HUMANPHYS: A HUMAN-CENTRIC PHYSICAL-WORLD PERCEPTION BENCHMARK FOR REAL-WORLD VISUAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce HumanPhys, a human-centric benchmark for real-world physical perception, and a training corpus emitted by the same pipeline. Each question targets a physically grounded quantity related to people, including depth, inter-person distance, gaze angle, displacement, velocity, heading-step trajectory, blink count, and contact-deformation strength. The rendered probes additionally test illumination inference from body and face cues and camera ego-motion, both with and without a person in the frame. All labels are derived programmatically from sensor calibration, source metadata, or renderer state, rather than written by annotators or produced by estimators on the benchmark frames. HumanPhys systematically probes six dimensions of human-centric physical perception: spatial state, kinematics, blink events, deformation, illumination, and camera pose. To the best of our knowledge, no existing MLLM benchmark-and-training corpus jointly supervises this set of human-centered physical quantities under direct metadata- or renderer-derived labels. The six dimensions appear only partially dissociable across the tested MLLMs. We benchmark a range of open- and closed-source MLLMs on HumanPhys and find that all six capability dimensions remain challenging for current models. We hope HumanPhys will support future research on multimodal systems that perceive and reason about people in the physical world.

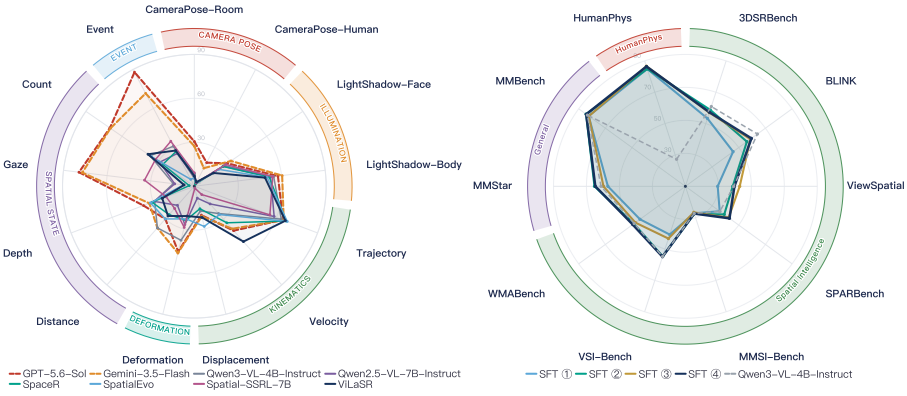


Figure 1: **Left**: radar of the eight MLLMs on HumanPhys, with one axis for each sub-task, grouped into the six taxonomy categories of Figure 2; the two proprietary models are dashed (numerical scores in Table 3). **Right**: checkpoint ablation of fine-tuning on HumanPhys and complementary spatial data across ten benchmarks, with Qwen3-VL-4B-Instruct dashed as reference; this data-mixture ablation is tabulated in Table 5 and analysed in §4.3 (Preserving General Spatial Abilities).

1 INTRODUCTION

The recent rise of spatial intelligence benchmarks has raised a natural question: can Multimodal Large Language Models (MLLMs) [2, 13, 23, 25, 62, 67], which perform strongly on 2D visual understanding, also perceive and reason about the physical world? Recent work has consequently explored physical spatial perception from object attributes (count, size) to inter-object

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.

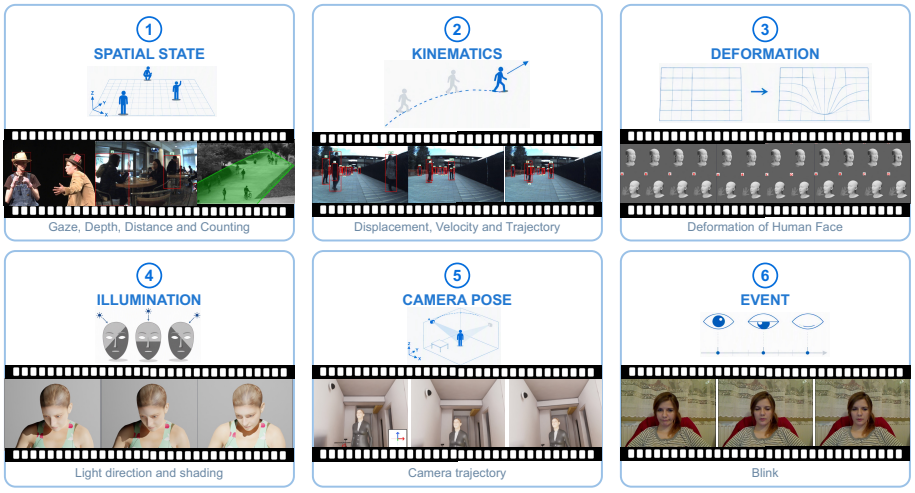


Figure 2: Overview of the HumanPhys taxonomy.

relations (position, distance) [19, 27, 40, 56, 59, 60]. These efforts have progressively expanded spatial-intelligence evaluation from semantic recognition toward explicit geometric understanding of objects and their spatial configurations. In parallel, recent training efforts have tried to elicit such reasoning through reinforcement learning over spatial rewards or tool-augmented inference [4, 14, 24, 38, 43, 57].

However, understanding the physical world requires more than object and scene geometry: humans are dynamic physical entities whose states and temporal dynamics must be perceived quantitatively. Fine-grained cues such as facial deformation under hand contact, eye blinking, and illumination on the face or body provide additional physical information, and public human-centric resources already capture some of these cues at the frame level [5, 47, 48], but recovering such quantitative physical information from visual observations remains beyond semantic recognition.

Despite this progress, coverage of quantitative human physical perception remains narrow and uneven rather than absent: video benchmarks mainly emphasize activities, actions, and events [7, 18, 28, 37, 65]; spatial benchmarks focus on object- and scene-centric attributes and relations [21, 32, 63], and even when humans appear they usually serve as semantic subjects or action performers rather than as physical entities whose states can be measured quantitatively, so it remains unclear whether MLLMs can reliably recover such physically grounded information from real-world videos. HumanMoveVQA [22] is the closest prior benchmark to ours and focuses on geometric reasoning about the root translation and orientation of a single target person through discretized multiple-choice labels derived from estimated mesh tracks (§2.2). Existing benchmarks still do not jointly measure a broader range of human-centered physical quantities in scenes, including metric relations between people, transient physiological events, non-rigid responses to contact, illumination inference from body and face cues, and camera ego-motion with a person in the frame. HumanPhys instead evaluates whether an MLLM can recover such quantities from visual input. The ground truth is derived directly from metadata or renderer state, rather than written by annotators or produced by estimators on the benchmark frames.

To bridge this gap, we introduce **HumanPhys**, a benchmark and corresponding corpus for evaluating human-centered physical perception in MLLMs on real-world and rendered videos. HumanPhys systematically evaluates models across multiple dimensions of physically grounded perception about people, spanning spatial state, kinematics, blink events, deformation, illumination, and camera pose. HumanPhys therefore tests whether MLLMs can move beyond semantic description toward quantitative perception of people in the physical world.

In summary, our contributions are as follows. **(1) Benchmark and training corpus.** We construct and release HumanPhys, a 13-sub-task benchmark spanning eight real and rendered sources, together with a 386K-sample training corpus generated by the same pipeline. All ground truth is derived directly from sensor calibration, 3D-box metadata, or renderer state, rather than written by annotators or produced by estimators on the benchmark frames. **(2) Evaluation and transfer.** We benchmark a diverse set of open- and closed-source MLLMs on HumanPhys and find that it

Table 1: Comparison of representative spatial reasoning benchmarks. **Human-Centric**: humans are the targets of perception; **Input Type**: accepted modalities, "Video" means temporally ordered sampled frames derived from source clips; **Source**: provenance of the visual data, real or simulated; **Task Taxonomy**: static or dynamic queried target; **Train Data Ava.**: whether a training corpus is released. For HumanPhys, Static and Dynamic correspond to its frame-level and temporal questions (§3.1). For HumanMoveVQA all questions are video-level multiple-choice questions about human motion, so Static is marked ✗, and *Train Data Ava.* denotes the authors' own training split rather than a separately released corpus; it is compared with HumanPhys axis by axis in §2.2 and Appendix B.

