This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

QUANTIPHY: A Quantitative Benchmark Evaluating Physical Reasoning Abilities of Vision-Language Models

Puyin Li[1*] Tiange Xiang[1*] Ella Mao[1*] Shirley Wei[1] Xinye Chen[1] Adnan Masood[2]
Li Fei-Fei[1†] Ehsan Adeli[1†]
[1]Stanford University [2]UST



Figure 1. On a crowded city street, a bird's nest falls from a branch, a car rushes by, an eagle soars past a building, and a person walks in a crosswalk — the real world is full of complex physical motion. To enable AI to understand and navigate this environment, it is essential for generalist embodied systems to reason about physical properties quantitatively. Because objects obey common laws of physics, their kinematic properties (such as size, velocity, and acceleration) are interrelated. This interdependence makes it possible for visual AI to systematically reason about these properties with respect to available priors. In this work, we present QUANTIPHY, the first benchmark to evaluate the reasoning ability of AI models on quantitative kinematic inference tasks.

Abstract

Understanding the physical world is essential for generalist AI agents. However, it remains unclear whether state-of-the-art vision perception models (e.g., large VLMs) can reason physical properties quantitatively. Existing evaluations are predominantly VQA-based and qualitative, offering limited insight into whether these models can infer the kinematic quantities of moving objects from video observations. To address this, we present QuantiPhy, the first benchmark designed to quantitatively measure a VLM's physical reasoning ability. Comprising more than 3.3K video–text instances with numerical ground truth, QuantiPhy evaluates a VLM's performance on estimating an object's size, velocity, and acceleration at a given timestamp, using one of these properties as an input prior. The benchmark standardizes prompts and scoring to assess numerical accuracy, enabling fair comparisons across models. Our experiments on state-of-the-art VLMs reveal a consistent gap between their qualitative plausibility and actual numerical correctness. We further provide an in-depth analysis of key factors like background noise, counterfactual priors, and strategic prompting and find that state-of-the-art VLMs lean heavily on pre-trained world knowledge rather than faithfully using the provided visual and textual inputs as references when reasoning kinematic properties quantitatively. QuantiPhy offers the first rigorous, scalable testbed to move VLMs beyond mere verbal plausibility toward a numerically grounded physical understanding.

1. Introduction

Understanding the physical world has long been a challenge for artificial intelligence. Humans inhabit a world governed by physical laws. From an apple falling from a tree to the trajectory of a thrown ball, we have developed mathematical tools to measure and calculate physical attributes quantitatively. This quantitative understanding forms the foundation for all modern scientific advances.

Visual intelligence systems, such as Vision-Language Models (VLMs), are developed in a manner that differs significantly from that of humans. They are trained to fit vast amounts of real-world data, which implicitly contain the physical principles behind visual observations. Thus, assessing the numerical accuracy of VLMs' reasoning about physical properties is a necessary next step. It is crucial for deploying applications such as embodied AI [16, 43], AR/VR [21, 29, 48–50], and autonomous driving [41].

Reasoning about quantitative physical properties from visual data can be challenging. For example, advanced methods like FoundationPose [45] require extensive prior information, including color, depth, object meshes, and camera parameters, to accurately localize an object in 3D

*Equal contribution.
†Equal mentorship.

space. Unfortunately, most of these priors are unavailable in "in-the-wild" captures. These challenges make it natural to ask whether large VLMs can leverage their rich implicit priors to reason end-to-end about precise kinematic and geometric properties. Moreover, a robust evaluation of VLM reasoning has the potential to help assess and improve the physical realism of videos created by generative models.

Studies of VLM physical understanding are not new. Existing benchmarks cover kinematics [57], dynamics [52], relationships [14], and scene understanding [54]. However, most are *VQA-based and qualitative*: ground-truth answers are constrained by the prompt, and models are evaluated with multiple-choice questions. This setup does not provide fine-grained evaluation of physical understanding. For example, if a model is asked to infer the size of a car from a video (ground truth: 3 meters), answers of 3.1 meters and 31 meters are both marked wrong in multiple-choice evaluation, even though the latter is quantitatively 10× worse. Capturing this numerical gap is crucial for advancing VLMs toward real-world use.

