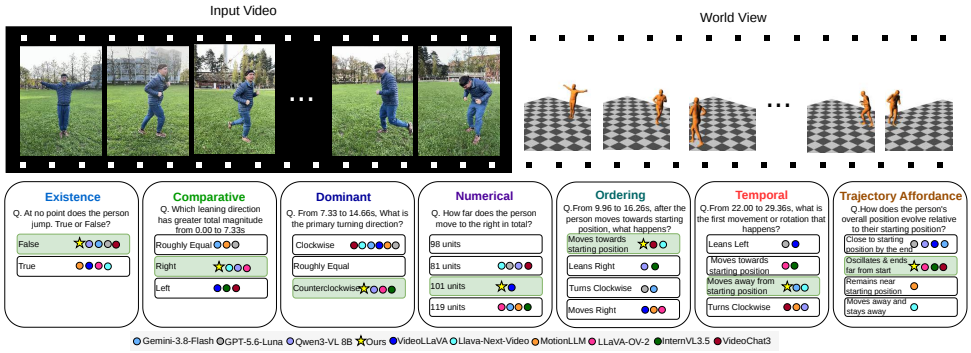
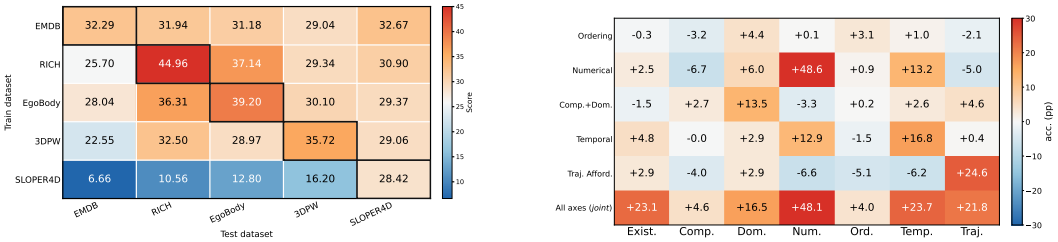


Figure 4: Qualitative results on **EMDB**. The world view shows the extracted 3D SMPL-X poses for illustration only; models do not see it. Predictions from several MLLMs are shown for all seven categories, with the correct option in green. More examples in Figure 6.



(a) Cross-dataset generalisation: training on one dataset, testing on another.

(b) Per-category vs. joint SFT: change in accuracy (pp) per category, averaged over datasets.

Figure 5: Effect of training data (a) and training categories (b) on SFT performance.

Takeaways.

- **A large gap to humans.** The best zero-shot Score is 14.01 against 78.74 for humans, with Ordering and Trajectory Affordance near chance for every zero-shot model.
- **Models reason in the image, not in the world.** Zero-shot Scores vary by up to 16 points across views of the same motion in RICH, although the answers are identical, so the spread reflects a failure to infer world motion from the image; fine-tuning reduces it to about 7 points (Section A.2).
- **The gap is learnable but not closed.** World-consistent supervision roughly triples the Score of Qwen3-VL, but the best fine-tuned model (38.96) remains well below humans, and joint-level supervision (MotionLLM) does not help.
- **Temporal ordering is the hardest category.** Ordering gains only $+2.6$ pp after fine-tuning and stays about 40 points below humans. It requires detecting and sequencing several short events, and is the category most affected by reconstruction noise (Section A.2) and sparse frame sampling (Section A.4).

4.2 Ablations

We study two SFT design choices: whether training on one dataset transfers to others, and whether all seven categories must be trained to improve all of them. Ablations on frame count, resolution, reasoning supervision, paraphrasing, out-of-domain evaluation and video length are in Section A.4.

Cross-dataset generalisation. We train on each dataset and evaluate on all five (Table 5, Figure 5a). In-domain training is best on four of the five test sets, reflecting differences in motion distribution; on SLOPER4D, training on EMDB is better (32.67 vs. 28.42). Across the other test sets, EMDB (31.2), EgoBody (31.0) and RICH (30.8) transfer similarly, although EMDB has ten times fewer training questions than RICH (3,949 vs. 41,429), which we attribute to its motion diversity (63.8 events per video). SLOPER4D transfers least (11.6), mainly due to its small training set (499 questions from 13 videos). Numerical transfers most stably (only -2.87 pp), while Ordering remains hardest in- and cross-domain. As no single dataset is best everywhere, we train jointly on all five.

Per-category vs. joint training. We compare joint training with five category-specific models (Numerical, Ordering, Temporal, Trajectory Affordance, and Comparative with Dominant; Table 26, Figure 5b). Each specialist improves its own category, most strongly Numerical ($+45.6$ pp on

Table 5: **Cross-Dataset Evaluation**. In-domain and cross-domain performance of Qwen3-VL 4B when trained on different dataset subsets. Results highlight generalization across datasets; green indicates the best result on the test set, and orange indicates the second best.

Train → Test	Existence	Comparative	Dominant	Numerical	Ordering	Temporal	Traj. Afford.	Score
EMDB → EMDB	85.31	41.82	50.91	64.55	31.82	48.18	42.67	32.29
RICH → EMDB	75.52	36.36	49.09	71.82	33.64	32.73	37.33	25.70
EgoBody → EMDB	72.73	42.73	47.27	73.64	32.73	41.82	38.67	28.04
3DPW → EMDB	76.92	37.27	44.55	58.18	27.27	38.18	37.33	22.55
SLOPER4D → EMDB	37.76	33.64	51.82	51.82	19.09	37.27	24.00	6.66
EMDB → RICH	85.82	35.33	50.95	71.25	38.54	48.74	33.33	31.94
RICH → RICH	89.90	49.84	60.44	79.69	45.83	52.20	49.38	44.96
EgoBody → RICH	83.65	44.48	53.48	81.88	38.89	44.65	39.51	36.31
3DPW → RICH	85.10	35.33	46.84	72.81	36.81	46.86	44.03	32.50
SLOPER4D → RICH	48.56	34.70	46.52	57.81	29.17	28.30	25.93	10.56
EMDB → EgoBody	75.33	40.65	57.83	66.51	34.22	44.79	44.40	31.18
RICH → EgoBody	79.83	45.30	56.43	80.66	34.47	47.37	48.27	37.14
EgoBody → EgoBody	83.89	50.84	57.83	79.43	34.09	45.38	48.80	39.20
3DPW → EgoBody	76.12	40.06	49.80	66.23	32.83	42.80	44.93	28.97
SLOPER4D → EgoBody	50.94	35.61	53.51	56.51	24.62	29.39	30.00	12.80
EMDB → 3DPW	81.78	35.00	53.59	61.05	35.88	40.53	42.65	29.04
RICH → 3DPW	84.21	39.44	50.28	73.16	27.06	36.84	39.71	29.34
EgoBody → 3DPW	81.78	41.11	50.28	74.74	27.65	38.95	41.18	30.10
3DPW → 3DPW	80.97	42.22	59.67	67.89	38.82	48.42	46.32	35.72
SLOPER4D → 3DPW	52.63	35.00	55.25	57.37	30.00	32.63	34.56	16.20
EMDB → SLOPER4D	81.03	41.26	53.79	67.33	35.07	35.81	54.84	32.67
RICH → SLOPER4D	75.90	47.55	53.79	80.67	31.34	35.81	36.56	30.90
EgoBody → SLOPER4D	75.38	43.36	44.83	81.33	29.85	39.86	40.86	29.37
3DPW → SLOPER4D	72.82	39.16	49.66	70.67	26.12	43.92	52.69	29.06
SLOPER4D → SLOPER4D	66.67	39.16	57.93	66.67	32.09	38.51	52.69	28.42

average) and Trajectory Affordance ($+19.0$ pp), often at the expense of others: Numerical-only training lowers Comparative by 8.2 pp on EMDB and 8.8 pp on RICH. Dominant benefits most consistently as a side effect, though it drops by 9.4 pp under Numerical-only training on 3DPW. Joint training gives the best Score on every dataset, exceeding the strongest specialist by at least 14 points and zero-shot by 17.9 to 32.0 points, consistent with positive transfer between categories, although it also sees more training questions.

5 Conclusion

We introduced HumanMoveVQA, a benchmark for evaluating whether Video MLLMs can reason about a person's global root trajectory and orientation in 3D, relative to their first-frame position and heading. A scalable pipeline recovers world-space motion from monocular video and deterministically generates 12,282 test questions in seven categories, with labels validated against motion-capture ground truth. Current models fall far below humans (Score 14.01 vs. 78.74), with Ordering and Trajectory Affordance near chance, and their scores vary by up to 16 points across camera views of the same motion. Joint-level motion supervision does not help, whereas world-consistent supervision roughly triples the Score of Qwen3-VL and reduces this viewpoint sensitivity; joint training across categories and answer-only supervision work best. Even so, the best fine-tuned model reaches only 38.96, and temporal ordering of motion events remains particularly challenging.

Limitations and future work. Labels come from monocular reconstruction, so some noise remains; about 10% of answers change under ground-truth tracks (Section A.2). HumanMoveVQA covers one person's global trajectory and orientation, not interactions or intent, and our fine-tuned model drops on ActionArt's joint-level categories (Section A.4), indicating that the two need complementary supervision. It also evaluates observed rather than unobserved motion, such as forecasting or occlusion, which future work can address alongside multi-person interaction and denser temporal encoding for event ordering.

AI USE STATEMENT

We used generative AI tools to assist with implementation, experimental analysis, and interpretation of results. We additionally used generative AI tools for code development, literature search, compiling results, and drafting and editing parts of the manuscript. All AI-assisted work was reviewed and verified by the authors. We take responsibility for the final content of this work.

ETHICS STATEMENT

All authors have read and comply with the ICLR Code of Ethics. This work does not collect new human-motion data and builds on existing datasets under their respective terms of use. Human evaluators participated with informed consent. Our benchmark evaluates geometric properties of human motion and does not infer sensitive attributes, identity, or intent.

REPRODUCIBILITY STATEMENT

We release the complete benchmark-generation and evaluation pipeline, experimental code, and LoRA fine-tuning configurations at `https://github.com/humanmovevqa/HumanMove_VQA`, and the benchmark data at `https://huggingface.co/datasets/humanmovevqa/HumanMoveVQA`. The data curation pipeline is described in Section 3.2 and further detailed in the Appendix, including label generation, evaluation, and reliability analysis. These resources enable reproduction of the benchmark construction, fine-tuning, and experimental results.

REFERENCES

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Kaichen Zhang, Wenkang Zhang, Zheng Cheng, Nansen Zhang, Chunsheng Wu, Chunjiang Ge, Zimin Ran, Dehua Song, Chunyuan Li, Shikun Feng, Ming Hu, Zhangquan Chen, Junbo Niu, Bo Li, Ziyong Feng, Ziwei Liu, Zongyuan Ge, and Jiankang Deng. Llava-onevision-2: Towards next-generation perceptual intelligence. In *arXiv*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1728–1738, 2021.

Hunar Batra, Haoqin Tu, Hardy Chen, Yuanze Lin, Cihang Xie, and Ronald Clark. Spatialthinker: Reinforcing 3d reasoning in multimodal llms via spatial rewards, 2025.

Andy Bonnetto, Haozhe Qi, Franklin Leong, Matea Tashkovska, Mahdi Rad, Solaiman Shokur, Friedhelm Hummel, Silvestro Micera, Marc Pollefeys, and Alexander Mathis. Epfl-smart-kitchen-30: Densely annotated cooking dataset with 3d kinematics to challenge video and language models, 2025.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, June 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.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. In *NeurIPS*, 2024.

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.

Yudi Dai, Yitai Lin, Xiping Lin, Chenglu Wen, Lan Xu, Hongwei Yi, Siqi Shen, Yuexin Ma, and Cheng Wang. Sloper4d: A scene-aware dataset for global 4d human pose estimation in urban environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 682–692, June 2023.

Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. Posescript: 3d human poses from natural language. In *ECCV*, 2022.

Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J. Black. Chatpose: Chatting about 3d human pose. In *CVPR*, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *CVPR*, 2025.

Google DeepMind. Gemini 3.8 flash model card. `https://deepmind.google/models/model-cards/gemini-3-8-flash/`, September 2026. Published: 2026-09-02.

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, 2025.

Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. In *International Conference on Computer Vision (ICCV)*, 2023.

Chuhan Li, Ruilin Han, Joy Hsu, Yongyuan Liang, Rajiv Dhawan, Jiajun Wu, Ming-Hsuan Yang, and Xin Eric Wang. Learning situated awareness in the real world, 2026a.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Renjie Li, Ruijie Ye, Mingyang Wu, Hao Frank Yang, Zhiwen Fan, Hezhen Hu, and Zhengzhong Tu. Mmhu: A massive-scale multimodal benchmark for human behavior understanding, 2025.

Xinhao Li, Yuhan Zhu, Xiangyu Zeng, Yuhao Dong, Haoning Wu, Zhiqiu Zhang, Yuandong Yang, Changlian Ma, Qingyu Zhang, Yansong Shi, et al. Videochat3: Fully open video mllm for efficient and generalist video understanding. *arXiv preprint arXiv:2607.14935*, 2026b.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36:25268–25280, 2023.

Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso M de Melo, Jianwen Xie, and Alan Yuille. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=IDJx97BC38.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman H. Khan, and Fahad Shahbaz Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *ArXiv*, abs/2406.09418, 2024.

OpenAI. GPT-5.6 system card. https://deploymentsafety.openai.com/gpt-5-6, July 2026. Published July 9, 2026.

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 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 10975–10985, 2019.

Yi-Xing Peng, Qize Yang, Yu-Ming Tang, Shenghao Fu, Kun-Yu Lin, Xihan Wei, and Wei-Shi Zheng. Actionart: Advancing multimodal large models for fine-grained human-centric video understanding. *arXiv preprint arXiv:2504.18152*, 2025.

Haozhe Qi, Shaokai Ye, Alexander Mathis, and Mackenzie W. Mathis. Llavaction: evaluating and training multi-modal large language models for action understanding. *ICLR*, 2026.

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

Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Lvbench: An extreme long video understanding benchmark, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Haoran Hao, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Ying Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kai Zhang, Hui Deng, Biqing Qi, Biqing Qi, Qipeng Guo, Wenwei Zhang, Yuzhe Gu, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Bowen Zhou, Weijie Su, Kaiming Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *ArXiv*, abs/2508.18265, 2025a.

Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael John Black, and Muhammed Kocabas. Prompthmr: Promptable human mesh recovery. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1148–1159, 2025b.

Bo Wu and Shoubin Yu. Star: A benchmark for situated reasoning in real-world videos. *ArXiv*, abs/2405.09711, 2024.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-MLLM: Boosting MLLM capabilities in visual-based spatial intelligence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024. URL https://arxiv.org/abs/2407.15754.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Liang Xu, Shaoyang Hua, Zili Lin, Yifan Liu, Feipeng Ma, Yichao Yan, Xin Jin, Xiaokang Yang, and Wenjun Zeng. Motionbank: A large-scale video motion benchmark with disentangled rule-based annotations. *arXiv preprint arXiv:2410.13790*, 2024.

Chiao-An Yang, Ryo Hachiuma, Sifei Liu, Subhashree Radhakrishnan, Raymond A Yeh, Yu-Chiang Frank Wang, and Min-Hung Chen. 4d-rgpt: Toward region-level 4d understanding via perceptual distillation. *arXiv preprint arXiv:2512.17012*, 2025.

Jihan Yang, Shusheng Yang, Anjali Gupta, Rilyn Han, Fei-Fei Li, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643, 2024.

Payam Jome Yazdian, Rachel Lagasse, Hamid Mohammadi, Eric Liu, Li Cheng, and Angelica Lim. Motionscript: Natural language descriptions for expressive 3d human motions. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025.

Xingyilang Yin, Chengzhengxu Li, Jiahao Chang, Chi-Man Pun, and Xiaodong Cun. Mllm-4d: Towards visual-based spatial-temporal intelligence. *arXiv preprint arXiv:2603.00515*, 2026.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: a dataset for understanding complex web videos via question answering. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.33019127. URL https://doi.org/10.1609/aaai.v33i01.33019127.

Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European Conference on Computer Vision*, 2022.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024a.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *Trans. Mach. Learn. Res.*, 2025, 2024b.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. Swift:a scalable lightweight infrastructure for fine-tuning, 2024.

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3D LLM: Learning position-aware video representation for 3D scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Wang HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment, 2023.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. LLaVA-3D: A simple yet effective pathway to empowering LMMs with 3D capabilities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

APPENDIX INDEX

A.1 EXPERIMENTS

- Evaluation Setup
- Implementation Details
- Results
- Takeaways

A.2 LABEL RELIABILITY

- Reconstruction Error Measured Against Ground-Truth Tracks
- Camera Viewpoint Sensitivity

A.3 CONFIDENCE INTERVALS AND ROBUSTNESS

- Pooled and Split-Averaged Scores

A.4 ABLATION

- Effect of Frame Count
- Effect of Resolution
- Reasoning Supervision
- Template Robustness
- Out-of-Domain Evaluation
- Effect of Video Length

A.5 UNITS AND DISCRETIZATION

- World-space quantities
- Discretization
- Velocity
- Frame rate

A.6 BROADER IMPACT

A APPENDIX

A.1 EXPERIMENTS

Evaluation Setup – We evaluate a diverse suite of models, including proprietary, open-source, and motion-specialized Video MLLMs. For proprietary models, we test GPT-5.6-Luna and Gemini-3.8-flash, the latter of which is evaluated in both a standard video-text setting and a text-only setting to quantify the reliance on visual information. Open-source baselines include Qwen3-VL (Bai et al., 2025), LLaVA-OV-2 (An et al., 2026), InternVL3.5 Wang et al. (2025a), VideoChat3 (Li et al., 2026b), Molmo2 (Clark et al., 2026), LLaVA-NeXT-Video Zhang et al. (2024b), Video-LLaVA Lin et al. (2024) and VideoGPT+ (Maaz et al., 2024). We also evaluate MotionLLM (Chen et al., 2025), which incorporates an explicit motion encoder trained on joint-level captions. Alongside these zero-shot baselines we report the fine-tuned variants described below, namely our Qwen3-VL 4B SFT, Qwen3-VL 8B SFT and InternVL3.5-4B SFT. All zero-shot baselines are evaluated using recommended configurations. Additionally, we conduct human evaluation on a small subset of the dataset. Participants were given one question of each type per video for evaluation. We report per category Accuracy percentage and Normalized Score.

Table 6: Evaluation of video MLLM models on the **EMDB** split. Results are reported across seven reasoning categories and an overall normalized score. Green, orange, and blue indicate the best, second-best, and third-best results per column, respectively.

Model	Frames	Existence	Comparative	Dominant	Numerical	Ordering	Temporal	Traj. Afford.	Score
Random Chance	–	50.00	33.33	33.33	25.00	25.00	25.00	25.00	0.00
Human (subset)	–	100.00	78.67	83.33	85.50	75.50	95.00	75.00	78.71
Proprietary Models									
GPT-5.6-Luna	32	71.33	43.64	45.45	28.18	25.45	25.45	12.00	9.20
Gemini-3.8-flash	2 FPS	71.33	47.27	30.00	32.73	20.00	27.27	24.00	9.13
Gemini-3.8-flash text	–	62.24	36.36	30.00	26.36	23.64	27.27	16.00	2.15
Open-Source MLLM									
VideoChat3 4B	32	69.23	36.36	44.55	33.64	30.00	25.45	22.67	10.79
Qwen3-VL 8B	32	62.24	39.09	44.55	33.64	25.45	29.09	30.67	10.73
Qwen3-VL 4B	32	66.43	44.55	38.18	31.82	27.27	24.55	26.67	10.10
InternVL3_5-8B	8	69.23	46.36	38.18	26.36	24.55	27.27	25.33	9.99
Molmo2 4B	32	67.13	31.82	45.45	31.82	28.18	28.18	13.33	7.46
LLaVA-OV-2 8B	32	57.34	32.73	43.64	30.00	23.64	32.73	24.00	6.15
LLaVA-NeXT-Video-7B	64	46.15	39.09	44.55	29.09	28.18	23.64	28.00	4.24
Video-LLaVA 7B	8	59.44	37.27	44.55	26.36	22.73	22.73	13.33	3.12
VideoGPT+ 4B	16	58.04	29.09	48.18	17.27	23.64	20.91	29.33	2.88
Motion-Specialized Model									
MotionLLM 7B	8	44.76	30.00	31.82	22.73	20.91	20.00	10.67	-7.43
Qwen3-VL 8B SFT	32	84.62	39.09	60.91	73.64	27.27	46.36	46.67	34.93
Qwen3-VL 4B SFT (Ours)	32	83.92	40.91	60.00	77.27	35.45	51.82	44.00	37.70
InternVL3.5-4B SFT	8	74.13	44.55	53.64	79.09	30.91	39.09	42.67	31.13

Implementation Details – We fine-tune three models on a shared set of 86,671 answer-only QA pairs generated from the PromptHMR tracks of 1,111 training videos of all five splits : RICH 41,429 from 705 videos, EgoBody 37,024 from 300 videos, EMDB 3,949 from 57 videos, 3DPW 3,770 from 40 videos and SLOPER4D 499 from 13 videos, with 11,065 questions held out for validation . Our two Qwen3-VL models (4B and 8B) are trained with LoRA (rank 16, alpha 32, dropout 0.05) via LlamaFactory Zheng et al. (2024) for 2.75 epochs at a learning rate of 1×10^{-4} with cosine decay and 0.1 warmup, using an effective batch size of 8 and 32 uniformly sampled frames per video at 256×256 resolution. The InternVL3.5-4B model is trained with the same LoRA settings via ms-swift Zhao et al. (2024) for 3 epochs at an effective batch size of 32, with the vision encoder and aligner frozen and a smaller input policy of 8 frames. Training was carried out on a single NVIDIA H200 GPU and took two days. The system prompt is shared between training and evaluation and is shown in Figure 13. We additionally provide qualitative results for each source dataset as videos with 3D motion visualizations in the supplementary material.

A.1.1 Results

EMDB – The EMDB dataset consists of monocular videos of a single person captured with a dynamic handheld camera in indoor and outdoor scenes. As shown in Table 6, VideoChat3 4B leads zero-shot models with a Score of 10.79, followed closely by Qwen3-VL 8B at 10.73. Among proprietary models, GPT-5.6-Luna and Gemini-3.8-flash are nearly tied at 9.20 and 9.13, with Gemini-3.8-flash the strongest model on Comparative (47.27). The text-only baseline performs near the Random Chance baseline, confirming the absence of exploitable linguistic priors in the benchmark.

MotionLLM remains the weakest model at -7.43, scoring below chance on all seven categories. Our Qwen3-VL 4B SFT model achieves the highest aggregate Score of 37.70, a gain of 27.60 pp over its base model, with the largest improvements in Numerical ($+45.45$ pp), Temporal ($+27.27$ pp) and Dominant ($+21.82$ pp). The larger Qwen3-VL 8B SFT reaches 34.93 and InternVL3.5-4B SFT 31.13, the latter leading every model on Numerical (79.09). Qualitative results are illustrated in Figure 6.

RICH – The RICH dataset consists of short clips (10–20s) captured by static cameras, with fewer motion events compared to EMDB. As shown in Table 7, among proprietary models Gemini-3.8-flash reaches 11.75, ahead of GPT-5.6-Luna at 9.99, while the text-only setting falls to 2.24, confirming

that the benchmark requires visual information rather than textual priors. Open-source models edge ahead of them, with Qwen3-VL 8B at 13.22, LLaVA-OV-2 8B at 11.95 and Qwen3-VL 4B at 11.86.