Benchmark	Human-Centric	Input Type		Source		Task Taxonomy		Train Data Ava.
		Image	Video	Real	Simulation	Static	Dynamic	
VSI-Bench [59]	✗	✗	✓	✓	✗	✓	✗	✗
STI-Bench [32]	✗	✗	✓	✓	✗	✓	✓	✗
3DSRBench [40]	✗	✓	✗	✓	✓	✓	✗	✗
BLINK [19]	✗	✓	✗	✓	✓	✓	✗	✗
ViewSpatial [27]	✗	✓	✗	✓	✗	✓	✗	✗
SPARBench [63]	✗	✓	✗	✓	✗	✓	✗	✗
MMSI-Bench [60]	✗	✓	✗	✓	✗	✓	✓	✗
WMABench [21]	✗	✓	✓	✗	✓	✓	✓	✗
SpaceR [43]	✗	✗	✓	✓	✗	✓	✗	✓
SpatialScore [56]	✗	✓	✓	✓	✓	✓	✓	✓
HumanMoveVQA [22]	✓	✗	✓	✓	✗	✗	✓	✓
HumanPhys	✓	✓	✓	✓	✓	✓	✓	✓

remains challenging: overall scores span 19.74–46.94, and the four checkpoints tuned on existing spatial corpora reach only 20.43–27.75. The best model lands close to an indicative, non-matched human reference, so we treat that comparison as context only. Supervising on the HumanPhys corpus also transfers outward, lifting Qwen3-VL-4B-Instruct by 45.16 and 41.57 points on the single- and multi-view distance sub-tasks of SPAR-Bench (Table 4). **(3) Post-training experiments.** We conduct preliminary post-training experiments on HumanPhys and study whether mixing complementary public spatial corpora [43, 56] helps preserve general spatial ability.

2 Related Work

2.1 Spatial Intelligence

Spatial intelligence probes whether MLLMs can recover geometric and physical properties beyond semantic recognition: object positions, sizes, orientations, distances and metric QA [16, 19, 27, 34, 40, 42, 54, 59]; simulator- or 3D-annotation-based supervision enables scalable construction and fine-tuning [6, 9, 33, 35, 52, 54–56]; region-level cues support object-centric analysis [10]; video extensions infer layouts and dynamics from frame sequences [31, 32, 43, 53, 59]; and recent benchmarks stress spatial intelligence under visual degradation and low-level embodied viewpoints [64, 66].

Despite these advances, existing spatial data remain predominantly object- or scene-centric, including comprehensive consolidations such as SpatialScore [56], and therefore rarely treat humans as physical entities whose states, motion, and interactions can be quantitatively perceived from videos; HumanPhys is designed to target this gap. Concretely, SpatialScore consolidates 30 task categories and releases a 331K-sample training corpus that includes counting, depth estimation, object distance, object size, object location, object motion, and camera pose and motion, but its measurement target remains object and scene geometry rather than properties of the human body. Inter-person metric distance, physiological event counts, contact-deformation strength, and hidden-light localization fall outside the supervision scope of SpatialCorpus and SpaceR-151K, which is why we use both corpora as external complements in §4.3 rather than position HumanPhys as a competing breadth-oriented resource.

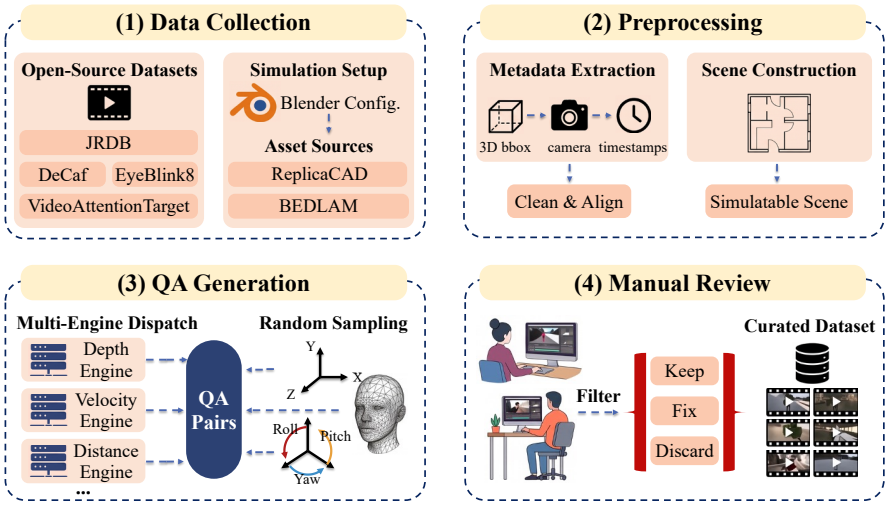


Figure 3: The construction pipeline of HumanPhys: (1) data collection from open-source datasets and Blender-based simulation; (2) preprocessing with metadata extraction and scene construction; (3) multi-engine QA generation with random sampling; (4) manual review and iterative refinement.

2.2 Human-Centric Perception and Understanding

Existing human-centric video benchmarks mainly target semantic, behavioral, or cognitive understanding, while HumanMoveVQA [22] is the closest prior work to ours because it asks geometric questions about human motion. HumanMoveVQA evaluates global trajectory and orientation reasoning for a single target person from exocentric video through discretized multiple-choice questions derived from estimated SMPL-X tracks. HumanPhys is complementary rather than redundant and differs from HumanMoveVQA in measured quantities, answer format and scoring, ground-truth provenance, and generalization protocol, as detailed in Table 7 of Appendix B. Table 1 positions HumanPhys among them.

3 HumanPhys

HumanPhys evaluates whether an MLLM can recover physical information from human-centric images and videos along six dimensions: spatial state, kinematics, event, deformation, illumination (light–shadow), and camera pose (Figure 2). Every answer is computed programmatically from source metadata and renderer state, including 3D boxes, camera parameters, timestamps, and source annotations for the real-world datasets and full scene state for the synthetic ones, so the ground truth is never written by a human annotator viewing the same frames as the model (provenance for each sub-task in §3.1 and Appendix A). HumanPhys spans **13 sub-tasks** over **8 sources** (5 real-world video datasets and 3 rendered scene collections), and includes a **2,500-question evaluation split drawn by equal per-sub-task quota sampling (200 questions, 100 for *Deform*)**, plus a **385,702-question** training pool emitted by the same construction pipeline that shares no question with the split. Quotas are sampled at the *question* level, not the scene level, so scenes may recur in both; the eight models are nevertheless all evaluated zero-shot.

The six dimensions are not merely a topic list but a compact operational decomposition of how a human body functions in visual physical perception, separated by the role the body plays: (a) a *rigid target* localized and tracked in the world frame (spatial state, kinematics); (b) a *non-rigid, living target* showing transient physiological events and contact response (blink events, deformation); (c) a *photometric probe* whose facial and body shading reveals a hidden light source (illumination); and (d) a *reference for the observer*, since camera ego-motion is measured with and without a person in frame (camera pose). The unifying criterion is measurability: every answer is a quantity in physical units or a discrete event count read programmatically from metadata or renderer state, and every sub-task is anchored on a person (*Count* counts humans under an explicit metric

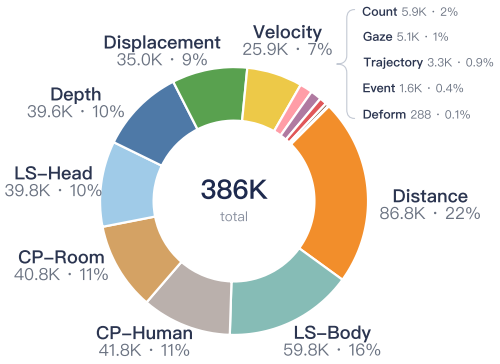


Figure 4: Task distribution of the HumanPhys training corpus (about 386K samples), distance-related tasks dominate, with long-tail categories such as Count, Gaze, Trajectory and Deformation.