Our contributions are four-fold: (I) We propose a quantitative paradigm for evaluating VLM physical reasoning, moving beyond qualitative VQA. (II) We define a kinematic inference task for videos, where correlated quantities such as size, velocity, and acceleration are inferred from a single physical prior in real-world units. (III) We present QUANTIPHY, the first benchmark for systematically evaluating VLMs' quantitative reasoning about object kinematics in videos, with standardized metrics, prompting, and a leaderboard over state-of-the-art models. (IV) We provide analysis of factors affecting VLM reasoning, including scene complexity, video availability, counterfactual priors, and chain-of-thought prompting. Overall, we find that existing VLMs often rely on pre-trained world knowledge rather than truly **inferring from** the reference video and text.

2. Related Work

2.1. Benchmarks for VLM Physical Understanding

For over a decade, researchers have recognized the need for AI to comprehend real-world physics, spurring a series of benchmarks assessing the qualitative physical reasoning abilities of VLMs. Early efforts primarily focused on evaluating a model's capacity to describe, predict, and explain basic physical events, such as collisions, falling, and rebounding in controlled synthetic environments [8, 57]. Subsequent works expanded this scope to reasoning about inherent physical properties of objects such as mass, friction, elasticity, and deformability [42]. Recently, comprehensive benchmarks like PhysBench [14] and STAR [46] extended coverage to object relationships, physical scene understanding, complex physics-based dynamics, and physics-situated actions in more realistic or diverse environments [22, 60], aiming to simulate the real-world reasoning demands that embodied AI systems will face.

Despite these advances, the vast majority of existing benchmarks adopt the VQA framework [5, 40, 51, 56], asking models for multiple-choice selections or descriptive explanations of physical events. One relevant work, VSI-Bench [54], provides preliminary studies on basic spatial understanding using numerical metrics, but is limited to static objects, focusing more on the models' perceptual ability than on emergent reasoning. Super-VSI [55] subsequently demonstrated that equipping VLMs with numerical spatial understanding yields strong potential for embodied AI. Therefore, a need still exists for benchmarking VLM's ability to perform quantitative reasoning over the geometric and kinematic properties of moving objects (e.g., estimating size, velocity, or acceleration in metric terms).

2.2. Physical Reasoning Models

Although AI models have achieved impressive results on general tasks, their ability to reason about the physical world remains limited. For VLMs, state-of-the-art models such as GPT-4o and Gemini-1.5 Pro achieve only around 60% accuracy on the PhysBench benchmark, far below human-level performance (∼95%) [14]. Generative models also struggle. Recent work shows that generated videos often violate basic physical laws [9], as these models are more sensitive to low-level visual cues (e.g., color, shape) than to underlying physical properties [25]. In embodied AI, recent systems have demonstrated basic physical reasoning abilities. They can interact with the environment [6], but their reasoning remains largely qualitative, lacking a precise and quantitative understanding of physical dynamics. These gaps underscore the need for benchmarks and models that move beyond qualitative judgment toward accurate quantitative physical reasoning.

2.3. Physical Understanding and Reasoning

In computer vision, techniques such as optical flow [11, 24], combined with object detection and tracking models like YOLO [34] and ByteTrack [59], enable accurate analysis of object motion across broad applications [18, 37]. These methods demonstrate that quantitative information (e.g., velocity, displacement) can be reliably inferred from video inputs. In cognitive science, research shows that humans possess strong abilities to reason about scale, from microscopic to astronomical, by leveraging relational cues and prior knowledge [36]. Even in visually unfamiliar or simulated environments, people can extract physical rules and perform causal and predictive reasoning about object motion and interactions [3, 17, 26]. This highlights the potential for robust physical reasoning ability to emerge in large AI models from visual inputs, a crucial step toward grounded quantitative understanding.