Table 7: Evaluation of video MLLM models on the **RICH** split. Results are reported across seven reasoning categories and an overall normalized score. Green, orange, and blue indicate the best, second-best, and third-best results per column, respectively.

Model	Frames	Existence	Comparative	Dominant	Numerical	Ordering	Temporal	Traj. Afford.	Score
Random Chance	–	50.00	33.33	33.33	25.00	25.00	25.00	25.00	0.00
Human (subset)	–	100.00	83.40	88.56	80.50	75.50	90.00	75.50	79.04
Proprietary Models									
Gemini-3.8-flash	2 FPS	67.31	35.02	38.29	40.94	26.39	33.33	27.57	11.75
GPT-5.6-Luna	32	65.62	36.59	48.42	29.38	21.18	31.45	26.34	9.99
Gemini-3.8-flash text	–	58.65	31.55	37.34	24.69	33.33	21.38	16.87	2.24
Open-Source MLLM									
Qwen3-VL 8B	32	68.27	35.02	44.30	35.94	30.90	38.68	22.22	13.22
LLaVA-OV-2 8B	32	66.35	30.28	46.20	49.69	22.92	29.87	24.69	11.95
Qwen3-VL 4B	32	63.22	41.96	45.57	30.31	38.54	28.30	21.81	11.86
InternVL3_5-8B	8	63.70	33.75	50.32	29.06	27.78	27.04	25.93	9.51
VideoChat3 4B	32	61.54	32.81	44.94	25.00	29.17	25.79	20.99	5.85
Molmo2 4B	32	56.97	35.96	48.10	28.12	30.56	21.38	17.28	5.21
LLaVA-NeXT-Video-7B	64	48.32	31.86	44.94	26.56	30.21	28.30	27.57	4.10
VideoGPT+ 4B	16	48.56	34.38	49.37	15.00	25.35	22.64	28.40	1.61
Video-LLaVA 7B	8	48.80	31.23	43.99	22.50	29.86	25.16	22.63	1.52
Motion-Specialized Model									
MotionLLM 7B	8	43.99	25.55	42.41	22.81	18.75	19.50	8.23	-7.29
Qwen3-VL 8B SFT	32	89.42	55.52	62.34	82.50	42.01	45.60	51.85	45.46
Qwen3-VL 4B SFT (Ours)	32	87.74	52.37	63.61	80.62	38.89	52.52	46.09	43.85
InternVL3.5-4B SFT	8	85.82	39.43	54.43	81.56	40.28	48.74	40.74	37.27

MotionLLM remains the worst performer at -7.29, with its Trajectory Affordance accuracy (8.23) the lowest cell in the table, far below the 25.00 chance baseline, mirroring its consistent failure in trajectory-based reasoning. Fine-tuning reshapes the top of the table: our Qwen3-VL 4B SFT lifts its base model from 11.86 to 43.85, led by Numerical ($+50.31$ pp) with Existence, Trajectory Affordance and Temporal clustered near $+24$ pp, while Ordering barely moves ($+0.35$ pp), being the strongest zero-shot model in that column (38.54). The larger Qwen3-VL 8B SFT reaches 45.46 and InternVL3.5-4B SFT 37.27. Qualitative results are illustrated in Figure 7.

EgoBody – The EgoBody consists of static multi-view recordings of two interacting individuals in sequences ranging from 30-60 seconds. On this dataset, we only evaluate individual motion rather than social interactions to maintain consistency with our single-person trajectory focus; however, the presence of a second subject introduces significant complexity for localization and tracking.

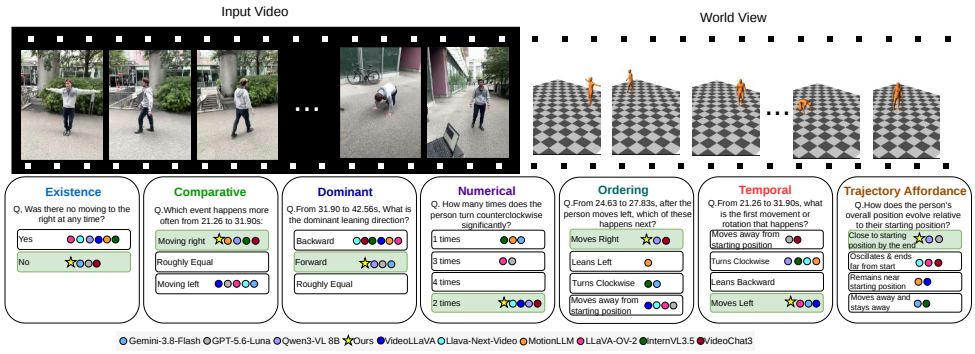


Figure 6: Qualitative results on the **EMDB** dataset. The world-view visualization depicts extracted 3D SMPL-X poses and is shown for illustration only (not provided to the models). Green denotes the correct option. We compare predictions from multiple MLLMs across seven reasoning categories, where our model demonstrates more accurate and consistent responses.

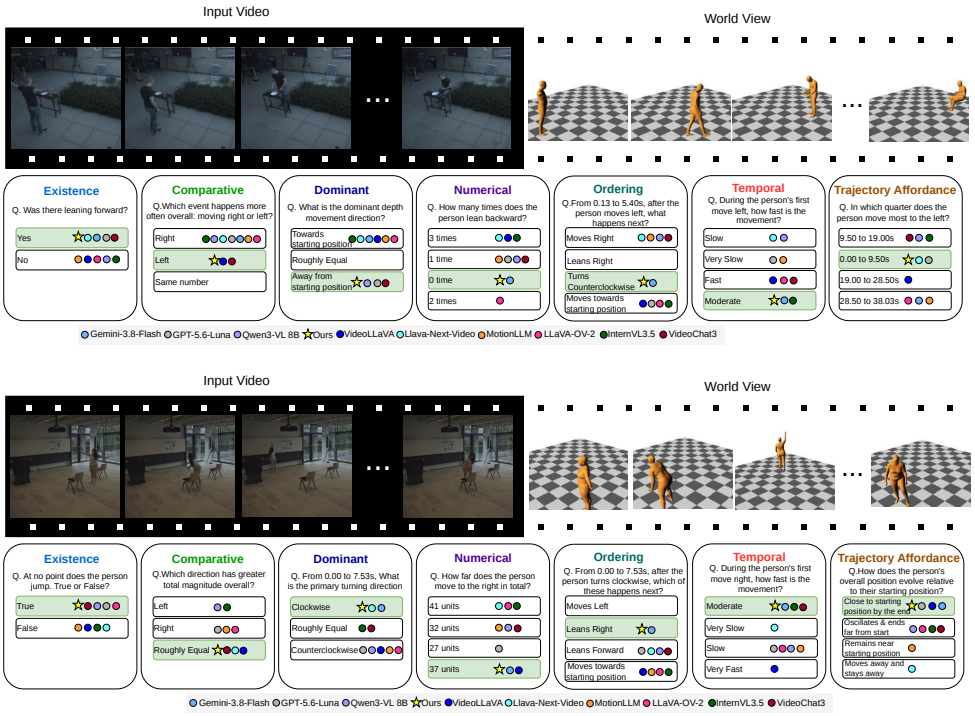


Figure 7: Qualitative results on the **RICH** dataset. The world-view visualization depicts extracted 3D SMPL-X poses and is shown for illustration only (not provided to the models). Green denotes the correct option. We compare predictions from multiple MLLMs across seven reasoning categories, where our model demonstrates more accurate and consistent responses.

As shown in Table 8, among proprietary models Gemini-3.8-flash reaches 11.79, ahead of GPT-5.6-Luna at 10.78, while the text-only setting drops to 7.50. That text-only model is nevertheless the strongest on Ordering (35.61), above every video-based model including our own SFT, which suggests that the ordering questions on this split are partly answerable from the question text alone. Open-source models lead the zero-shot field, with Qwen3-VL 8B at 16.25 and LLaVA-OV-2 8B at 14.62.

MotionLLM remains the worst performer at -4.94, with Numerical and Ordering falling below chance. Fine-tuning transforms the split: our Qwen3-VL 4B SFT lifts its base model from 11.24 to 40.75, led by Numerical ($+48.68$ pp), with Existence and Trajectory Affordance also gaining more than 20 pp, while Ordering moves least ($+3.41$ pp), consistent with the text-only shortcut noted above. The three fine-tuned models are separated by only 1.63 points (40.75, 40.73 and 39.12), despite InternVL3.5-4B SFT seeing 8 frames against 32 for the other two. Figure 8 demonstrates qualitative results.

3DPW – The 3DPW split consists of monocular videos captured with a moving camera, each featuring one or two target people; some sequences contain many other people as distractors, so the model must first resolve which person a question refers to. As shown in Table 9, among proprietary models GPT-5.6-Luna reaches 9.97, ahead of Gemini-3.8-flash at 9.26, while the text-only setting reaches 7.38. Open-source models lead by a wider margin here: LLaVA-OV-2 8B reaches 18.10, more than three points clear of the next zero-shot model, and it is also the strongest zero-shot model on Existence (76.11) and Numerical (37.37).

MotionLLM remains the worst performer at -8.73, below chance on six of the seven categories. Fine-tuning again drives the top of the table: our Qwen3-VL 4B SFT lifts its base model from 12.75 to 39.24, led by Numerical ($+47.89$ pp) and Temporal ($+23.15$ pp), while Ordering barely moves ($+1.77$ pp). The Qwen3-VL 8B SFT is marginally ahead at 39.37, a gap of only 0.13 points, and InternVL3.5-4B SFT reaches 33.39 with just 8 frames. Qualitative results are illustrated in Figure 9.

SLOPER4D – The SLOPER4D dataset consists of subjects playing football, running, and moving through library and garden scenes, filmed with a dynamic camera over long distances. As shown in

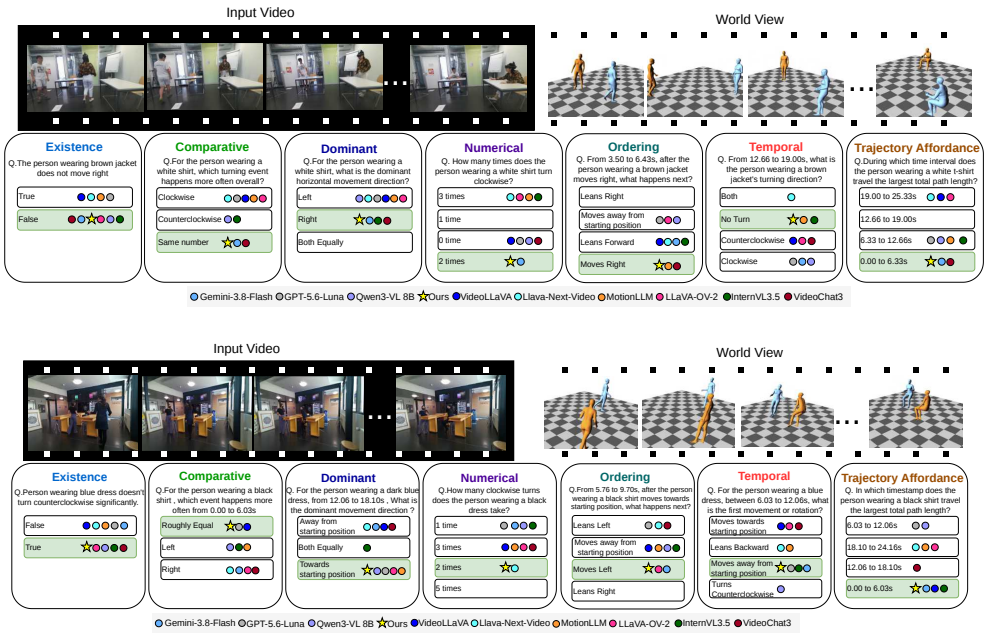


Figure 8: Qualitative results on the **EgoBody** dataset. The world-view visualization depicts extracted 3D SMPL-X poses and is shown for illustration only (not provided to the models). Green denotes the correct option. We compare predictions from multiple MLLMs across seven reasoning categories, where our model demonstrates more accurate and consistent responses.