Group	Dimension	Sub-task	Source	Input	Format
I	Spatial state	Depth	JRDB-Activity	frame/video	Num
		Distance	JRDB-Activity	frame/video	Num
		Count	DanceTrack, UCSD	frame/video	Num
		Gaze	VideoAttentionTarget	frame	Num
	Kinematics	Displacement	JRDB-Activity	video	Num
		Velocity	JRDB-Activity	video	Num
		Trajectory	JRDB-Activity	video	Tuple
	Event	Blink	EyeBlink8	video	Num
II	Deformation	Deform	DECAF (FLAME, MANO)	video	MC
	Illumination	LS-Head	200 BEDLAM identities	video	Tuple
		LS-Body	BEDLAM, 8 wall cams	video	Tuple
	Camera pose	CP-Room (ctrl)	ReplicaCAD (84 stages)	video	Tuple
		CP-Human (anchored)	ReplicaCAD (84 stages)	video	Tuple

Table 2: **HumanPhys composition. Group**: I, measured human cues on real-world footage; II, controlled rendered probes for phenomena real footage cannot be labelled for reliably. **Dimension**: taxonomy category of Figure 2; **Sub-task**: the queried quantity; **Source**: dataset or renderer. **Input**: *frame*, one image; *video*, ordered frames. **Format**: Num, numeric scored by MRA; MC, option accuracy; Tuple, sequence scorers of §4.2.

criterion, *LS-Head*/*LS-Body* localize light from shading on skin and clothing, *CP-Human* asks for ego-motion while a person anchors the frame). Two properties make the taxonomy diagnostic rather than merely broad: the evaluated dimensions show different cross-model profiles (§4.2), and the camera-pose pair serves as a controlled human-presence comparison (§4.4), with *CP-Room* retained as the human-free control condition.

3.1 Task Definition

The 13 sub-tasks fall into two groups with complementary roles. **Group I covers *measured* human-centered physical cues in real-world video**: depth, inter-person distance, counting, gaze, displacement, velocity, trajectory and blink events, each recovered from the sensor metadata or annotations of actual footage, which keeps the benchmark tied to real visual data. **Group II provides *controlled probes*** for phenomena that cannot be reliably labeled in wild footage: the strength of hand–face contact deformation, the lattice walk of a hidden point light source observable only through body and facial shading and cast-shadow motion, and the sequence of atomic camera actions underlying an ego-motion clip. None of these is directly observable or reliably annotatable in real video; rendering (or, for *Deform*, re-rendering FLAME/MANO panels over DECAF's fits to real footage) makes the scene state the exact ground truth and lets us control each sub-task's difficulty. In short, Group I tests whether a model can recover physical cues from real people, while Group II isolates cues whose ground truth is otherwise difficult to obtain. Table 2 summarises sources, input types and answer formats; the evaluation split keeps the equal quota stated above (*Count*: 69 frame, 131 video questions).

Group I: real-world measured human-centered physical cues.

- **Spatial state**: *Depth* — distance of people to the reference sensor; *Distance* — pairwise inter-person distance at the foot point; *Count* — humans satisfying an explicitly stated spatial criterion; *Gaze* — clockwise image-plane angle from upward to the head-centre–to–fixation vector.
- **Kinematics**: *Displacement* — cumulative ground-plane path length of a tracklet; *Velocity* — that path length over the clip's elapsed time; *Trajectory* — heading changes and step lengths as (turn_deg, distance_m) tuples.
- **Event**: *Blink* — number of discrete eye-blink events in a clip, integrating evidence across frames.

Group II: controlled physically rendered probes.

Table 3: **Main results on HumanPhys.** Scores of eight MLLMs on the 13 sub-tasks (best per column among the eight models in bold), grouped by the six taxonomy dimensions of Figure 2. Column keys: Dist = Distance, Disp = Displacement, Traj = Trajectory, LS = light-shadow, CP = camera-pose. Row keys: SSRL = Spatial-SSRL-7B; SpaceR/SpatialEvo/SSRL/ViLaSR are Qwen2.5-VL-7B-Instruct fine-tunes; the shaded group rows separate the closed-source (blue) from the open-source (green) checkpoints. "Overall" (our primary metric) is the question-weighted composite $\sum_k n_k \bar{s}_k / 2500$ over the full item bank (n_k=200; 100 for *Deform*), scoring an answer that cannot be parsed into the required type 0 and keeping it in the denominator, as `lmms-eval` does; the 13 sub-task columns instead average only the answers a model produced in parseable form, so they are diagnostic rather than additive, exceeding "Overall" by up to 4.06 points where truncated generations are frequent. The *Human* row is indicative only: ten colleagues each answered an availability-selected, non-uniform ≈40% subset (mean times in Table 6, Appendix A); the *Random* row is a uniform random baseline over the full question bank; the *Text Only* row answers from the question text alone, without any visual input.

	Spatial state				Kinematics			Event	Deform.	Illumination		Camera pose		
Model	Depth	Dist	Count	Gaze	Disp	Vel	Traj	Blink	Deform	LS-Head	LS-Body	CP-Room	CP-Human	Overall
Random	7.20	8.30	23.40	28.20	5.80	12.20	37.80	22.00	26.20	13.40	48.10	6.10	2.80	18.60
Text Only	20.60	24.00	0.00	32.55	0.00	14.80	54.03	0.00	28.00	24.58	54.19	4.62	2.32	18.03
Human	25.20	52.00	79.20	60.40	36.50	30.30	63.50	87.40	36.10	50.60	70.10	24.90	12.20	48.30
Closed-source models														
GPT-5.6-Sol	**37.15**	29.90	**70.05**	**79.35**	20.35	40.50	64.72	**87.95**	46.00	27.51	57.48	**30.55**	**18.15**	**46.94**
Gemini-3.5-Flash	32.90	**38.05**	69.60	77.05	21.15	38.85	65.28	71.60	**47.00**	**30.39**	**60.39**	27.36	13.91	45.61
Open-source models														
Qwen3-VL-4B	29.00	**38.05**	36.65	14.50	17.40	24.95	63.48	30.85	38.00	24.82	54.62	5.36	2.43	27.27
Qwen2.5-VL-7B	29.25	17.40	28.80	13.50	8.40	16.15	58.29	25.20	27.00	23.94	52.48	6.44	2.56	19.74
SpaceR	30.90	25.95	26.15	3.50	15.85	33.40	67.07	26.85	24.00	23.78	51.54	8.44	1.68	26.14
SpatialEvo	32.10	29.85	35.70	7.45	**28.30**	25.90	**67.85**	5.30	23.00	20.27	50.07	5.60	2.86	25.21
SSRL	20.55	20.30	32.20	34.30	1.40	7.60	55.14	34.65	29.00	26.20	53.97	9.17	2.57	20.43
ViLaSR	23.45	27.15	38.45	7.05	22.00	**50.55**	66.51	27.55	21.00	15.97	48.58	6.95	3.20	27.75

- **Deformation** (*Deform*): among a 2×4 panel of a FLAME head deformed by MANO hand contact at four strengths, pick the strength whose deformation best matches the physically plausible one.
- **Illumination** (*LS-Head*, *LS-Body*): infer the lattice-coordinate walk of a single hidden point light from shadow motion on the face (slow single-axis walks) or on the body and floor (fast multi-axis walks).
- **Camera pose** (*CP-Room*, *CP-Human*): recover the sequence of atomic camera actions — lattice translations or in-place rotations; *CP-Human* places a SMPL-X human and samples paths that keep the face framed, while *CP-Room* is the corresponding human-free control rendered with the same scene family, lattice step, action grammar, and scorer, so the CP-Human/CP-Room gap is used as a controlled comparison for the effect of human presence on ego-motion recovery (§4.4) rather than leaving *CP-Room* as a generic spatial task appended to the human-centered ones.

Appendix A provides detailed construction procedures for each sub-task, including ego-motion filtering, coordinate and angle conventions, blink-window compression.