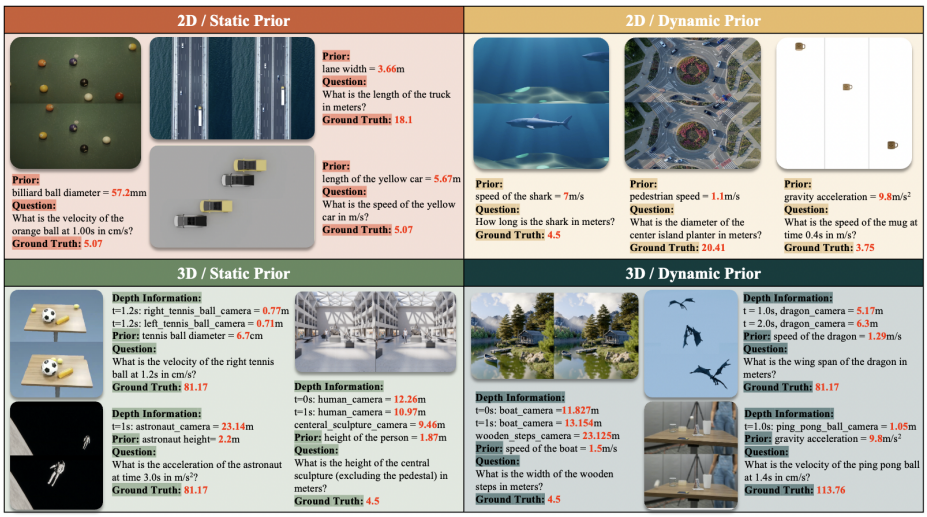


Figure 2. Sample examples from QUANTIPHY, illustrating the four core task combinations defined by `Dimensionality` {`2D, 3D`} and `Physical Prior` {`Static, Dynamic`}, as described in Section 3.1. Our collected data is diverse in nature, and each video is paired with multiple (prior, question, ground truth) triplets. Please see the supplementary materials for more data examples.

3. Methods

We introduce **QUANTIPHY**, the first benchmark designed to *quantitatively* evaluate VLMs' physical reasoning ability on moving objects. Section 3.1 first defines the basic object kinematic properties assessed in QUANTIPHY and presents key dataset statistics. Section 3.2 then details the benchmark construction and data collection process.

3.1. Overview of QUANTIPHY

The task of kinematic inference. The purpose of QUANTIPHY is to evaluate whether VLMs can utilize prior knowledge to reason about objects' kinematic properties with numerical accuracy, which forms the foundation for a more sophisticated understanding of physics. Specifically, we focus on the *translational movements* of various objects.[1]

> Given a video, we provide the VLM with a single physical prior for a source object (from the set $\{\mathbf{S}^{\text{world}}, \mathbf{V}_t^{\text{world}}, \mathbf{A}_t^{\text{world}}\}$, in real-world units) as textual input. The model is then prompted to *quantitatively determine* requested kinematic properties of a target object (which may be the same as or different from the source object) in world space.

We distinguish *pixel space*, where quantities are measured in pixels ([pixel], [pixel/s], [pixel/s^2]), from *world space*, where they are expressed in physical units (e.g., [m], [m/s], [m/s^2]). Consider a fixed-camera video capturing the translational motion of an object. At any time t, the object's location in pixel space, $\mathbf{X}_t^{\text{pixel}}$, can be obtained from the frames. From the discrete trajectory, we compute pixel-space velocity and acceleration via finite differences:

$$\mathbf{V}_t^{pixel} \approx \frac{\mathbf{X}_{t+\mathrm{d}t}^{\text{pixel}} - \mathbf{X}_t^{\text{pixel}}}{\mathrm{d}t}; \quad \mathbf{A}_t^{pixel} \approx \frac{\mathbf{X}_{t+2\mathrm{d}t}^{\text{pixel}} - 2\mathbf{X}_{t+\mathrm{d}t}^{\text{pixel}} + \mathbf{X}_t^{\text{pixel}}}{\mathrm{d}t^2}.$$

The video thus defines the kinematics only in pixel units. To relate them to world space, we assume an unknown scalar scale factor $\gamma > 0$ (in units [world length/pixel]) such that, along the motion direction, $\mathbf{S}^{\text{world}} = \gamma \mathbf{S}^{\text{pixel}}$, $\mathbf{V}_t^{\text{world}} = \gamma \mathbf{V}_t^{\text{pixel}}$, $\mathbf{A}_t^{\text{world}} = \gamma\, \mathbf{A}_t^{\text{pixel}}$, where $\mathbf{S}^{\text{pixel}}$ and $\mathbf{S}^{\text{world}}$ denote the same object size in pixel and physical units, respectively. When a single prior in world space is provided (object size $\mathbf{S}^{\text{world}}$, velocity $\mathbf{V}_t^{\text{world}}$, or acceleration $\mathbf{A}_t^{\text{world}}$ at some time t), together with the corresponding pixel-space quantity from the video, γ can be determined, and any other kinematic property in world space follows by rescaling its pixel-space counterpart.