Table 8: Evaluation of video MLLM models on the **EgoBody** split. Results are reported across seven reasoning categories and an overall normalized score. Green, orange, and blue indicate the best, second-best, and third-best results per column, respectively.

Model	Frames	Existence	Comparative	Dominant	Numerical	Ordering	Temporal	Traj. Afford.	Score
Random Chance	–	50.00	33.33	33.33	25.00	25.00	25.00	25.00	0.00
Human (subset)	–	100.00	76.33	83.67	85.00	85.00	90.00	75.00	79.05
Proprietary Models									
Gemini-3.8-flash	2 FPS	60.30	32.05	39.06	44.72	31.31	34.86	30.53	11.79
GPT-5.6-Luna	32	59.87	38.87	46.59	28.77	25.38	31.68	34.80	10.78
Gemini-3.8-flash text	–	57.55	33.23	43.98	26.60	35.61	25.32	28.67	7.50
Open-Source MLLM									
Qwen3-VL 8B	32	69.01	44.41	52.21	31.89	31.94	30.88	28.40	16.25
LLaVA-OV-2 8B	32	63.57	36.99	50.10	42.55	32.45	33.07	25.33	14.62
InternVL3_5-8B	8	61.76	40.45	50.10	32.08	30.05	29.00	30.67	12.63
VideoChat3 4B	32	58.71	43.52	50.90	29.15	32.70	27.21	30.67	12.19
Qwen3-VL 4B	32	55.08	42.14	50.30	31.32	32.07	29.79	29.20	11.24
Molmo2 4B	32	59.87	42.14	51.51	29.34	30.68	23.93	27.07	10.70
LLaVA-NeXT-Video-7B	64	52.32	38.97	43.88	31.04	27.78	24.03	23.73	5.39
VideoGPT+ 4B	16	51.23	41.44	50.20	16.04	32.58	21.55	24.53	4.69
Video-LLaVA 7B	8	45.07	36.00	47.69	21.32	27.02	28.50	16.53	0.98
Motion-Specialized Model									
MotionLLM 7B	8	46.66	26.01	39.76	21.23	20.83	21.75	16.27	-4.94
Qwen3-VL 8B SFT	32	81.79	50.05	63.25	79.43	33.33	46.77	54.13	40.73
Qwen3-VL 4B SFT (Ours)	32	82.29	49.06	60.04	80.00	35.48	49.35	52.93	40.75
InternVL3.5-4B SFT	8	82.66	48.76	59.04	79.34	34.09	48.56	48.13	39.12

Table 10, among proprietary models Gemini-3.8-flash reaches 9.19, ahead of GPT-5.6-Luna at 8.51, while the text-only setting reaches 7.04. Open-source models are tightly grouped at the top of the zero-shot field, with LLaVA-OV-2 8B at 15.72, Qwen3-VL 4B at 15.30 and Qwen3-VL 8B at 15.18.

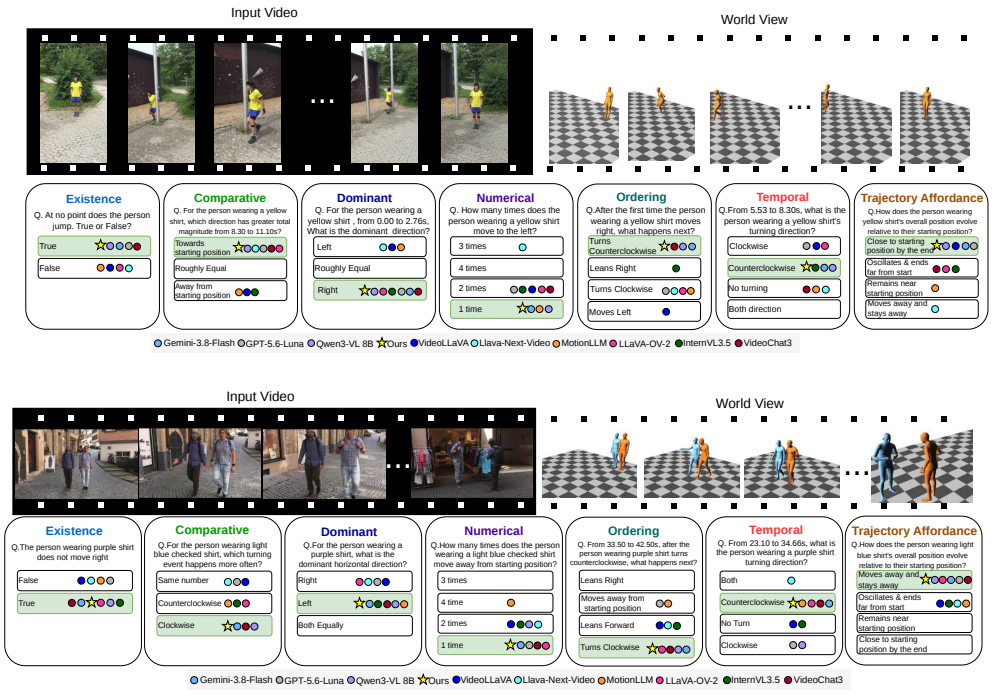


Figure 9: Qualitative results on the **3DPW** dataset. The world-view visualization depicts extracted 3D SMPL-X poses and is shown for illustration only (not provided to the models). Green denotes the correct option. We compare predictions from multiple MLLMs across seven reasoning categories, where our model demonstrates more accurate and consistent responses.

MotionLLM remains the worst performer at -8.21. Fine-tuning lifts our Qwen3-VL 4B SFT from 15.30 to 33.24, the smallest gain of the five splits, led by Numerical ($+33.33$ pp) and Existence ($+14.87$ pp), with Trajectory Affordance improving only $+4.30$ pp and Ordering the sole regression (-0.75 pp). The smaller gain likely reflects the subject-disjoint benchmark, on which the fine-tuned models are evaluated on people absent from training. Qwen3-VL 8B SFT score reaches 31.08 while InternVL3.5-4B SFT scores 28.72. Thus training on our dataset improves the performance of these MLLMs. Figure 10 shows qualitative results.

Takeaways – Across the five splits the best proprietary models are Gemini-3.8-flash and GPT-5.6-Luna, which trade places from split to split and average 10.22 and 9.69, while Gemini's text-only variant collapses to 5.26, confirming that the benchmark cannot be answered from the question text alone. Open-source models take the top zero-shot position on every split, led by Qwen3-VL 8B at a mean of 14.01 and LLaVA-OV-2 8B at 13.31, with MotionLLM last on all five at -7.32. Fine-tuning dominates both groups by a wide margin: our Qwen3-VL 4B SFT averages 38.96 and the Qwen3-VL 8B SFT 38.32, close to triple the best zero-shot model, while InternVL3.5-4B SFT reaches 33.93 on only 8 frames per video.

Figure 11 visualizes the quantitative results across datasets. We provide qualitative video results across all five datasets in the supplementary material.

A.2 Label Reliability

The Human Score reported in the main tables is chance-normalized as $(\text{Acc} - \text{chance})/(1 - \text{chance})$ and averaged over categories, where chance $= 1/k$ and k is the number of options. To understand the average human accuracy across the EMDB, EgoBody, RICH, 3DPW and SLOPER4D splits, we measure the per-category inter-annotator scores on the small subset we curate. Table 11 reports Fleiss' κ across three annotators together with the raw accuracy per category.

As shown in Table 11, human agreement is highest for Existence and Temporal questions ($\kappa \geq 0.8$), suggesting these categories are reliably interpretable. Lower agreement is observed for Ordering and Trajectory Affordance, motivating further inspection. Manual analysis indicates that disagreements

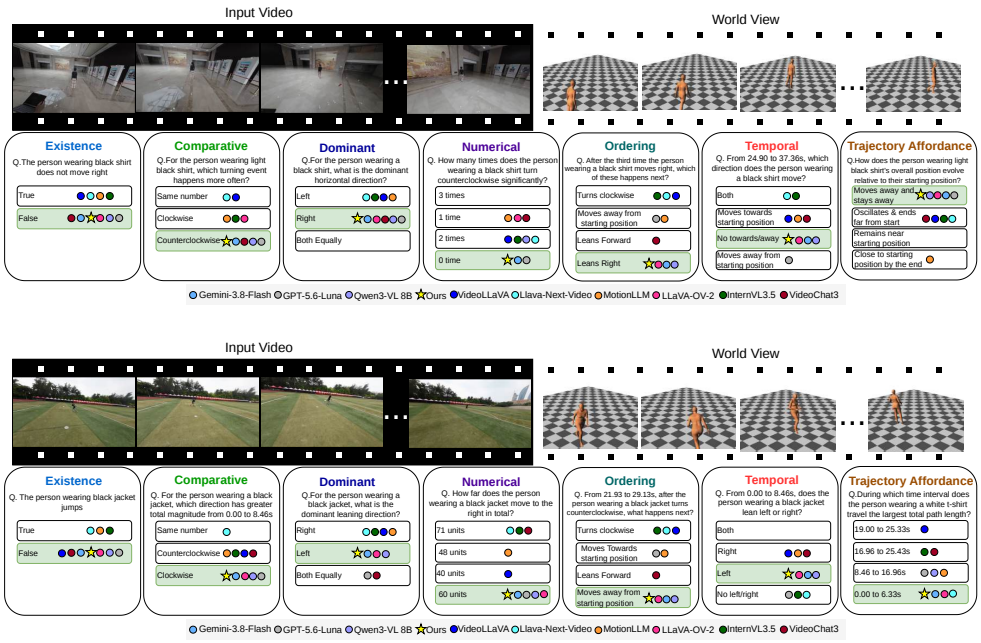


Figure 10: Qualitative results on the **SLOPER4D** dataset. The world-view visualization depicts extracted 3D SMPL-X poses and is shown for illustration only (not provided to the models). Green denotes the correct option. We compare predictions from multiple MLLMs across seven reasoning categories, where our model demonstrates more accurate and consistent responses.

Table 9: Evaluation of video MLLM models on the **3DPW** split. Results are reported across seven reasoning categories and an overall normalized score. Green, orange, and blue indicate the best, second-best, and third-best results per column, respectively.

Model	Frames	Existence	Comparative	Dominant	Numerical	Ordering	Temporal	Traj. Afford.	Score
Random Chance	–	50.00	33.33	33.33	25.00	25.00	25.00	25.00	0.00
Human (subset)	–	100.00	85.71	85.71	92.85	71.42	89.28	75.00	80.27
Proprietary Models									
GPT-5.6-Luna	32	66.40	38.89	49.72	25.26	16.47	36.32	25.00	9.97
Gemini-3.8-flash	2 FPS	64.78	35.56	30.94	37.37	31.18	31.58	26.47	9.26
Gemini-3.8-flash text	–	60.32	41.67	38.67	27.89	31.18	25.26	23.53	7.38
Open-Source MLLM									
LLaVA-OV-2 8B	32	76.11	41.11	49.17	37.37	28.82	33.68	29.41	18.10
Qwen3-VL 8B	32	68.42	40.00	48.62	31.05	31.18	27.89	34.56	14.67
VideoChat3 4B	32	64.78	46.11	50.83	25.26	30.00	30.53	33.82	14.45
Qwen3-VL 4B	32	63.16	37.78	50.83	30.00	32.35	26.32	33.82	12.75
InternVL3_5-8B	8	65.99	42.22	46.41	26.84	22.94	33.68	25.74	11.03
Molmo2 4B	32	62.75	43.33	49.17	26.84	33.53	26.32	21.32	10.71
LLaVA-NeXT-Video-7B	64	47.37	34.44	44.75	27.37	23.53	21.58	24.26	1.31
Video-LLaVA 7B	8	44.53	41.67	51.38	17.37	23.53	26.84	14.71	0.75
VideoGPT+ 4B	16	46.56	32.78	36.46	17.37	27.65	22.63	16.91	-3.37
Motion-Specialized Model									
MotionLLM 7B	8	41.70	31.11	35.36	19.47	21.18	17.37	8.82	-8.73
Qwen3-VL 8B SFT	32	86.23	48.33	58.56	74.74	33.53	47.37	51.47	39.37
Qwen3-VL 4B SFT (Ours)	32	82.19	44.44	64.64	77.89	34.12	49.47	48.53	39.24
InternVL3.5-4B SFT	8	76.11	43.33	55.25	75.79	32.35	42.11	50.00	33.39

primarily arise from temporal boundary cases, visually ambiguous trajectories, and occasional reconstruction artefacts, rather than widespread annotation inconsistency.