3.2 Benchmark Construction

As illustrated in Figure 3, the construction of HumanPhys follows a four-stage pipeline: data collection, preprocessing, QA generation, and manual review. In the first stage, we aggregate open-source human-centric datasets (e.g., JRDB [17], DECAF [47], EyeBlink8 [48], VideoAttentionTarget [11], DanceTrack [49] and UCSD [8]) together with synthetic scenes configured in Blender and using assets from sources such as ReplicaCAD [50] in Habitat-Sim [46] and BEDLAM [5]. The collected videos are then preprocessed by extracting metadata (3D bounding boxes, camera parameters, and timestamps) and constructing simulatable scenes, followed by cleaning and alignment. Finally, because answers are derived from source metadata or renderer state rather than written by annotators, manual review only flags rendering artefacts, mis-aligned metadata or ill-posed prompts for regeneration and never alters a ground-truth value. Figure 4 shows the training-corpus task distribution.

4 EXPERIMENTS

4.1 SETUP

Training setup. We use Qwen3-VL-4B-Instruct as the base model and fine-tune it for one epoch with a cosine scheduler, 5% warm-up, a global batch size of 1024, and a learning rate of 1×10^{-5} on 256 Ascend HPU-910C.

Inference setup. We implement all inference on top of lmms-eval, extending it with prompt templates and answer parsers for each sub-task. Every model uses a generation budget of up to 1,024 new tokens and default sampling; open-source checkpoints (Qwen3-VL-4B-Instruct [2], Qwen2.5-VL [3], SpaceR [43], Spatial-SSRL [38], SpatialEvo [26] and ViLaSR [57]) are served locally and proprietary models (GPT-5.6-Sol, Gemini-3.5-Flash) queried through official APIs.

4.2 EVALUATION

For **Multiple-choice answers** (*Deform*), we assign a score of 1 to an exact option match and 0 otherwise. **Numeric answers** are scored by mean relative accuracy over a threshold set, as in VSI-Bench [59]:

$$C = \{0.50, 0.55, \ldots, 0.95\}, \qquad \text{MRA} = \frac{1}{|C|} \sum_{\theta \in C} \mathbb{1}\left[\frac{|\hat{y} - y|}{|y|} < 1 - \theta \right],$$

This grants partial credit for continuous quantities while assigning 0 to missing or unparseable predictions.

Trajectory tuples accumulate headings from a shared origin and score each step by the law-of-cosines vector difference

$$e_i = \sqrt{a_i^2 + b_i^2 - 2a_i b_i \cos(h_i^{\text{gt}} - h_i^{\text{pred}})}, \qquad \bar{e} = \frac{1}{m}\Big(\sum_{i < \min(n,m)} e_i + \sum_{i=n}^{m-1} a_i \Big), \qquad s = \frac{1}{1+\bar{e}},$$

Accumulating headings from a shared origin fixes the world frame of every step, so an early turn error compounds into later positions; m and n are the gold and predicted step counts, so missing steps are charged the full gold step length and extra steps are ignored. Each step yields an independent credit in $[0, 1]$ and the sequence score is their mean.

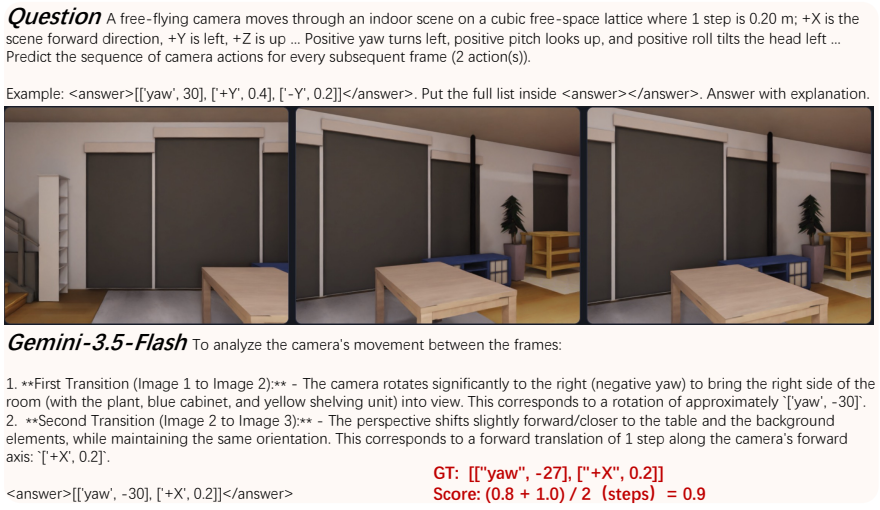


Figure 5: **One scored *CP-Room* QA.** Step 1 shares the gold axis (yaw) but not the magnitude ($-30°$ vs. $-27°$), passing 8 of the 10 MRA thresholds for $8/10 = 0.8$; Step 2 matches exactly for 1.0; the sequence score averages over gold steps: $(0.8 + 1.0)/2 = 0.9$.

Table 4: **Transfer to existing benchmarks for related sub-tasks.** Scores of Qwen3-VL-4B-Instruct before (Baseline) and after training with HumanPhys only (SFT ① in Table 5) on selected sub-tasks of SPAR-Bench and MMSI-Bench; ↑ marks the gain over the baseline. The consistent gains on distance and depth sub-tasks indicate that HumanPhys supervision transfers effectively to related tasks in existing benchmarks, while the change on *Motion(Cam.)* (22.97 vs. 32.43, i.e. 17/74 vs. 24/74 correct; CLT 1σ of ± 4.89 and ± 5.44 pp, 95% CI $[-4.9, +23.8]$, so this n=74 category is not individually significant) is suggestive only. `_mv` marks the multi-view variant of the sub-task.

Benchmark	Subtask	Baseline	SFT
SPAR-Bench	`distance_prediction_oc`	11.20	56.36 ↑
	`distance_prediction_oc_mv`	18.05	59.62 ↑
	`depth_prediction_oc`	33.47	59.39 ↑
	`depth_prediction_oc_mv`	37.66	58.53 ↑
MMSI-Bench	`Motion(Cam.)`	22.97	32.43 ↑

For **Camera-pose actions**, we assign 0 when the predicted axis is wrong; otherwise, the magnitude is scored using the MRA rule (CP-Room translations ignore magnitude once the direction is right), averaged over the gold step count; we additionally report the axis hit rate. Figure 5 shows one scored *CP-Room* QA.

Light-shadow coordinates are scored by

$$s = \mathrm{clip}\left(1 - \frac{\bar{e}}{2\,C_{\mathrm{mode}}},\, 0,\, 1\right), \qquad C_{\mathrm{mode}} = 1 \text{ for LS-Head, } \sqrt{3} \text{ for LS-Body,}$$

where $\bar{e}$ is the mean per-step Euclidean coordinate error in lattice units and C_{mode} is the maximum legal step length of the walk mode: 1 for the slow single-axis *LS-Head* walk and $\sqrt{3}$ for the fast *LS-Body* walk, whose steps may change up to three lattice axes at once. The clip keeps every score in $[0, 1]$, charging any prediction whose mean error exceeds $2\,C_{\mathrm{mode}}$ the full zero.

Taken together, the five scorers are deterministic, bounded in $[0, 1]$, and require no human judgment at scoring time. This design also makes HumanPhys a plausible candidate for reward-based post-training, because the same scoring functions used for evaluation could in principle serve as reward signals for RLVR-style training [1, 12, 15, 20, 30, 36, 39, 41, 51, 58, 61] on HumanPhys.

Object-centric spatial training transfers only weakly. The four checkpoints released for object- and scene-centric spatial reasoning reach 20.43–27.75 overall, 19–27 points below GPT-5.6-Sol. Human-physical perception is therefore at best weakly induced by existing spatial corpora, motivating the HumanPhys training pool and our use of SpaceR-151K and SpatialCorpus-331K as complements in §4.3.

The six dimensions appear partially dissociable. Across the eight models, kinematics and illumination correlate at $r = -0.18$ and spatial state and kinematics at $r = 0.35$ (mean off-diagonal $r = 0.64$), and rankings reverse between dimensions (e.g. SpatialEvo is strongest on displacement and trajectory yet is the weakest of all eight on blink events, while Spatial-SSRL leads the open-

Table 5: **Data-mixture ablation.** Scores (best per column in bold) of the Qwen3-VL-4B-Instruct base model and four SFT checkpoints on HumanPhys, seven spatial-intelligence benchmarks and two general multimodal benchmarks. The Data columns mark which training corpora each run mixes (HP = HumanPhys corpus, SpR = SpaceR-151K, SpC = SpatialCorpus-331K); SFT ④ is the final model.