See Figure 2 for examples of the kinematic tasks included in the benchmark. The performance of VLMs is measured by the numerical error between the VLM's prediction and the annotated ground truth.

Benchmark setup. For a comprehensive evaluation of the kinematic movements above, QUANTIPHY is designed to include video-question pairs categorized along three primary axes. We first describe the two axes that define the core reasoning task:

- `Dimensionality`: {`2D`, `3D`}. `2D` movement assumes the object moves strictly in the xy-plane with no depth change relative to the camera. `3D` motion includes movement along the z-axis, making the problem more challenging due to depth variation. In `3D` scenes, pixel displacements do not map uniquely to physical motion because of

[1] Rotational movements is deferred to future work; see section 6.

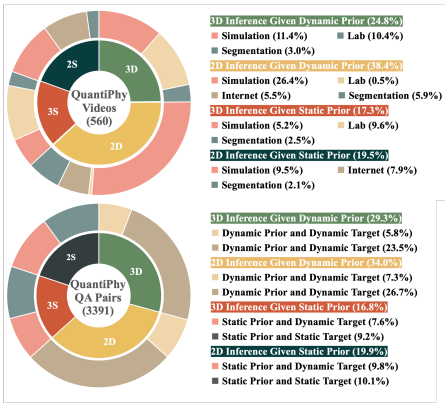


Figure 3. **QUANTIPHY Statistics.** The collected data and curated QA pairs are among four main setups with further breakdowns.

perspective projection, so depth information is needed to resolve this ambiguity. In our experiments, depth priors are provided in the same way as ground-truth priors in `2D` cases.

- `Physical prior`: {`Static`, `Dynamic`}. The `Static` prior indicates the provision of the object size $\mathbf{S}^{\text{world}}$, which is constant throughout the video. While `Dynamic` prior indicates velocity $\mathbf{V}_t^{\text{world}}$ or acceleration $\mathbf{A}_t^{\text{world}}$ at a given timestamp t.

Together, these two axes yield four task categories: `2D-Static`, `2D-Dynamic`, `3D-Static`, and `3D-Dynamic`.[2]

Data statistics. Following [14, 54], QUANTIPHY organizes data as (prompt, video, numerical ground truth) triplets, where each video may be paired with multiple questions. Across the four task categories, we collected 560 unique videos and 3391 questions from diverse sources. With proper post-processing, our collected videos typically have a duration of 2-3 seconds, totaling approximately 120MB of disk storage, making the benchmark suitable for use across various hardware settings. See Figure 3 for a breakdown.

3.2. Benchmark Construction

QUANTIPHY is constructed from a large-scale, diverse video dataset spanning multiple sources. The data collection process is visualized in Figure 4.

Data collection. Our data combines simulated and real-world sources for controllability and practical applicability.

- `Blender simulation`: Blender allows us to render scenes that are both visually realistic and physically plausible, encompassing `2D` and `3D` motion. Simulating object movements in Blender provides complete control over the environment and guarantees precise ground-truth annotations. This control allows us to probe VLM's reasoning capabilities by placing objects with different motion types within a single scene. It also enables us to study model robustness by holding motions fixed while systematically varying the background and visual noise. To enhance data diversity, we include simulations of scenes that are difficult or impossible to record in the real world, ranging from microscopic scales (e.g., red blood cells in vessels) to astronomical scales (e.g., galaxy movements).
- `Lab capturing`: Reconstructing objects in 4D (3D space + time) lets us annotate real-world object motion using a multi-view stereo setup. Our captures feature a wide diversity of movements, including free fall, sliding down slopes, pendulum motion, and bouncing. We also varied the physical properties of the objects, including those that are rollable and deformable. Further details on our setup can be found in the supplementary materials.
- `Internet scraping`: To extend our dataset to more out-of-distribution scenarios, we leverage the internet as a source for in-the-wild data. However, not all internet data is suitable for our benchmark, as our benchmark specifically requires consistent coverage of moving objects captured by a relatively static camera, along with reference objects of known dimensions (e.g., a standard coin) to serve as priors. Hence, we narrowed our scraping scope to websites hosting high-quality monocular videos and then manually inspected them to select those meeting our stringent requirements.

We provide a detailed description of the data collection process in the supplementary materials Appendix B.