We further isolate the sources of error and label noise, which split into two kinds:

Table 10: Evaluation of video MLLM models on the **SLOPER4D** split. Results are reported across seven reasoning categories and an overall normalized score. Green, orange, and blue indicate the best, second-best, and third-best results per column, respectively.

Model	Frames	Existence	Comparative	Dominant	Numerical	Ordering	Temporal	Traj. Afford.	Score
Random Chance	–	50.00	33.33	33.33	25.00	25.00	25.00	25.00	0.00
Human (subset)	–	100.00	78.67	82.14	92.85	71.42	85.71	71.42	76.63
Proprietary Models									
GPT-5.6-Luna	32	59.49	38.46	42.76	35.33	18.66	25.68	34.41	8.51
Gemini-3.8-flash	2 FPS	51.28	34.97	40.69	46.00	25.37	30.41	34.41	9.19
Gemini-3.8-flash text	–	56.41	31.47	46.21	26.00	33.58	26.35	29.03	7.04
Open-Source MLLM									
LLaVA-OV-2 8B	32	63.59	30.07	45.52	54.67	30.60	27.03	39.78	15.72
Qwen3-VL 4B	32	57.44	30.77	44.14	50.00	35.82	31.08	43.01	15.30
Qwen3-VL 8B	32	62.56	27.97	42.76	41.33	36.57	31.08	47.31	15.18
VideoChat3 4B	32	57.44	33.57	52.41	35.33	33.58	24.32	32.26	11.12
InternVL3_5-8B	8	53.85	25.17	48.28	45.33	32.09	31.08	27.96	9.50
Molmo2 4B	32	57.44	38.46	50.34	27.33	30.60	21.62	25.81	7.89
LLaVA-NeXT-Video-7B	64	49.23	34.97	48.28	26.00	24.63	21.62	29.03	3.58
VideoGPT+ 4B	16	49.23	44.76	50.34	14.00	24.63	22.97	19.35	2.25
Video-LLaVA 7B	8	43.59	30.77	53.79	14.00	19.40	24.32	12.90	-3.59
Motion-Specialized Model									
MotionLLM 7B	8	46.15	22.38	36.55	20.00	25.37	14.19	11.83	-8.21
Qwen3-VL 8B SFT	32	70.26	40.56	50.34	80.00	36.57	45.95	43.01	31.08
Qwen3-VL 4B SFT (Ours)	32	72.31	44.76	51.03	83.33	35.07	42.57	47.31	33.24
InternVL3.5-4B SFT	8	70.26	44.76	46.21	82.67	37.31	36.49	36.56	28.72

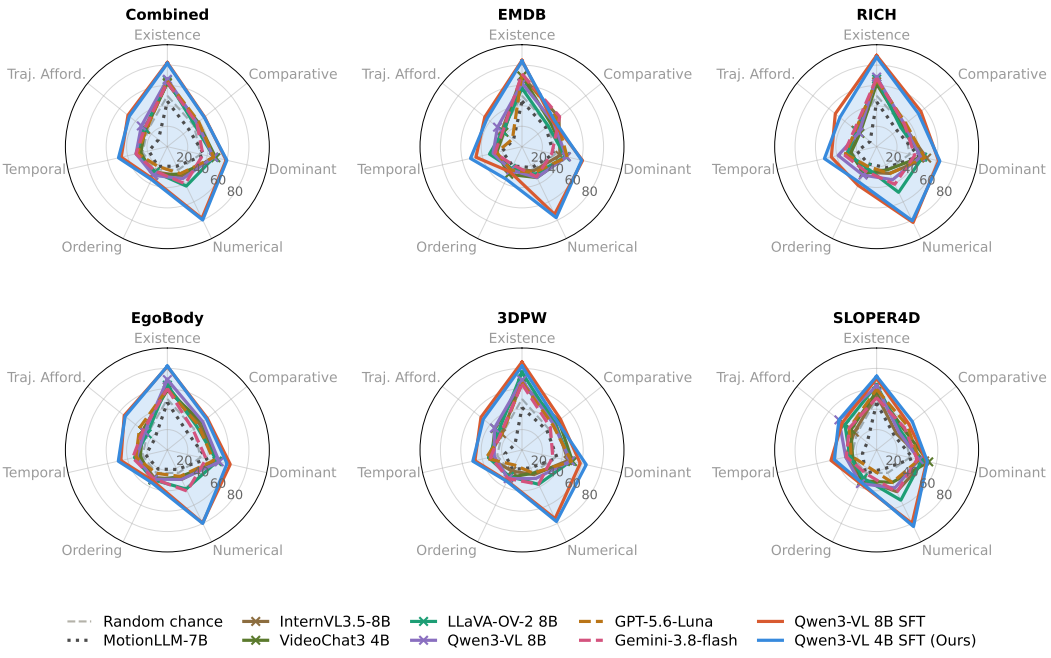


Figure 11: Performance comparison across reasoning categories on the HumanMoveVQAbenchmark. Radar plots show model accuracy on seven categories for Combined Avg and per dataset split (EMDB, RICH, EgoBody,3DPW,SLOPER4D) . Our model (Qwen3-VL 4B SFT) consistently outperforms baselines across most categories, with particularly strong gains in numerical, temporal, and trajectory reasoning.

1. **The label is wrong.**

Table 11: Human accuracy and inter-annotator agreement by category and dataset. Acc is raw accuracy; κ is Fleiss' kappa across three annotators; k is the number of options. The All row is the unweighted mean of the seven category values and is not the chance-normalized Score reported in the main tables.

Category	k	EMDB		EgoBody		RICH		3DPW		SLOPER4D	
		Acc	κ	Acc	κ	Acc	κ	Acc	κ	Acc	κ
Existence	2	100.00	1.00	100.00	1.00	100.00	1.00	100.00	1.00	100.00	1.00
Comparative	3	78.67	0.64	76.33	0.52	83.40	0.60	85.71	0.70	78.67	0.64
Dominant	3	83.33	0.64	83.67	0.60	88.56	0.70	85.71	0.66	82.14	0.62
Numerical	4	85.50	0.62	85.00	0.68	80.50	0.65	92.85	0.72	92.85	0.72
Ordering	4	75.50	0.48	85.00	0.70	75.50	0.57	71.42	0.45	71.42	0.45
Temporal	4	95.00	0.90	90.00	0.80	90.00	0.80	89.28	0.78	85.71	0.75
Traj. Afford.	4	75.00	0.50	75.00	0.60	75.50	0.55	75.00	0.55	71.42	0.45
All	–	84.71	0.68	85.00	0.70	84.78	0.70	85.71	0.69	83.17	0.66

   (a) *Reconstruction error.* We use PromptHMR, so the pipeline runs on any monocular video without motion capture. Reconstruction is implicit, so some tracks may carry error that transfers into the label.
   (b) *Discretization boundaries.* MotionScript converts numerical quantities into ordinal categories. A value near a boundary can fall into either bucket, e.g. 89 degrees and 91 degrees.

2. **The label is right and the annotator answered differently.**
   (a) *Question design.* Some categories apply tolerance rules. Near-ties are skipped and events spanning a timestamp interval are excluded, so ambiguous items are removed at generation time. An annotator reasoning without those tolerances can still land on a different answer for a borderline item.
   (b) *Reference frame.* Answers are defined relative to the person's position in the first frame, and the prompt states this. This framing differs from the camera-observed setting and requires tracking in case of a dynamic camera.
   (c) *Estimation difficulty.* Measuring how far a person exactly travelled, and which timestamp interval held the most movement, is hard for someone observing once. This is task difficulty, and it is the quantity the benchmark is meant to evaluate.

Kind 1 errors affect the label quality while Kind 2 errors are the ones the benchmark is designed to measure.

Reconstruction Error Measured Against Ground-Truth Tracks – This isolates Kind 1. We generate labels from ground-truth (GT) tracks and from PromptHMR tracks using the identical generator, then compare the answers. Every question whose answer changes is attributable to reconstruction error. Table 12 reports the flip rate by category and dataset.

From Table 12, the average flip rate is similar across datasets, ranging from 9.74% on RICH to 10.93% on EMDB. Existence is the most stable category across all four datasets (2–3%), as it only requires detecting whether an event occurs. In contrast, categories that depend on precise temporal or trajectory information exhibit higher sensitivity. Ordering has the highest flip rate on EMDB (15.84%), EgoBody (13.42%), and RICH (14.15%), while Trajectory Affordance is highest on SLOPER4D (20.00%). Comparative and Dominant questions are less affected because they depend primarily on relative differences rather than exact values. Overall, reconstruction changes the semantic answer for approximately 10% of comparable queries. The flip rate also captures the discretization effect described above, since reconstruction error changes a label only when it moves the underlying quantity across a decision boundary.

The flip rate bounds how much reconstruction error can matter, but it does not tell us whether any conclusion in the paper depends on it. To measure this directly, we rescore every model under both key sets and report the average Score across the four datasets. These two analyses isolate the effect of reconstruction noise on the benchmark. When evaluated using GT-derived rather than PromptHMR-derived keys, the zero-shot Qwen3-VL 4B score changes by only -0.21 points and Gemini-3.8-flash by -1.71 points, while the SFT checkpoint changes

Table 12: Reconstruction flip rate by category and dataset. Flip rate is the share of comparable anchored queries whose semantic answer changes between the PromptHMR track and the ground-truth track, under the same generator and the same frozen options.

Category	EMDB	EgoBody	RICH	SLOPER4D
Existence	2.00%	3.00%	2.5%	2.33%
Comparative	8.00%	10.00%	6.33%	7.45%
Dominant	10.5%	12.37%	8.88%	8.05%
Numerical	14.43%	10.90%	12.31%	8.09%
Ordering	15.84%	13.42%	14.15%	13.37%
Temporal	10.26%	8.29%	12.03%	15.07%
Traj. Afford.	15.45%	12.89%	12.00%	20.00%
All	10.93%	10.12%	9.74%	10.62%

Table 13: Model scores under PromptHMR-derived vs ground-truth-derived keys. Values are the mean chance-normalized Score across the EMDB, EgoBody, RICH and SLOPER4D splits.

Model	PromptHMR keys	GT keys	Δ
Qwen3-VL 4B SFT	38.22	33.82	-4.40
Qwen3-VL 4B	11.94	11.73	-0.21
Gemini-3.8-flash	11.45	9.74	-1.71

by -4.40 points. The larger sensitivity of the SFT checkpoint is consistent with its having been trained on PromptHMR-derived supervision. Importantly, the main SFT improvement remains under GT-derived keys: the SFT model scores 33.82 compared with 11.73 for the corresponding zero-shot Qwen3-VL 4B model. Thus, while reconstruction introduces measurable label noise, the principal conclusions of our benchmark remain unchanged when evaluated against ground-truth-derived labels.