	Data			HumanPhys	Spatial intelligence							General	
Model	HP	SpR	SpC	Overall	3DSRB.	BLINK	ViewSp.	SPAR	MMSI	VSI	WMA	MMStar	MMB.
Qwen3-VL-4B	✗	✗	✗	27.27	**60.91**	**64.01**	39.29	35.81	27.30	54.74	48.37	61.18	83.51
SFT ①	✓	✗	✗	84.92	53.31	45.78	29.59	35.71	27.20	40.77	44.06	57.00	82.56
SFT ②	✓	✓	✗	84.89	58.79	56.11	38.73	39.09	**27.50**	54.65	48.59	**65.13**	**84.45**
SFT ③	✓	✗	✓	**86.58**	57.03	58.73	**42.96**	42.37	26.60	43.30	47.48	59.72	82.39
SFT ④	✓	✓	✓	86.40	57.28	59.48	39.76	**43.05**	27.40	**54.90**	**49.46**	64.64	84.36

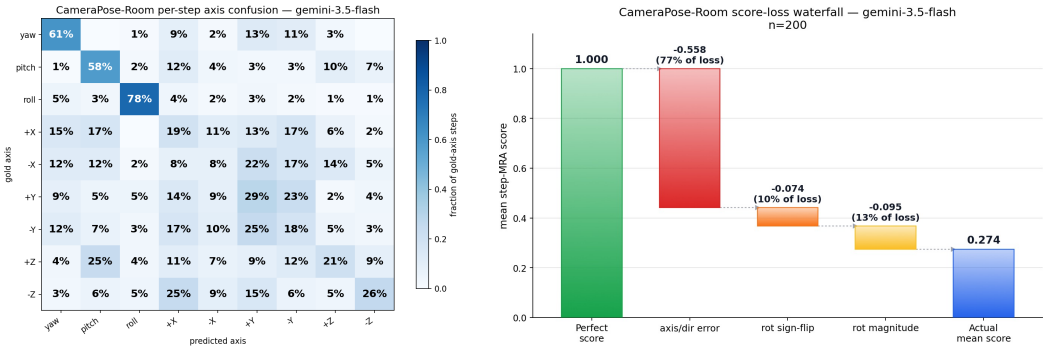


Figure 6: **Diagnosis of *CP-Room* errors (Gemini-3.5-Flash,** $n = 200$**).** Left: per-step gold-vs-predicted axis confusion. Right: score-loss waterfall from a perfect mean step-MRA of 1.0 to the actual 0.274.

source models on gaze and *CP-Room*). An aggregate score would hide these differences; the profiles of Figure 1 attribute each failure to a measurement target.

4.3 Preserving General Spatial Abilities

Supervised training on the HumanPhys corpus introduces a clear specialization trade-off: SFT ① (HumanPhys only) raises Qwen3-VL-4B-Instruct from 27.27 to 84.92 on HumanPhys but degrades general spatial benchmarks relative to the untouched base model (e.g. 45.78 vs. 64.01 on BLINK, 40.77 vs. 54.74 on VSI-Bench; Table 5), because human-centric cues alone shift the model away from broader object- and scene-centric supervision. To preserve it we mix SpaceR-151K [43] and SpatialCorpus-331K [56] into the pool, separately and jointly. The joint mixture, SFT ④, reaches 86.40 on HumanPhys, 1.48 above SFT ①, while matching or exceeding the base model on seven of the nine non-HumanPhys benchmarks and reducing the BLINK gap from 18.23 to 4.53: HumanPhys supervision is highly effective for the target capability but too narrow on its own to preserve broad spatial reasoning.

4.4 Diagnosis

Figure 6 analyses the failure modes of Gemini-3.5-Flash on *CP-Room* (n = 200). Rotational axes are recovered much more reliably than translational ones (diagonal confusion 77%, 64%, 61% for roll, pitch, yaw); translational motions are often confused with rotations (e.g. gold +X predicted as yaw in 28% of steps). The score breakdown attributes most of the error to choosing the wrong axis or motion type (69% of the total loss), then sign errors (22%) and only then magnitude errors (9%): identifying the motion type, not its magnitude, is the main challenge.

5 Conclusion

We introduced HumanPhys, a human-centric evaluation suite and training corpus for physical perception from images and videos, spanning 13 sub-tasks with labels derived programmatically from source metadata or renderer state. Our experiments show that these capabilities remain challenging for current MLLMs, while training on HumanPhys substantially improves held-out performance and benefits from being combined with complementary spatial data. More broadly, HumanPhys provides a structured setting for studying how multimodal models perceive people as physical entities rather than only as semantic subjects. We hope it will support future work on more robust evaluation and on active, tool-augmented physical reasoning.

ETHICS STATEMENT

HumanPhys is built entirely by secondary use of publicly released research datasets (JRDB, DanceTrack, UCSD Ped, VideoAttentionTarget, EyeBlink8, DECAF) and of synthetic assets (FLAME/MANO/SMPL-X parametric bodies, BEDLAM identities, ReplicaCAD scenes); we collect no new human-subjects data, so the release itself required no IRB review. Ten colleagues, all adult volunteers known to the authors, additionally answered a partial subset of the released benchmark questions on a voluntary, unpaid basis to produce the indicative *Human* row of Table 3; only anonymised per-question answers and response times were recorded, and no identity-linked or sensitive attribute was collected, stored or released. Because each participant's subset was availability-selected rather than uniform, we treat that row as context only and draw no statistical conclusion from it. We inherit and enforce the upstream licence terms: the release comprises only the trimmed video segments each question consumes, together with derived question–answer pairs, calibration metadata and renderer-generated sequences, all under non-commercial research terms that preserve upstream attribution; for sources whose terms do not clearly permit re-hosting we ship fetch scripts pointing to the upstream archives instead, plus a per-segment provenance manifest and an expedited takedown procedure that honours any upstream, creator or subject request. Concretely, JRDB clips are shared under CC BY-NC-SA 3.0 and DanceTrack annotations under CC BY 4.0; the minimal trimmed segments of DanceTrack, UCSD Ped, VideoAttentionTarget and EyeBlink8 are hosted strictly for non-commercial research — the legally most delicate part of the release, since DanceTrack's footage is internet-sourced and its creators disclaim ownership while the other three publish no redistributable terms — with upstream attribution, spans limited to what each question requires, and the takedown procedure above. All rendered Group-II material inherits the MPI non-commercial research licence of its parametric bodies and scenes, including the explicit MPI prohibition of surveillance use, which we forward to every downstream user, and ships with regeneration scripts so that asset bytes never change hands. Our labels describe geometry and transient physiological cues (depth, distance, image-plane gaze angle, blink count) rather than identity or sensitive attributes: the release contains no identity labels and we make no attempt to re-identify any individual. We acknowledge a dual-use risk — fine-grained physical reading of people could in principle support surveillance-adjacent applications — and mitigate it by non-commercial licensing, by withholding identity-linked attributes, and by restricting every task to short, context-free clips without narrative or behavioural profiling. All authors have read the ICLR Code of Ethics and adhere to it.

REPRODUCIBILITY STATEMENT

Evaluation runs on the open-source lmms-eval framework with prompt templates for each sub-task and the deterministic, judgement-free scorers of §4.2; generation is capped at 1,024 tokens with otherwise default sampling. As in §3.1, every question carries its ordered sampled frames (*frame*: one image; *video*: the clip's full sampled frame list), so no frame-rate resampling and no per-video frame cap is ever applied to a HumanPhys question, identically in training and evaluation; the only videos in our pipeline belong to the external SpaceR-151K training mixture, where each video is uniformly sub-sampled under a per-video cap of 16 frames and a shorter clip may yield fewer frames. The SFT checkpoints update all parameters of Qwen3-VL-4B-Instruct with learning rate 1×10^{-5}, a cosine schedule with 5% warm-up and global batch size 1024 for one epoch on 256 Ascend HPU-910C accelerators, mixing the indicated corpora in full (§4.3). Construction is fully specified in §3: source datasets, filtering thresholds, walk grammars, scorer definitions and the stratified evaluation split. Code, prompt templates, scorers and the evaluation split will be released upon acceptance to allow exact replication.