Additional segmented data. We apply SAM [33] to segment moving objects against solid backgrounds, doubling our dataset without additional annotation while enabling controlled analysis of background complexity effects.

Data annotation and question design. Blender simulations provide direct extraction of sizes, velocities, and accelerations via automation scripts. We construct Blender scenes using open-source 3D assets from Sketchfab [38] and BlenderKit. Automation scripts were developed for each scene to extract accurate, time-specific physical properties for any given object.

However, data annotation is non-trivial for the other two sources. For lab captures, we utilize metric depth directly from depth cameras combined with multi-view stereo to achieve a full 4D reconstruction of the scene. We then apply per-pixel segmentation masks to outline the objects of interest in each video. For each video, we select a primary camera and use the object's metric depth from that viewpoint as the depth prior provided to the VLM. Object movements are then computed in world coordinates. For internet data, since the videos are monocular and lack multi-view information, we manually annotate sizes and displacements in pixel space and use reference objects with known priors to obtain the mapping to world scale.

We provide a detailed description of the data annotation process in the supplementary materials Appendix D.

Prompt design. As mentioned, the input to the VLM consists of a video paired with a single physical prior. For sim-

[2] In figures and tables, these are abbreviated as `2S`, `2D`, `3S`, and `3D`.

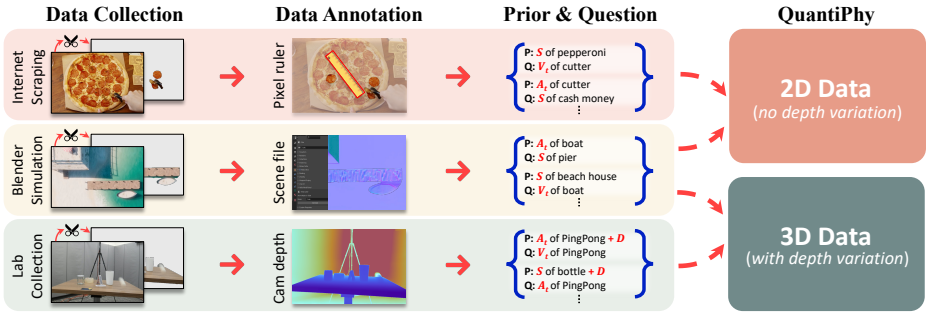


Figure 4. The construction of QUANTIPHY proceeds in three sequential stages. First, we collect diverse raw videos from three different sources. Additionally, we segment these videos with solid plain background (described in subsection 3.2). Second, we obtain high-quality annotations, employing distinct labeling methods tailored to each data source to accurately capture the object's physical properties. Finally, we formulate the benchmark tasks by associating each video with multiple (prior, question, ground truth) triplets. Each triplet is then categorized as either `2D` or `3D`, depending on the object's movement relative to the camera.

plicity and effective integration, this prior is provided as a textual description (see Figure 2 for examples). In addition, we include textual cues in the prompt, such as 'analyze the video and calculate the answer carefully'. We also add output constraints, instructing the model to 'output only the numerical answer and unit' in its final response.

We provide a detailed description of the prompt design process in the supplementary materials Appendix F.

4. Evaluation on QUANTIPHY

4.1. Evaluation Setup

Benchmark models. With different architecture designs, training data and protocals, different VLMs may have varying ability of physical understanding from videos. For better comprehensiveness, we include the evaluations of a total number of 21 state-of-the-art VLMs and variants in this work, consisting of 6 proprietary models include ChatGPT-5.1 [32], ChatGPT-5 [31], Gemini-2.5 Flash [19], Gemini-2.5 Pro [20], Grok-4.1 [47], and Claude-4.5 Sonnet [4] and 15 open-sourced models (as listed in Table 1). The code supports multiple providers (OpenAI, Gemini, xAI, Anthropic, and Replicate API) with provider-specific parameters. Temperature is typically 0–0.1 for deterministic outputs. Token limits vary: OpenAI models use up to 10,000 tokens due to longer thinking steps, and open-sourced models with around 500–2,048.

Human studies. To complement our model evaluation and establish a reference point for human-level performance on quantitative physical reasoning, we conducted a survey study. Participants watched 18 videos and answered 1–3 quantitative kinematic questions using the same priors and task definitions as our VLM evaluation.