Camera Viewpoint Sensitivity – Our training data and benchmark labels are built entirely from 3D world-space tracks and discretization using MotionScript. The labels therefore do not change with camera placement, camera motion, or apparent motion magnitude in the video. In terms of a model answering these questions from RGB, our per-dataset results show the performance. EMDB,SLOPER4D and 3DPW have large displacements captured by a moving camera, while EgoBody and RICH are static multi-view captures. Table 14 groups the five splits by camera type and number of people, and Tables 15a and 15b break the multi-view sets down by camera name in EgoBody and RICH. The camera names are descriptive and not fixed across captures. Across all groups, we observe a significant SFT gain over zero-shot Qwen3-VL 4B inference.

Table 14: Camera setup and person count per split, with the aggregate Score for Qwen3-VL 4B zero-shot and our joint SFT model. Score is the chance-normalized aggregate averaged over the seven categories.

Split	Camera	People	Zero-shot	SFT
EMDB	moving	single	10.10	37.70
RICH	static multi-view	single	11.86	43.85
EgoBody	static multi-view	two	11.24	40.75
3DPW	moving	one or two	12.75	39.24
SLOPER4D	moving	single	15.30	33.24

The zero-shot model is strongest on SLOPER4D (15.30) and 3DPW (12.75), the two splits with the largest apparent motion, and weakest on EMDB (10.10). Fine-tuning reverses this ordering: the largest gains are on RICH ($+31.99$) and EgoBody ($+29.51$), and the smallest on SLOPER4D ($+17.94$). Camera type alone does not explain the pattern, since the moving-camera splits span both

Table 15: View-averaged Scores on the two multi-view splits. Performance of Qwen3-VL 4B zero-shot and our joint SFT model, broken down by camera view on EgoBody (a) and by camera on RICH (b). Score is the chance-normalized aggregate averaged over the seven categories.

View	Zero-shot	SFT	Gain
master	14.92	41.82	+26.90
sub1	14.93	47.59	+32.67
sub2	10.66	39.40	+28.74
Mean	13.50	42.94	+29.44

(a) EgoBody

Camera	Zero-shot	SFT	Gain
cam_00	14.28	45.11	+30.83
cam_01	21.94	38.32	+16.38
cam_02	15.22	45.55	+30.33
cam_03	9.60	45.17	+35.57
cam_04	13.90	44.43	+30.53
cam_05	5.79	40.40	+34.61
cam_06	6.24	44.64	+38.40
cam_10	6.91	45.37	+38.46
Mean	11.73	43.62	+31.89

(b) RICH

the smallest gain (SLOPER4D) and a large one (EMDB, $+27.60$). EgoBody and 3DPW contain videos with two people, yet both also show substantial gains after fine-tuning.

Since the target labels are defined in world space, differences across camera views reflect the model's sensitivity to the RGB observation rather than changes in the underlying labels. This is particularly evident on RICH, where zero-shot performance varies from 5.79 to 21.94, a spread of 16.15 points across views. After fine-tuning, this spread reduces to 7.23 points (38.32–45.55). On EgoBody, zero-shot performance varies from 10.66 to 14.93, while SFT performance ranges from 39.40 to 47.59. More importantly, every individual view improves substantially, with gains from $+26.90$ to $+32.67$ on EgoBody and $+16.38$ to $+38.46$ on RICH. This shows that the improvement from world-space supervision is not tied to a particular camera view, while the strong viewpoint sensitivity observed zero-shot on RICH is substantially reduced after fine-tuning.

A.3 Confidence Intervals and Robustness

The benchmark contains $12{,}282$ questions from 127 clips. Since questions from the same clip share the same video and reconstruction, we treat the clip rather than each question as the independent unit. We bootstrap clips with replacement, keeping all questions from each sampled clip, and use $10{,}000$ resamples to compute the confidence intervals.

We report both micro and macro results. Micro pools questions across clips, while macro first computes accuracy per clip and then averages across clips. Both are chance-normalized and averaged over the seven categories. For model comparisons, we use a paired bootstrap with the same sampled clips for both models and compute the confidence interval directly on their difference.

Table 16 reports the gain of our Qwen3-VL 4B SFT model over its zero-shot base and the two strongest baselines. The gain holds across every split, with all confidence intervals above zero. Against the zero-shot base, the pooled gain is $+28.62$ $[+26.39, +30.84]$, with the smallest gain on SLOPER4D at $+17.94$ $[+10.72, +24.42]$. We also observe gains of $+28.58$ over Gemini-3.8-flash and $+26.37$ over LLaVA-OV-2 8B. Micro and macro estimates remain within 0.8 points, showing that the gains are not driven by clips with more questions.

Pooled and Split-Averaged Scores – The pooled bootstrap gain ($+28.62$) differs from the main-table gain ($+26.71$) because the former pools questions across all splits, while the latter averages the Score across datasets. Overall, the gains remain consistent across datasets and after accounting for questions sharing the same clips, showing that the improvement is systematic rather than an artefact of question-level aggregation.

A.4 Ablation

We investigate the following six questions here in the appendix:

Table 16: Paired cluster-bootstrap confidence intervals for SFT gains. Clips are treated as the independent unit and resampled with replacement using 10,000 resamples. Micro and macro report question-level and clip-level aggregation, respectively.

Comparison	Scope	Clips	Micro Δ [95% CI]	Macro Δ [95% CI]
SFT − Qwen3-VL 4B	All five splits	127	$+28.62$ $[+26.39, +30.84]$	$+28.77$ $[+26.44, +31.02]$
SFT − Gemini-3.8-flash	All five splits	127	$+28.58$ $[+26.27, +30.87]$	$+28.77$ $[+26.25, +31.24]$
SFT − LLaVA-OV-2 8B	All five splits	127	$+26.37$ $[+24.00, +28.67]$	$+27.02$ $[+24.64, +29.40]$
SFT − Qwen3-VL 4B	EMDB	11	$+27.60$ $[+21.28, +33.51]$	$+27.59$ $[+21.26, +33.46]$
SFT − Qwen3-VL 4B	RICH	32	$+31.99$ $[+27.67, +36.21]$	$+32.61$ $[+28.07, +37.10]$
SFT − Qwen3-VL 4B	EgoBody	54	$+29.51$ $[+26.48, +32.59]$	$+30.10$ $[+27.02, +33.21]$
SFT − Qwen3-VL 4B	3DPW	15	$+26.50$ $[+19.58, +34.06]$	$+27.32$ $[+20.23, +34.96]$
SFT − Qwen3-VL 4B	SLOPER4D	15	$+17.94$ $[+10.72, +24.42]$	$+18.13$ $[+10.79, +24.69]$

Table 17: Effect of frame count and input resolution. Performance of Qwen3-VL 4B fine-tuned (SFT) with varying numbers of video frames (a) and varying input resolution (b). Each setting is trained and evaluated at the same frame count or resolution. Mean is the chance-normalized aggregate averaged over the five splits.

Dataset	16	32	64
EMDB	36.04	37.70	37.05
RICH	43.13	43.85	42.79
EgoBody	39.02	40.75	40.55
3DPW	36.61	39.24	38.78
SLOPER4D	31.26	33.24	32.62
Mean	37.21	38.96	38.36

(a) Frame Count

Dataset	128	256	384
EMDB	31.01	37.70	36.41
RICH	42.11	43.85	42.86
EgoBody	38.82	40.75	39.97
3DPW	34.44	39.24	38.26
SLOPER4D	32.54	33.24	34.88
Mean	36.98	38.96	38.48

(b) Input Resolution ($n \times n$)

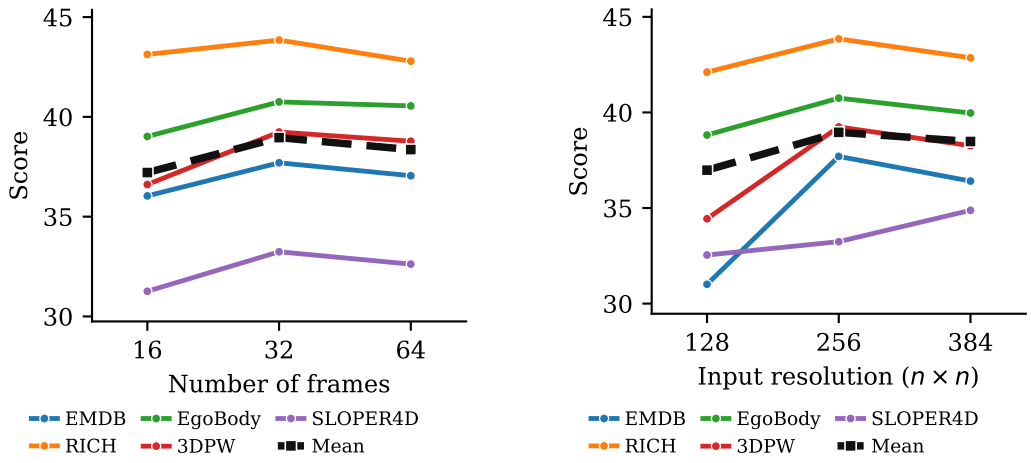
1. How sensitive is the training to the number of frames?
2. How sensitive is the training to the resolution of the videos?
3. How does it perform if trained with a short reasoning trace?
4. How does it perform if the questions are rephrased?
5. How does our model generalize to joint-level reasoning benchmarks like ActionArt?
6. Does the length of the source video affect the score?

Effect of Frame Count – We evaluated models trained with 16, 32, and 64 uniformly sampled frames to test the impact of temporal resolution. Table 17a and Figure 12a report the aggregate Score for each setting.

32 frames is the best setting on all five splits and 16 frames the worst, so the choice does not trade one split off against another. The gain from 16 to 32 frames is uneven, from $+0.72$ points on RICH to $+2.63$ on 3DPW, with EgoBody ($+1.73$) and SLOPER4D ($+1.98$) in between. Increasing temporal resolution to 64 frames does not improve results. EgoBody score drops by 0.20 points. Overall, SFT training on 32 frames reaches 38.96, ahead of 64 at 38.36 and 16 at 37.21. More frames likely dilute the amount of useful signal.

Effect of Resolution – We ablated input resolutions (128×128, 256×256, 384×384) at 32 frames. Table 17b and Figure 12b report the aggregate Score for each setting.

Four of the five splits peak at 256×256. SLOPER4D is the exception, peaking at 384×384, and is also the least responsive split, gaining only 0.70 points from 128×128 to 256×256. The low-resolution penalty is far larger elsewhere, at 6.69 points on EMDB and 4.80 on 3DPW, with RICH (1.74) and EgoBody (1.93) in between. Past 256, the four splits that peak there give back only a little, between 0.78 points on EgoBody and 1.29 on EMDB, so the curve rises sharply and then flattens rather than degrading symmetrically. Overall SFT training with image resolution 256×256



(a) Frame count: 16, 32 and 64 uniformly sampled frames.

(b) Input resolution: 128×128, 256×256 and 384×384 at 32 frames.

Figure 12: Effect of the training input budget. Qwen3-VL 4B is fine-tuned (SFT) at each setting and evaluated at the matching setting. The coloured lines are the per-dataset Scores and the dashed black line is their mean.

is the best setting at 38.96, against 38.48 at 384×384 and 36.98 at 128×128. Higher resolution likely dilutes the motion signal by amplifying background tokens.

Reasoning Supervision – We tested whether adding a short reasoning trace during SFT improves performance. To do this, the SFT Reason is trained with the same training samples and LoRA configuration. However we add a short reasoning trace to the training samples using the same *SpatialCodes* that we have. It follows the template of <observe></observe><thinking></thinking><answer>X</answer> and the reasoning trace has structured keys describing the relevant motion before coming to an answer.

Table 18: Reasoning supervision. Results comparing SFT fine-tuning with answer options only and with short reasoning traces.