AI USE STATEMENT

In this work, we used generative AI tools only for tasks with recommended disclosure: editing portions of the manuscript for readability and scripting auxiliary data-processing code. Every AI-assisted passage and script was reviewed, verified and approved by the authors. We have not used generative AI tools for any task with required disclosure: the HumanPhys corpus is generated programmatically from source-dataset metadata and simulator/renderer state rather than by an LLM,

and no mathematical claims, hypotheses, experimental designs or result interpretations were produced by generative models. The large vision–language models evaluated in our experiments (e.g. GPT-5.6-Sol, Gemini-3.5-Flash, Qwen3-VL) serve solely as study subjects, not as assistants in the research workflow. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

[1] Alon Albalak, Duy Phung, Nathan Lile, Rafael Rafailov, Kanishk Gandhi, Louis Castricato, Anikait Singh, Chase Blagden, Violet Xiang, Dakota Mahan, and Nick Haber. Big-Math: A large-scale, high-quality math dataset for reinforcement learning in language models. *arXiv preprint arXiv:2502.17387*, 2025.

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[4] Hunar Batra, Haoqin Tu, Hardy Chen, Yuanze Lin, Cihang Xie, and Ronald Clark. SpatialThinker: Reinforcing scene graph-grounded spatial reasoning via dense rewards. *arXiv preprint arXiv:2511.07403*, 2025.

[5] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. arXiv:2306.16940.

[6] Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. SpatialBot: Precise spatial understanding with vision language models. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025. doi: 10.1109/ICRA55743.2025.1112867. arXiv:2406.13642.

[7] Yuxuan Cai, Jiangning Zhang, Zhenye Gan, Qingdong He, Xiaobin Hu, Junwei Zhu, Yabiao Wang, Chengjie Wang, Zhucun Xue, Chaoyou Fu, Xinwei He, and Xiang Bai. HumanVideo-MME: Benchmarking MLLMs for human-centric video understanding. *arXiv preprint arXiv:2507.04909*, 2025.

[8] Antoni B. Chan, Zeng-Sheng Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[9] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, 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)*, 2024. doi: 10.1109/CVPR52733.2024.01370. arXiv:2401.12168.

[10] 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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.01584.

[11] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M. Rehg. Detecting attended visual targets in video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. arXiv:2003.02501.

[12] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2501.17161.

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

[14] Yalun Dai, Hao Li, Shulin Tian, Runmao Yao, Yuhao Dong, Fangzhou Hong, Zhaoxi Chen, Fangfu Liu, Tao Wang, Kim-Hui Yap, and Ziwei Liu. S-Agent: Spatial tool-use elicits reasoning for spatial intelligence. *arXiv preprint arXiv:2606.20515*, 2026.

[15] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaosong Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[16] Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. EmbSpatial-Bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2024. doi: 10.18653/v1/2024.acl-short.33. arXiv:2406.05756.

[17] Mahsa Ehsanpour, Fatemeh Saleh, Silvio Savarese, Ian Reid, and Hamid Rezatofighi. JRDB-Act: A large-scale dataset for spatio-temporal action, social group and activity detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[18] Gueter Josmy Faure, Min-Hung Chen, Jia-Fong Yeh, Hung-Ting Su, and Winston H. Hsu. FineBench: Benchmarking and enhancing vision-language models for fine-grained human activity understanding. *arXiv preprint arXiv:2605.19846*, 2026.

[19] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision (ECCV)*, 2024. doi: 10.1007/978-3-031-73337-6_9. arXiv:2404.12390.

[20] Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. Omni-MATH: A universal olympiad level mathematic benchmark for large language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.07985.

[21] Qiyue Gao, Xinyu Pi, Kevin Liu, Junrong Chen, Ruolan Yang, Xinqi Huang, Xinyu Fang, Lu Sun, Gautham Kishore, Bo Ai, Stone Tao, Mengyang Liu, Jiaxi Yang, Chao-Jung Lai, Chuanyang Jin, Jiannan Xiang, Benhao Huang, Zeming Chen, David Danks, Hao Su, Tianmin Shu, Ziqiao Ma, Lianhui Qin, and Zhiting Hu. Do vision-language models have internal world models? towards an atomic evaluation. *arXiv preprint arXiv:2506.21876*, 2025.

[22] Pulkit Gera, Faegheh Sardari, Asmar Nadeem, Valentina Bono, Padraig Boulton, Adrian Hilton, and Armin Mustafa. HumanMoveVQA: Can video MLLMs reason about human movement in videos? *arXiv preprint arXiv:2606.27999*, 2026.

[23] GLM-V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, et al. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

[24] GLM-V Team, Wenyi Hong, Xiaotao Gu, Ziyang Pan, Zhen Yang, Yuting Wang, Yue Wang, et al. GLM-5V-Turbo: Toward a native foundation model for multimodal agents. *arXiv preprint arXiv:2604.26752*, 2026.

[25] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-VL technical report. *arXiv preprint arXiv:2504.07491*, 2025.

[26] Dinging Li, Yingxiu Zhao, Xinrui Cheng, Kangheng Lin, Hongbo Peng, Hongxing Li, Zixuan Wang, Yuhong Dai, et al. SpatialEvo: Self-evolving spatial intelligence via deterministic geometric environments. *arXiv preprint arXiv:2604.14144*, 2026.

[27] Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. ViewSpatial-Bench: Evaluating multi-perspective spatial localization in vision-language models. In *European Conference on Computer Vision (ECCV)*, 2026. doi: 10.1007/978-3-032-37098-3_6. arXiv:2505.21500.

[28] Kun Li, Jihao Gu, Fei Wang, Zhiliang Wu, Hehe Fan, and Dan Guo. MA-Bench: Towards fine-grained micro-action understanding. *arXiv preprint arXiv:2603.26586*, 2026.

[29] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics (TOG)*, 2017.

[30] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. VideoChat-R1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

[31] Yifei Li, Pengyiang Liu, Yuhang Zang, Zhongyue Shi, Qi Fu, Hongye Hao, and Jiwen Lu. OVO-S-Bench: A hierarchical benchmark for streaming spatial intelligence in multimodal LLMs. *arXiv preprint arXiv:2606.03890*, 2026.

[32] Yun Li, Yiming Zhang, Tao Lin, Xiangrui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. STI-Bench: Are MLLMs ready for precise spatial-temporal world understanding? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. doi: 10.1109/ICCV51701.2025.00533. arXiv:2503.23765.

[33] Lucy Lin, Ayush Jain, Yifan Liu, and Katerina Fragkiadaki. Qwen-3D: A generalist 3D vision-language model for spatial understanding. *arXiv preprint arXiv:2608.02980*, 2026.

[34] Michelle Lin. StateSight: Benchmarking latent spatial-state reconstruction in vision-language models. *arXiv preprint arXiv:2608.20414*, 2026.

[35] Jianhui Liu, Haoze Sun, Wenbo Li, Yanbing Zhang, Rui Yang, Zhiliang Zhu, Yijun Yang, Shenghe Zheng, Nan Jiang, et al. OpenSpatial: A principled data engine for empowering spatial intelligence. *arXiv preprint arXiv:2604.07296*, 2026.

[36] Ming Liu, Yunbei Zhang, Shilong Liu, Liwen Wang, and Wensheng Zhang. Wan-R1: Verifiable-reinforcement learning for video reasoning. *arXiv preprint arXiv:2603.27866*, 2026.

[37] Yuansen Liu, Haiming Tang, Jinlong Peng, Jiangning Zhang, Xiaozhong Ji, Qingdong He, Wenbin Wu, Donghao Luo, Zhenye Gan, Junwei Zhu, et al. Human-MME: A holistic evaluation benchmark for human-centric multimodal large language models. *arXiv preprint arXiv:2509.26165*, 2025.