Importantly, humans and VLMs perceive and process the inputs in very different ways. VLMs operate directly on pixel-accurate video tensors, while human participants rely on visual perception, intuition, and coarse approximations.

Metric design. All tasks in QUANTIPHY require models to output numerical values. Following VSI-Bench [54], we use *Mean Relative Accuracy* (MRA) as the primary metric to measure the proximity between model predictions and ground-truth answers. While accuracy based on exact matching is a simple baseline but too brittle for continuous, noisy physical measurements in QUANTIPHY, MRA evaluates whether the *relative error* falls below a set of tolerance thresholds and averages the resulting accuracies, offering a more calibrated and robust notion of when models are "accurate enough" for physically grounded reasoning (see subsection A.2 in the supplementary material for further discussion).

Concretely, we consider a set of confidence thresholds $\mathcal{C} = \{0.1, 0.2, \ldots, 0.9, 0.95\}$, and define MRA for a prediction $\hat{y}$ with ground truth y as

$$\text{MRA} = \frac{1}{10} \sum_{\theta \in \mathcal{C}} \mathbb{1}\left(\frac{|\hat{y} - y|}{|y|} < 1 - \theta \right),$$

where $\mathbb{1}(\cdot)$ is the indicator function. Intuitively, larger θ corresponds to a stricter tolerance $1 - \theta$ on the relative error, and MRA averages accuracy across a spectrum of such tolerances, providing a more informative measure than a single-threshold accuracy.

Our benchmark organizes questions into four kinematic categories, `2D-Static`, `2D-Dynamic`, `3D-Static`, and `3D-Dynamic`. For each model and each category, we compute the category-level score by averaging the question-level MRA over all questions in that category for which the model produces a *valid numerical answer*. The overall score of a model is then obtained as the unweighted mean of its four category-level scores.

A practical challenge is that not all VLMs consistently output sensible numerical predictions for our reasoning tasks. For every video–question pair, we query a model up

to five times with the same prompt, stopping early if any response contains a parseable numerical value. If none of the five responses yields a valid number, we regard the model as *failing to answer* that question.

4.2. Main Results

Table 1 summarizes performance on QUANTIPHY across four task types (2D-Static, 2D-Dynamic, 3D-Static, and 3D-Dynamic). Overall, we find that quantitative kinematic inference remains challenging for current VLMs: even the best systems do not yet reach human performance, despite in principle having access to more precise information.

Human baseline. Human annotators achieve an average MRA of 55.6 across all categories, with scores between 50.0 (2D-Static) and 59.1 (2D-Dynamic). This range is consistent with the fact that humans do not have direct access to pixel-level measurements and must instead rely on coarse visual estimation (e.g., counting grid lines or comparing to familiar objects), but still demonstrates that the tasks are solvable with reasonable accuracy.

Proprietary models. ChatGPT-5.1 attains the highest overall score with 53.1 MRA, followed by Gemini-2.5-Pro at 49.6. ChatGPT-5.1 slightly surpasses humans on the 2D-Dynamic category, yet none of them surpass the human average. Other closed models such as GPT-5 and Claude Sonnet 4.5 are substantially weaker, with overall MRA around 32.6 and 22.8, respectively.

Open-weight models. Open-weight models exhibit a wide performance spread. The best open-weight system, Qwen3-VL-Instruct-32B, reaches 46.0 overall MRA, with strong 2D-Dynamic and 3D-Dynamic scores, followed by InternVL-3.5-30B (40.7) and Qwen3-VL-Instruct-8B (38.8). These models are clearly below the top proprietary ones, but already comparable to mid-tier closed models. Smaller open-weight models such as Phi-4 Multimodal and SmolVLM-256M still achieve non-trivial MRA, yet remain far from both human and large-model performance.

Scaling effects. To better understand the role of model scale, we compare models within the same family across different parameter sizes. Within the Qwen3-VL family, average MRA increases from 29.0 at 2B parameters to 38.8 at 8B and further to 46.0 at 32B. A similar trend is observed for InternVL, where InternVL-3.5-30B (40.7) substantially outperforms its 8B (35.4) and 2B (25.0) variants. Notably, scaling benefits are most pronounced on dynamic categories (2D-Dynamic and 3D-Dynamic), suggesting that larger models are better able to integrate temporal information for quantitative inference. However, these gains exhibit diminishing returns and do not close the gap to either top proprietary models or human performance, indicating that scale alone is insufficient for faithful physical reasoning.