Dataset	SFT	SFT Reason
EMDB	37.70	35.01
RICH	43.85	38.87
EgoBody	40.75	36.17
3DPW	39.24	31.91
SLOPER4D	33.24	26.32

Table 18 shows that this "SFT Reason" model underperforms the option-only SFT on every split, by 7.70 points on EMDB, 12.98 on RICH, 9.58 on EgoBody, 14.33 on 3DPW and 6.93 on SLOPER4D, a mean drop of 10.3 points. The added supervision raises the entropy of the output. The option-only model returns one of four strings of nine characters, whereas the SFT Reason model produces a trace of roughly 1,100 characters and 96% of its responses are unique, so it must commit to a long high-entropy span before it reaches the option and the variation accumulated in that span propagates into the final choice. Figure 15 shows examples of output reasoning traces. The model was trained with similar reasoning traces.

Template Robustness – We evaluate whether the model answers the task or the template. We sample up to three questions per video and category to build a subset of 2,622 questions across the five splits, then reword each question while holding the video, the options and the gold answer fixed. Rewriting uses Qwen3.5-9B. We run inference on both the original and the paraphrased questions with our Qwen3-VL 4B SFT model and compare the paired answers. Table 19 reports the result.

The score moves from 57.39 to 55.93. Most answers remain the same. Scores on RICH remain unchanged, SLOPER4D improves slightly while EMDB loses 4.79 points. SFT gains are not carried by template language

Table 19: Template robustness: original versus paraphrased questions, paired on the same rows. Accuracy is the percentage of rows answered correctly; Δ is paraphrased minus original.

Dataset	N	Original	Paraphrased	Δ
EMDB	231	58.87	54.08	-4.79
RICH	663	60.78	60.78	$+0.00$
EgoBody	1,116	56.54	54.48	-2.06
3DPW	312	59.29	57.05	-2.24
SLOPER4D	297	49.83	50.84	$+1.01$
All	2,619	57.39	55.93	-1.46

Out-of-Domain Evaluation – We evaluated our Qwen3-VL 4B SFT model (trained on HumanMoveVQA) on the ActionArt Peng et al. (2025) benchmark, which is the closest existing benchmark with similar categories such as Action Direction (AD), Action Count (AC), Action Sequence (AS), Global Spatial (GS), Local Spatial (LS), and Temporal Localization (TL), among others, but focuses on joint-level pose reasoning. The results are shown in Table 20.

Table 20: Evaluation on the ActionArt benchmark. The ActionArt score is taken from the original paper, while all other results are computed in our evaluation. Column names follow the ActionArt convention.

Model	AC	MV	AS	HO	AR	GS	LS	TL	Avg
VideoLLaVA	33.33	37.33	37.93	55.68	38.67	62.18	34.22	36.28	41.95
MotionLLM	24.60	32.96	27.27	28.09	23.47	52.30	21.37	21.58	28.96
VideoGPT+	37.30	44.44	44.03	59.55	56.59	61.51	52.73	31.71	48.48
ActionArt	45.20	59.60	61.80	83.10	82.60	85.50	75.70	42.50	69.40
Qwen3VL-4B	52.80	48.74	53.35	72.29	71.57	83.40	66.10	39.08	60.92
Qwen3VL-4B SFT	40.00	39.50	44.41	46.99	43.46	77.55	43.44	37.54	46.61

Our model scores 46.61, lower than the zero-shot Qwen3-VL 4B baseline, but substantially higher than MotionLLM (28.96). The per-category results provide a clearer picture of this difference. Our supervision contains no joint-level pose or object-interaction information, and correspondingly Action Recognition, Human-Object Interaction, and Local Spatial reasoning drop after fine-tuning, while Global Spatial reasoning remains strong. The degradation is therefore not uniform across the eight subtasks, but concentrated in categories requiring information absent from our supervision.

This suggests that trajectory-level supervision does not substitute for fine-grained joint-pose and object-interaction information. Instead, trajectory-level and joint-level understanding are complementary skills requiring distinct supervision signals. Our SFT teaches the model to reason about whole-body displacement, orientation, and temporal sequencing, while ActionArt additionally requires fine-grained joint and object-level understanding.

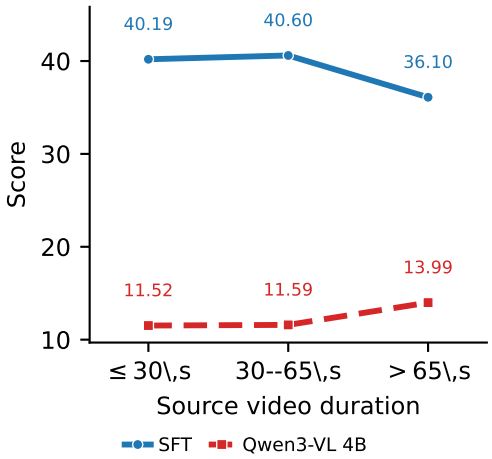
Effect of Video Length – Training and evaluation both sample a fixed 32 frames uniformly across the whole clip, so the interval between sampled frames grows with clip length, from roughly a third of a second on the shortest clips to three and a half seconds on the longest. Table 21 bins the 127 evaluation clips by the duration of their source video, recomputes the Score on the questions of the clips in each bin, and plots the pooled Scores in its right panel.

The two shorter bins are close, at 40.19 for clips up to 30 seconds and 40.60 for clips between 30 and 65 seconds, and the longest bin drops to 36.10. The bins are not balanced: they hold 13, 105 and 9 clips, and the longest carries only 628 questions. They also differ in which splits they draw from, since the shortest bin holds only EMDB, EgoBody and 3DPW clips and the longest only EMDB and RICH clips. Part of the drop in the longest bin is therefore a change of dataset rather than an effect of length.

Table 21: **Effect of source video length.** Clips are binned by source video duration and Score is recomputed per bin using the main-table metric. Both models are Qwen3-VL 4B: the SFT model is the combined five-split run. We report within-split comparisons in the text.

Duration	Clips	Questions	SFT	Zero-shot
≤ 30 s	13	1,371	40.19	11.52
30–65 s	105	10,283	40.60	11.59
> 65 s	9	628	36.10	13.99
All	127	12,282	40.35	11.73

(a) Score by duration bin



(b) Pooled Score by duration bin

Comparing within a single split removes this confound. The Score falls from 46.16 to 35.53 within RICH, from 42.62 to 40.51 within EgoBody and from 40.08 to 38.88 within 3DPW, so longer clips are harder in every split that spans two bins. EMDB does not follow this pattern. It is the only split that spans all three bins, and it is lowest on its shortest bin (31.52) and below its middle bin on its longest (37.84), though those two cells hold only 3 and 2 clips.

The zero-shot base moves the other way, rising from 11.52 and 11.59 in the two shorter bins to 13.99 in the longest. The longest clips are therefore easier for the base model and harder for the fine-tuned one, and the gain from fine-tuning is largest on medium-length clips (+29.01) and smallest on the longest (+22.11).

A.5 Units and Discretization

This section explains the units used in Tables 24 and 25. The band values reported in these tables are ordinal counts and not physical measurements.

World-space quantities – All motion quantities are computed from 3D world-space root translation and orientation, not from 2D image-space displacement. The pipeline lifts the subject to world coordinates and all thresholds are applied to these world-space quantities, so pixels are never used when generating labels. The labels are therefore defined in the world frame and do not depend on the camera observing the motion. Root translation is measured in metres and root orientation in degrees.

Discretization – Following MotionScript Yazdian et al. (2025), continuous translation and orientation are first discretized into categorical bins. Translation is measured in units where 1 unit $= 10$ cm, so the displacement thresholds in Table 24 correspond to a range of ± 1 m, and rotation is measured in degrees. A motion segment is then assigned a spatial band and a temporal band by comparing its accumulated displacement or rotation against those thresholds.

Velocity – The velocity carried on each motion event is the accumulated displacement per frame, and the temporal thresholds in Table 25 are applied to that quantity. The table reports each threshold both in the stored per-frame units and in metres per second, using the 10 cm unit and the fixed 30 fps frame rate.

Frame rate – All videos, posecodes, and frame-level velocities are defined at a fixed 30 fps.

Table 22: Per-dataset motion statistics over the set of Figure 3, one entry per video and view-averaged for the multi-view datasets(EgoBody and RICH) . Displacement and rotation events are mean motion event counts per video; Median path is the median geometric length of the XZ ground-plane trajectory.Green denotes the highest value in the column.

Dataset	Videos	Displacement events	Rotation events	Median path (m)	Mean duration (s)
EMDB	75	44.2	19.6	13.3	37.0
RICH	138	26.7	7.5	5.1	20.8
EgoBody	133	34.2	7.0	5.1	28.7
3DPW	63	32.4	9.9	9.0	38.3
SLOPER4D	36	23.9	8.9	44.0	42.6

Table 23: Distractor generation rules by category.

Category	Correct answer from	Distractors
Existence	Yes or No from event count, epsilon, or dominance predicate	Logical complement only
Comparative	Event count, total magnitude, or first-event speed level	Fixed 3-way of direction A, direction B, tie. Tie if difference $\leq 0.2 \cdot \max$, tie rate capped at 30%
Dominant	Summed absolute intensity per direction	Fixed 3-way of direction A, direction B, both equally. Tie if difference $< 0.2 \cdot \max$, tie rate capped at 10%
Numerical	Filtered event count, or rounded sum of absolute intensities	Count: 3 ints in $[c-3, c+3]$ excluding c. Magnitude: 3 ints in $c \pm \max(0.2c, 10)$ excluding c
Ordering	Earliest event B with $B.t_0 \geq A.t_1 + 10$ frames in the same interval	3 real alternate event labels from the same interval; the four options hold 2 displacement and 2 rotation labels
Temporal	Movement within a timestamped interval, the first movement in an interval, or the first-event speed bin	Interval: all four interval labels. First motion: 2 displacement and 2 rotation labels observed in the video. Speed: 3 of the other 4 bins
Traj. Afford.	Argmax or argmin interval metric, endpoint bias, or shape rule	All four intervals, or the full mutually exclusive endpoint and shape taxonomy

Table 24: Spatial codes and their discretization into categorical bins. Translation is measured in units, where 1 unit $= 10$ cm, and rotation is measured in degrees. See Section A.5 for the full unit convention.

Axis	**Category**	**Range (units)**
Translation (Displacement)		
X-axis	very_long_left	$v < -10$
	long_left	$-10 \leq v < -8$
	moderate_left	$-8 \leq v < -5$
	short_left	$-5 \leq v < -3$
	very_short_left	$-3 \leq v < -1$
	no_action	$-1 \leq v \leq 1$
	very_short_right	$1 < v \leq 3$
	short_right	$3 < v \leq 5$
	moderate_right	$5 < v \leq 8$
	long_right	$8 < v \leq 10$
	very_long_right	$v > 10$
Y-axis	very_long_down	$v < -10$
	long_down	$-10 \leq v < -8$
	moderate_down	$-8 \leq v < -5$
	short_down	$-5 \leq v < -3$
	very_short_down	$-3 \leq v < -1$

Axis	Category	Range (units)
	no_action	$-1 \leq v \leq 1$
	very_short_up	$1 < v \leq 3$
	short_up	$3 < v \leq 5$
	moderate_up	$5 < v \leq 8$
	long_up	$8 < v \leq 10$
	very_long_up	$v > 10$
Z-axis	very_long_backward	$v < -10$
	long_backward	$-10 \leq v < -8$
	moderate_backward	$-8 \leq v < -5$
	short_backward	$-5 \leq v < -3$
	very_short_backward	$-3 \leq v < -1$
	no_action	$-1 \leq v \leq 1$
	very_short_forward	$1 < v \leq 3$
	short_forward	$3 < v \leq 5$
	moderate_forward	$5 < v \leq 8$
	long_forward	$8 < v \leq 10$
	very_long_forward	$v > 10$
Rotation (Orientation)		
Pitch	significant_leaning_backward	$v < -4$
	moderate_leaning_backward	$-4 \leq v < -3$
	slight_leaning_backward	$-3 \leq v < -2$
	no_action	$-2 \leq v \leq 0$
	slight_leaning_forward	$0 < v \leq 2$
	moderate_leaning_forward	$2 < v \leq 3$
	significant_leaning_forward	$v > 3$
Roll	significant_leaning_right	$v < -4$
	moderate_leaning_right	$-4 \leq v < -3$
	slight_leaning_right	$-3 \leq v < -2$
	no_action	$-2 \leq v \leq 0$
	slight_leaning_left	$0 < v \leq 2$
	moderate_leaning_left	$2 < v \leq 3$
	significant_leaning_left	$v > 3$
Yaw	significant_turn_clockwise	$v < -4$
	moderate_turn_clockwise	$-4 \leq v < -3$
	slight_turn_clockwise	$-3 \leq v < -2$
	no_action	$-2 \leq v \leq 0$
	slight_turn_counterclockwise	$0 < v \leq 2$
	moderate_turn_counterclockwise	$2 < v \leq 3$
	significant_turn_counterclockwise	$v > 3$

A.6 Broader Impact

This work introduces a benchmark for evaluating human trajectory and orientation reasoning with Video MLLMs. The dataset and benchmark can support the development of downstream models for applications such as sports analytics and industrial safety. However, models evaluated on this dataset could potentially be misused for human surveillance or behavior analysis in sensitive contexts. To

System Prompt
You are answering multiple-choice questions about a person's motion in a video.