[38] Yuhong Liu, Beichen Zhang, Yuhang Zang, Yuhang Cao, Long Xing, Xiaoyi Dong, Haodong Duan, Dahua Lin, and Jiaqi Wang. Spatial-SSRL: Enhancing spatial understanding via self-supervised reinforcement learning. *arXiv preprint arXiv:2510.27606*, 2025.

[39] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-RFT: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. doi: 10.1109/ICCV51701.2025.00197. arXiv:2503.01785.

[40] Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso M. de Melo, and Alan Yuille. 3DSRBench: A comprehensive 3D spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. doi: 10.1109/ICCV51701.2025.00651. arXiv:2412.07825.

[41] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. MM-Eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

[42] Hai Nguyen, Tung Vu, and Cong Tran. SpatialQuery: Benchmarking geometry-grounded multi-instance spatial reasoning in vision-language models. *arXiv preprint arXiv:2608.01709*, 2026.

[43] Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. SpaceR: Reinforcing MLLMs in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.

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

[45] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 2017.

[46] Manolis Savva et al. Habitat: A platform for embodied AI research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[47] Soshi Shimada, Vladislav Golyanik, Patrick Pérez, and Christian Theobalt. DECAF: Monocular deformation capture for face and hand interactions. *ACM Transactions on Graphics (TOG)*, 42(6), 2023. arXiv:2309.16670.

[48] Tereza Soukupová and Jan Čech. Real-time eye blink detection using facial landmarks. In *Computer Vision Winter Workshop (CVWW)*, 2016.

[49] Peixi Sun, Jinkun Cao, Yong Jiang, Zanjia Yuan, Song Bai, Kris Kitani, and Ping An. DanceTrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[50] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their Habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2106.14405.

[51] Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. OpenMathInstruct-2: Accelerating AI for math with massive open-source instruction data. *arXiv preprint arXiv:2410.01560*, 2024.

[52] Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, et al. SpatialVID: A large-scale video dataset with spatial annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2509.09676.

[53] Wei Wang, Yuqian Yuan, Tianwei Lin, Wenqiao Zhang, Siliang Tang, Jun Xiao, and Yueting Zhuang. CrossView suite: Harnessing cross-view spatial intelligence of MLLMs with dataset, model and benchmark. *arXiv preprint arXiv:2605.18621*, 2026.

[54] Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M. de Melo, Jieneng Chen, and Alan Yuille. Spatial457: A diagnostic benchmark for 6D spatial reasoning of large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. doi: 10.1109/CVPR52734.2025.02297. arXiv:2502.08636.

[55] 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. arXiv:2505.23747.

[56] Haoning Wu, Xiao Huang, Yaohui Chen, Ya Zhang, Yanfeng Wang, and Weidi Xie. SpatialScore: Towards comprehensive evaluation for spatial intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2505.17012.

[57] Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.09965.

[58] Jiahao Xie, Alessio Tonioni, Nathalie Rauschmayr, Federico Tombari, and Bernt Schiele. SSL-R1: Self-supervised visual reinforcement post-training for multimodal large language models. *arXiv preprint arXiv:2604.20705*, 2026.

[59] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. doi: 10.1109/CVPR52734.2025.00994. arXiv:2412.14171.

[60] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. MMSI-Bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025.

[61] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. R1-Onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. doi: 10.1109/ICCV51701.2025.00229. arXiv:2503.10615.

[62] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 2024. arXiv:2306.13549.

[63] Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. From flatland to space: Teaching vision-language models to perceive and reason in 3D. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2503.22976.

[64] Jiyao Zhang, Mingxu Zhang, Yitong Peng, Haoxuan Liu, Chenshuo Wang, Yuxing Long, Haoyang Huang, Dongjiang Li, et al. Embodied3DBench: Benchmarking low-level embodied spatial intelligence of vision language models. *arXiv preprint arXiv:2605.29074*, 2026.

[65] Ting Zhou, Daoyuan Chen, Qirui Jiao, Bolin Ding, Yaliang Li, and Ying Shen. HumanVBench: Probing human-centric video understanding in MLLMs with automatically synthesized benchmarks. *arXiv preprint arXiv:2412.17574*, 2024.

[66] Xiaolong Zhou, Yifei Liu, Ziyang Gong, Jiarui Li, Qiyue Zhao, Muyao Niu, Yuanyuan Gao, Le Ma, et al. SpaceDG: Benchmarking spatial intelligence under visual degradation. *arXiv preprint arXiv:2605.22536*, 2026.

[67] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A CONSTRUCTION DETAILS FOR EACH SUB-TASK

This appendix provides the full construction details summarized in §3.1: the filtering thresholds, coordinate conventions, walk grammars and worked examples behind each of the thirteen sub-tasks.

Group I: real-world measured human-centered physical cues.

- **Spatial state** (*Depth*, *Distance*, *Count*, *Gaze*). *Depth* and *Distance* are derived from JRDB [17], whose five cameras ride a moving robot: point-cloud registration between consecutive samples flags every range whose per-step camera displacement exceeds 0.20 m for two consecutive samples (or a whole sequence above 1.5 m cumulative), and questions are constructed only from the remaining static-camera frames. *Count* is read from the boxes of DanceTrack [49] and UCSD [8], and *Gaze* from the head boxes and fixation points of VideoAttentionTarget [11], as listed in Table 2. *Depth* is read directly from the calibration metadata, namely the distance to the reference sensor. *Distance* is the pairwise inter-person distance measured at the *foot* point $(c_x, c_y, c_z - \frac{h}{2})$ which reduces the effect of variation in human height. *Count* is the number of humans satisfying an explicit spatial criterion (everybody dancing in frame, or inside a labelled walkway ROI), explicitly stated in the question. *Gaze* is the clockwise angle $\mathrm{atan2}(\Delta y, \Delta x) + 90^\circ$ between the upward image direction and the head-centre-to-fixation-point vector; because this angle lives in the image plane it is camera-dependent by construction, we treat it as a viewpoint-conditioned physical cue, not an invariant head direction. *Ex: "What is the minimum distance between pedestrian 29 and pedestrian 46 in the video?"*
- **Kinematics** (*Displacement*, *Velocity*, *Trajectory*). *Displacement* is the cumulative ground-plane path length $\sum_i \sqrt{\Delta x_i^2 + \Delta y_i^2}$ between the first and last frame of the clip; *Velocity* divides that path length by the clip's elapsed time, with the sampling interval stated in the question so the model must bind visual motion to a stated time base; and *Trajectory* is the sequence of heading changes and step lengths as tuples $(\text{turn_deg}, \text{distance_m})$ with the first step anchored at 0° and positive turn meaning left. Curved paths are kept by a random forest that triages every candidate trajectory into straight/curved/abandon; 400 human labels (200 uniformly sampled, 200 mined from its highest-entropy predictions) retrain the classifier, and its outputs decide which trajectories become questions; the classifier's residual error is therefore a component of the *Trajectory* gold noise.
- **Event** (*Blink*). The number of discrete behaviour events (eye blinks) in a clip, requiring the model to integrate evidence across frames instead of reading a single canonical pose. Its source, EyeBlink8 [48], runs at 30 Hz, so a raw 32-frame window would almost never contain more than one blink; we therefore down-sample it, compressing each open-eye run to at most two evenly spaced frames and collapsing each closed-eye run to its middle frame, so that 32 frames span a far longer period and the gold counts spread over a wider range — while a count never depends on how long the eyes happened to stay shut. *Ex: "How many times did the person blink in this video?"*

Group II: controlled physically rendered probes.

- **Deformation** (*Deform*). A FLAME [29] head deformed by MANO [45] hand contact, captured by DECAF [47], is rendered as a 2×4 panel grid (two head-following views $\times$ four deformation strengths). Each option's strength α rescales the measured contact deformation by a Gaussian influence field $w_i = \exp(-d_i^2/2\sigma^2)$, $\mathbf{V}_i' = \mathbf{V}_i^{\text{und}} + (\alpha - 1) w_i \mathbf{D}_i$; exactly one option ($\alpha = 1.0$) matches physical reality, while the three distractor options sample their strength from $\alpha \in [0, 0.5] \cup [2, 3]$ under a deterministic per-(subject, window) seed, so every option stays physically legal yet separable from the true contact. DECAF's monocular fit is therefore the anchor of the α=1 option, so its fit error is a component of the *Deform* gold noise, exactly as the triage classifier's error is for *Trajectory*. *Ex: "Which face's deformation best matches physical reality?"*
- **Illumination** (*LS-Head*, *LS-Body*). A standing human (200 BEDLAM [5] identities) inside a light stage of 27 point lights on a $3 \times 3 \times 3$ lattice; exactly one light is on per frame and its coordinate follows a lattice walk of 3–6 frames that the model must infer from the shadow motion on the face and body. The light sources themselves are never visible in any rendered frame, so the walk can only be recovered from shading and cast-shadow motion. The two variants probe complementary cue regimes and walk grammars rather than two points on one difficulty scale.