Gap to super-human performance. By collecting results from both humans and VLMs, we observe an interesting pattern. Importantly, human performance does not represent the theoretical ceiling for QUANTIPHY. While humans must rely on coarse visual approximations, an ideal agent with precise frame-level access to pixel coordinates could recover the world–pixel scale and compute target quantities exactly. In theory, VLMs should be capable of significantly outperforming humans by leveraging this exact pixel information to perform precise algebraic operations. The fact that the best current systems cluster around 50% MRA, comparable to or below human baselines, suggests they still fundamentally under-utilize visual precision and physical priors, leaving robust and accurate quantitative reasoning for kinematic inference tasks as an open challenge.

5. Dissecting Quantitative Reasoning in VLMs

Beyond directly evaluating VLMs as black boxes on kinematic inference tasks, we next take an inside look at *how* they arrive at their quantitative conclusions. Specifically, we investigate three aspects: (i) how scene context, such that background complexity and the number of objects, modulates task difficulty; (ii) the extent to which VLMs faithfully use the provided video and priors, rather than relying on memorized world knowledge; and (iii) whether structured prompting with step-by-step guidance can systematically improve kinematic inference.

5.1. Effect of Scene Context

For a more fine-grained analysis, we further categorize each video-question pair along a third axis describing the visual environment: *Scene Difficulty* $\in \{\texttt{SX}, \texttt{MX}, \texttt{SS}, \texttt{MS}, \texttt{SC}, \texttt{MC}\}$. The first letter indicates whether there is a single (S) or multiple (M) moving objects. The second letter indicates the background: a solid plain color (X)[3], a simple but textured scene (S), or a visually complex scene (C).

Figure 5 summarizes MRA across these categories. Overall, we observe that background complexity has only a mild effect on model's performance on the task. Models perform slightly *better* in the SAM-denoised condition than in the simple-texture (S) background, suggesting that removing irrelevant clutter reduces distractions and stabilizes quantitative estimates. Interestingly, performance in visually complex scenes (C) is above the other two background conditions for most models. A plausible explanation is that realistic backgrounds provide additional reference cues (e.g., tiles, windows, or road markings) that help the model infer scale and motion.

In contrast, the number of objects in the scene exhibits a clearer trend: setups with multiple objects (MX, MS, MC) consistently yield higher MRA than their single object

[3]Some X-type scenes are constructed by segmenting the target object with SAM2 and compositing it onto a uniform background, i.e., a fully "denoised" variant of the original video.

Models	Size	Kinematic Categories 2S	2D	3S	3D	Avg.
Proprietary models						
ChatGPT-5.1 [32]	–	**46.3**	56.2	**51.5**	**58.3**	**53.1**
Gemini-2.5 Pro [20]	–	44.8	**57.5**	42.4	53.7	49.6
Gemini-2.5 Flash [19]	–	40.3	53.2	43.6	57.4	48.6
Grok 4.1 (Fast Reasoning) [47]	–	39.4	49.5	42.4	48.6	45.0
ChatGPT-5 [31]	–	36.6	35.0	25.9	33.1	32.6
Claude Sonnet 4.5 [4]	–	19.6	23.0	19.6	29.1	22.8
Open-weight models						
Qwen3-VL-Instruct-32B [7]	32B	35.8	**51.6**	**43.2**	**53.4**	**46.0**
InternVL-3.5-30B [44]	30B	**36.7**	45.4	38.6	42.0	40.7
Qwen3-VL-Instruct-8B [7]	8B	26.0	47.8	35.1	46.3	38.8
InternVL-3.5-8B [44]	8B	27.3	41.8	34.4	38.3	35.4
Molmo-7B [15]	7B	30.0	43.1	24.4	36.6	33.5
Phi-4-Multimodal-Instruct [2]	5.6B	33.4	42.3	25.4	28.4	32.4
Qwen3-VL-Instruct-2B [7]	2B	27.1	39.0	17.6	32.1	29.0
SmolVLM-Instruct [30]	0.26B	31.6	34.4	20.0	27.8	28.5
InternVL-3.5-2B [44]	2B	25.0	31.1	16.6	27.4	25.0
VILA-7B [27]	7B	23.0	29.8	14.4	23.0	22.6
CogVLM2 Video [23]	12B	19.4	28.7	12.7	27.9	22.2
Phi-3-Mini-128K-Instruct-3.8B [1]	3.8B	17.3	14.7	19.5	18.6	17.5
LLaVA-13B [28]	13B	14.4	22.1	8.0	16.5	15.2
MiniCPM-V 4.5 [58]	8B	27.6	26.3	0.4	0.0	13.6
Fuyu-8B [10]	8B	9.5	14.7	9.5	16.2	12.5
Human Baseline	–	50.0	59.1	55.2	57.9	55.6