Motion is described in a fixed world coordinate system:

- Motion is measured relative to the first frame of the video.
- The vertical axis (**Y**) points upward from the ground.
- The horizontal axis (**X**) points left–right relative to the person in the first frame.
- The depth axis (**Z**) points forward–backward relative to the person in the first frame.
- Turning left or right refers to rotation of the person's body, not camera motion.

Use the explicit timestamps in the question when reasoning about temporal windows. If the question identifies a person by appearance, answer only about that person.

Select the single best provided option. Respond exactly as "Answer: X", where X is the option letter. Do not provide an explanation.

Figure 13: System prompt used for training and evaluation of models.

Table 25: Temporal category thresholds. See Section A.5 for the unit convention.

Temporal category	Velocity range	m/s
very_slow	≤ 0.05	≤ 0.150
slow	$(0.05, 0.10]$	$(0.150, 0.300]$
moderate	$(0.10, 0.50]$	$(0.300, 1.500]$
fast	$(0.50, 0.80]$	$(1.500, 2.400]$
very_fast	> 0.80	> 2.400

mitigate this, the dataset only provides annotations and references to source data, requiring users to obtain the original datasets and comply with their respective licenses.

Table 27: Overview of the seven categories, highlighting the capability tested in each, along with a representative example question.

Category	Sub-axis	Example question
existence	Semantic existence	Is there walking in the video?
	Directional existence	Does the person move to the left at any point?
	Negation (logical complement)	The person did not move to the left at any point. True or false?
comparative	Count-based comparison	Which event happens more often overall: moving left or moving right?
	Magnitude-based comparison	Which direction has greater total magnitude overall: left or right?
	Speed-based comparison	Is the person's first move left faster or slower than their first move right?
dominant	Direction dominance (global)	What is the dominant horizontal movement direction, left or right?
	Direction dominance (interval-local)	From 3.16 to 6.33s, What is the dominant turning direction, clockwise or counterclockwise?
numerical	Event count (direction-specific)	How many times does the person move to the left?
	Total magnitude (direction-specific)	How far does the person move to the left in total?
	Strength-restricted count (moderate/significant)	How many times does the person turn clockwise significantly?

Category	Sub-axis	Example question
ordering	Next event after anchor within interval	From 0.56 to 3.43s, after the person leans to the left, which of these happens next?
temporal	Interval-local motion	From 9.50 to 12.70s, what horizontal movement does the person make?
	First event in interval	From 6.33 to 9.50s, what is the first movement or rotation that happens?
	Event speed classification	During the person's first move to the left, how fast is the movement?
trajectory affordance	Largest net displacement (interval)	During which time interval does the person move most to the right (largest net rightward displacement)?
	Smallest net displacement (interval)	Among the intervals where the person actually moved, during which interval was that displacement smallest?
	Largest path length (interval)	During which time interval does the person travel the largest total path length (sum of horizontal and depth movement)?
	Overall endpoint position	Where is the person horizontally relative to where they started at the end of the video?
	Trajectory shape	How does the person's overall position evolve relative to their starting position?

Table 26: Per-category vs. joint training. Performance of Qwen3-VL 4B fine-tuned (SFT) on individual categories and evaluated across datasets. Green, orange, and blue denote the best, second-best, and third-best results per dataset, respectively. Zero-shot indicates no task-specific training, while All denotes joint training on all categories.

Dataset	Setting	Existence	Comparative	Dominant	Numerical	Ordering	Temporal	Traj. Afford.	Score
EMDB	Zero-shot	66.43	44.55	38.18	31.82	27.27	24.55	26.67	10.10
	Ordering	65.03	39.09	44.55	31.82	32.73	26.36	21.33	10.26
	Numerical	62.94	36.36	56.36	77.27	35.45	40.00	22.67	23.64
	Comp.+Dom.	66.43	40.91	56.36	29.09	37.27	28.18	34.67	16.82
	Temporal	69.23	36.36	41.82	36.36	31.82	39.09	24.00	13.92
	Traj. Afford.	70.63	39.09	48.18	29.09	26.36	20.91	52.00	15.71
	All	83.92	40.91	60.00	77.27	35.45	51.82	44.00	37.70
RICH	Zero-shot	63.22	41.96	45.57	30.31	38.54	28.30	21.81	11.86
	Ordering	63.22	38.49	49.68	30.63	40.28	29.56	20.99	12.47
	Numerical	64.42	33.12	48.10	81.56	36.11	41.51	20.16	22.35
	Comp.+Dom.	62.74	47.63	60.44	25.62	30.56	30.82	29.63	15.68
	Temporal	66.83	46.37	53.16	45.94	30.21	46.86	26.34	21.25
	Traj. Afford.	63.70	35.96	45.57	24.38	28.47	22.01	48.56	11.56
	All	87.74	52.37	63.61	80.62	38.89	52.52	46.09	43.85
EgoBody	Zero-shot	55.08	42.14	50.30	31.32	32.07	29.79	29.20	11.24
	Ordering	55.59	41.54	52.91	31.23	34.09	29.69	29.07	12.14
	Numerical	64.73	39.17	47.59	80.38	28.91	40.62	19.73	21.78
	Comp.+Dom.	50.94	48.27	57.73	28.87	30.68	31.38	27.20	12.15
	Temporal	62.99	45.80	47.69	49.81	31.31	47.07	28.40	20.24
	Traj. Afford.	59.22	41.74	48.90	20.28	27.78	21.25	51.07	11.65
	All	82.29	49.06	60.04	80.00	35.48	49.35	52.93	40.75
3DPW	Zero-shot	63.16	37.78	50.83	30.00	32.35	26.32	33.82	12.75
	Ordering	63.97	32.78	45.86	27.37	33.53	22.11	27.94	8.64
	Numerical	61.54	38.89	41.44	79.47	34.71	31.58	22.79	19.28
	Comp.+Dom.	56.28	40.00	56.35	28.42	27.06	28.42	30.88	10.97
	Temporal	65.99	43.33	45.86	36.32	26.47	44.74	25.74	15.73
	Traj. Afford.	66.80	43.33	53.59	20.00	27.06	23.16	48.53	14.86
	All	82.19	44.44	64.64	77.89	34.12	49.47	48.53	39.24
SLOPER4D	Zero-shot	57.44	30.77	44.14	50.00	35.82	31.08	43.01	15.30
	Ordering	60.51	38.46	44.83	39.33	31.34	29.73	43.01	14.83
	Numerical	60.51	41.26	48.97	82.67	25.37	31.08	18.28	18.99
	Comp.+Dom.	62.05	39.86	50.34	24.67	21.64	32.43	39.78	12.01
	Temporal	66.67	38.46	35.86	54.00	22.39	37.84	34.41	15.67
	Traj. Afford.	56.92	34.97	46.21	28.67	19.40	23.65	49.46	9.12
	All	72.31	44.76	51.03	83.33	35.07	42.57	47.31	33.24

Clothing Captioning Prompt

Describe this person's clothing in a short noun phrase with NO leading verb (e.g. "a purple shirt and dark navy trousers", NOT "wearing a..." or "a person wearing..."). Mention the upper-body garment and the lower-body garment, their colors, and any visible pattern (check, plaid, striped, pinstripe, printed) if present. Do NOT mention shoes, sneakers, boots, sandals, or any footwear. Do NOT mention the person, pose, body position, camera, or background. Answer with just the noun phrase, under 60 characters.

Figure 14: Prompt used for clothing caption generation.

Reasoning traces (three representative categories)

Comparative

Question: Which leaning direction has greater total magnitude overall: right or left? *Options:* A. roughly equal B. left C. right

```
<observe>[rotation_roll(-7.00,0.21,significant_leaning_right,moderate,1.60,2.66),
...
  motion_total(7.00,-,left,-,1.60,15.33),

motion_total(18.00,-,right,-,1.60,15.33),tie_threshold(3.60,-,-,-,1.60,15.33)]</observe>
<thinking>
Step 1:  Objective | I need to determine which leaning direction has greater total
...
Step 2:  Start perception | At the beginning of the relevant interval, around 1.6s,
...
Step 3:  Temporal progression | The first lean right covers about 7.0 units, ...
Step 4:  End perception | By about 15.3s, the motion settles after the person has
...
Step 5:  Final synthesis | Given this motion sequence, the best matching option ...
</thinking>
<answer>C</answer>
```

Ground truth: C

Ordering

Question: From 8.16 to 13.63s, after the person moves right, which of these happens next? *Options:* A. moves left B. leans to the right C. leans to the left D. moves towards starting position

```
<observe>[displacement_x(10.00,0.27,long_right,moderate,8.16,9.36),
  displacement_x(-7.00,0.20,moderate_left,moderate,13.63,14.76)]</observe>
<thinking>
Step 1:  Objective | I need to determine from 8.16 to 13.63s, after the person
moves ...
Step 2:  Start perception | Within 8.16 to 13.63s, the sequence begins with the ...
Step 3:  Temporal progression | The motion unfolds in order:  first from about 8.2s
...
Step 4:  End perception | By the end of the sequence, the person has moved left.
Step 5:  Final synthesis | Putting the sequence together, the action that follows
the ...
</thinking>
<answer>A</answer>
```

Ground truth: A

Trajectory Affordance

Question: During which time interval does the person move most to the right (largest net rightward displacement)? *Options:* A. 9.23 to 13.83s B. 0.00 to 4.60s C. 4.60 to 9.23s D. 13.83 to 18.46s

```
<observe>[displacement_x(17.00,0.14,very_long_right,moderate,5.33,9.36), ...
  window_total(10.00,-,rightward,-,4.60,9.23), ...
  window_total(7.00,-,rightward,-,13.83,18.46)]</observe>
<thinking>
Step 1:  Objective | I need to determine during which time interval does the person
...
Step 2:  Start perception | At the beginning of the relevant interval, around 5.3s,
...
Step 3:  Temporal progression | Adding up the movement window by window:  between
...
Step 4:  End perception | By about 18.1s, the motion settles after the person has
...
Step 5:  Final synthesis | Because the interval with the largest rightward
displacement ...
</thinking>
<answer>C</answer>
```

Ground truth: C

Figure 15: Examples of the reasoning traces used as training targets, for three representative categories, on EMDB clip `01_mvs_b.mp4`. For each, we show the question, the answer choices, the trace the model is trained to produce, and the ground-truth answer. Each step is condensed onto one line as its heading followed by the opening of its text, long text is elided with an ellipsis, and the observe block is shown wrapped to fit.