LS-Head uses the frontal self-portrait camera, where the shading change on the face is subtle, and restricts each walk step to a single lattice axis ($\pm X$, $\pm Y$ or $\pm Z$ by one step), keeping the facial cue readable; *LS-Body* uses the 8 peripheral wall cameras, where body shading and cast shadows are far more informative, and runs the *fast* walk, whose steps may change up to three lattice axes at once (step lengths 1, $\sqrt{2}$ or $\sqrt{3}$), so its legal step set is larger and its per-axis shading signals are mixed. Because cue strength and step grammar change together, scores are comparable within a variant but not across variants; we therefore treat the two variants as separate sub-tasks throughout and caution against cross-variant comparison.

- **Camera pose** (*CP-Room*, *CP-Human*). A free-flying camera performs a random walk on a collision-checked cubic lattice (0.20 m step; lattice nodes that intersect scene geometry are discarded beforehand) through 84 baked indoor stages; every frame transition is exactly one *atomic action*: a translation along $\pm X/\pm Y/\pm Z$ by 1–3 steps, or an in-place yaw/pitch/roll rotation of $\pm 1^\circ$–$\pm 30^\circ$ (accumulated pitch capped at $+60^\circ$). Frame 1 shows a 3D axis gizmo plus the stated camera pose, and the question asks for the action sequence of all later frames. *CP-Human* plants a SMPL-X [44] human and samples paths that keep the face framed.

Table 6: **Human study per sub-task.** From top: uniform random baseline over the full bank; text-only baseline; mean human answering time per question in seconds, with cell shading proportional to time (darker = slower); and human scores on the same availability-selected, non-uniform ≈40% subset as the *Human* row of Table 3, reported as an indicative reference only.

	Depth	Dist	Count	Gaze	Disp	Vel	Traj	Blink	Deform	LS-Head	LS-Body	CP-Room	CP-Human
Random	7.2	8.3	23.4	28.2	5.8	12.2	37.8	22.0	26.2	13.4	48.1	6.1	2.8
Text Only	20.60	24.00	0.00	32.55	0.00	14.80	54.03	0.00	28.00	24.58	54.19	4.62	2.32
Time (s)	40.4	25.2	96.5	35.4	44.9	180.3	131.5	82.3	136.2	1141.8	533.7	303.4	823.8
Human	25.2	52.0	79.2	60.4	36.5	30.3	63.5	87.4	36.1	50.6	70.1	24.9	12.2

B Comparison with the Closest Benchmarks

This appendix makes the two comparisons promised in §2.2 explicit, so that our claims about prior work can be checked axis by axis rather than asserted. Table 7 contrasts HumanPhys with HumanMoveVQA [22], the only existing human-centric benchmark that asks geometric questions, and with SpatialScore [56], the broadest existing consolidation of spatial evaluation together with its SpatialCorpus training resource. Table 8 decomposes the first comparison by task category and states precisely how much of HumanPhys overlaps with it.

Two conclusions follow. First, the overlap with HumanMoveVQA is confined to three of our thirteen sub-tasks and even there the answer format is strictly harder: an open-ended metric sequence scored step by step, rather than an ordinal choice among four options derived from the same estimated tracks. Second, our differentiation from SpatialScore is about the referent and not about breadth: SpatialScore already spans 30 task categories, releases a 331K-sample corpus and fine-tunes on it, so we claim neither the paradigm nor the breadth — we claim that none of those categories measures a property of a human body, that human-physical perception is not obtained from that supervision (the four spatially tuned checkpoints in Table 3 reach 20.43–27.75), and that reading the profile per measurement target yields findings an aggregate score cannot: dissociable dimensions, a universal camera-pose failure mode, and a human-presence penalty on ego-motion recovery.

Table 7: **HumanPhys vs. the two closest benchmarks.** HumanMoveVQA is the only human-centric benchmark that asks geometric questions; SpatialScore is the broadest general spatial evaluation with an accompanying training corpus. Entries are taken from each work's published description; "—" means the axis does not apply.

	HumanPhys (ours)	**HumanMoveVQA**	**SpatialScore**
Measurement target	the human body as rigid/non-rigid target, photometric probe and observer reference	root translation and orientation of one target person	object and scene geometry, including the camera
Task inventory	13 sub-tasks in 6 dimensions	7 multiple-choice reasoning categories	30 tasks in about 10 capability groups
Answer units	m, m/s, °, event counts, lattice coords, `(turn_deg, distance_m)` tuples	ordinal bins (short/moderate/long, slow/fast)	mostly categorical, partly numeric
Format and scoring	open-ended; MRA (10 thresholds) + law-of-cosines sequence score; MC only for *Deform*	all multiple choice; chance-normalised accuracy	judgement, MC and open-ended accuracy
Gold provenance	sensor calibration, 3D-box metadata, renderer state; no estimator/annotator *on the benchmark frames* (*Trajectory* curation uses a random forest retrained on 400 labels; *Deform* anchors α=1 on DECAF's monocular fit)	estimated SMPL-X tracks, discretised into codes, manual QC	repurposed 3D annotations, manually verified
Coordinate convention	metric world frame; static-camera filter; foot-point distance; heading anchored at 0°, left positive	first-frame anchored, normalised to initial body-root state	per-task source conventions
Viewpoint and people	exocentric + face-level; multi-person (*Count*, *Distance*); probes with/without person	exocentric; one person per clip	mostly scene images, multi-view sets and videos
Visual source	5 real video datasets and 3 rendered collections	3 real video datasets (EMDB, RICH, EgoBody)	real, AIGC and 3D-annotation repurposing
Evaluation scale	2,500 evaluation questions, equal quota per sub-task (question-level split)	10,203 test questions on 97 videos	about 5K manually verified samples
Training resource	385,702 QA pairs, same pipeline, no question shared with eval	authors' own training split	SpatialCorpus (331K); SpatialAgent (training-free)
Post-training protocol	SFT of Qwen3-VL-4B, four corpus mixtures	SFT of Qwen3-VL-8B on its own data	SFT of Qwen3-VL on SpatialCorpus
Generalization analysis	transfer to SPAR/MMSI sub-tasks; specialization trade-off on 9 external benchmarks; mixture repair	cross-dataset/category transfer within its own family	gains on own benchmark after SFT/agent
Human-free conditions	*CP-Room* as matched control only	none; every clip has the target person	all conditions

Table 8: **Category-by-category overlap with HumanMoveVQA.** Ten of our thirteen sub-tasks have no counterpart there, and the three that do ask for open-ended metric sequences rather than ordinal multiple choice.

Their category	Nearest HumanPhys sub-task	What each asks for
Existence	(none)	they assert or deny one motion event (2-way MC); we count events (*Blink*) or ask for the magnitude itself
Comparative	(none)	they compare magnitudes along one axis; we ask for the metric value and score it by MRA
Dominant	*Trajectory* (weakly)	they ask the primary turning direction; we ask for the whole ordered turn/step sequence
Numerical	*Count*, *Displacement*	they count turning events and pick a total displacement from unitless bins; we count humans under an explicit metric criterion and ask displacement in metres
Ordering	*Trajectory* (order)	they identify the event following an anchor; we score every step of the tuple sequence by accumulated heading
Temporal	*Velocity*	they pick the earliest event or a slow/fast category; we ask m/s against a stated sampling interval
Trajectory affordance	*Displacement*, *Trajectory*	they pick net displacement or path shape from 4 options; we score per-step `(turn_deg, distance_m)` tuples continuously