Table 1. Evaluation results on QUANTIPHY. We report Mean Relative Accuracy (MRA %) on four kinematic categories (2S, 2D, 3S, 3D) and their average. **Dark** cell marks the best overall model and **light** cell marks the best open-weight model.

Video + Prior	Prior only	Counterfactual	CoT
56.1	39.0	15.4	27.7
60.9	46.1	29.9	49.8
49.8	36.1	14.4	22.4
47.5	44.3	31.6	39.5
34.2	50.8	29.6	53.7
25.4	16.6	11.6	25.9
50.1	37.2	34.0	23.1
45.4	–	12.1	17.6
40.5	24.9	12.0	21.0
37.0	19.3	29.7	18.2
39.8	–	14.7	15.9
40.0	20.1	9.2	23.5
34.9	28.2	25.2	25.6
38.9	–	14.3	17.8
32.7	25.1	22.5	21.5
31.8	–	14.1	10.0
28.5	–	9.5	26.4
11.1	10.3	8.4	7.2
20.2	–	13.9	14.4
29.7	–	19.9	24.1
14.3	–	9.0	21.1
–	–	–	–

Table 2. Extensive results on an analysis subset. We report Mean Relative Accuracy (MRA) in %. Rows follow the same model order as in Table 1.

counterparts (SX, SS, SC). Having more objects gives the model extra reference targets (e.g., another ball, a ruler-like structure), which can be used as implicit comparison standards for both size and speed. These observations highlight that VLMs robustly benefit from richer background information and relational structure in the scene.

5.2. Do VLMs Use Videos and Priors Faithfully?

QUANTIPHY is designed to assess whether VLMs can truly understand physical events captured by visual sensors. It is therefore crucial to investigate whether these systems genuinely comprehend the provided visual signals, rather than merely producing *plausible* guesses. Here, we re-evaluate all models on a controlled subset of 1612D video-prior pairs,[4] using two complementary probes.

Key finding: VLMs rely more on learned prior knowledge than visual inputs for physical reasoning. We first compare the default *video+prior* setting against a *prior-only* setting in which the video is removed but the prompt (including the physical prior, object description, and question) is kept unchanged. Hypothetically, without the reference video, we would expect an *evident performance drop*, since VLMs would be forced to guess; even if such guesses are reasonable, they may not correspond to the specific instance. However, across most models we observe only modest differences between the two conditions: on the 161-pair subset, MRA in the *prior-only* setting is often close to, and sometimes only slightly below, the *video+prior* setting, even for strong systems such as ChatGPT-5.1, Gemini 2.5 Pro/Flash, and Qwen3-VL-Instruct-32B. In other words, models can already obtain reasonably high scores by relying on their *internal* prior knowledge about typical object sizes and speeds, with limited added value from the actual video frames. This suggests that, in our tasks, many VLMs behave less like visual measurers and more like powerful guessers conditioned on textual hints.

Key finding: VLMs (mostly) do not reason but memorize. To further test whether models faithfully utilize the conditioning prior and video to infer the target physical quantity, we perform a counterfactual analysis. For each of the 161 instances, we construct a family of counterfactual prompts by multiplying the original physical prior by a scalar factor $\alpha \in \{0.001, 0.01, 0.1, 0.2, 5, 50, 100, 200, 500, 700\}$, while keeping the video and question unchanged. Hypothetically, if a model is capable of correct kinematic reasoning, its prediction should scale accordingly, tracking the *counterfactual ground truth* $y^{\text{cf}} = \alpha \cdot y$. As a result, we would expect no, or at least only modest, performance drop in this test. However, in practice, we find that even the best models obtain very low scores in this setting: most models' MRA

[4] Due to resource constraints, the experiments in subsection 5.2 and subsection 5.3 are run on a 2D-only subset of 161 instances, and MRA values therefore differ slightly from those in Table 1.

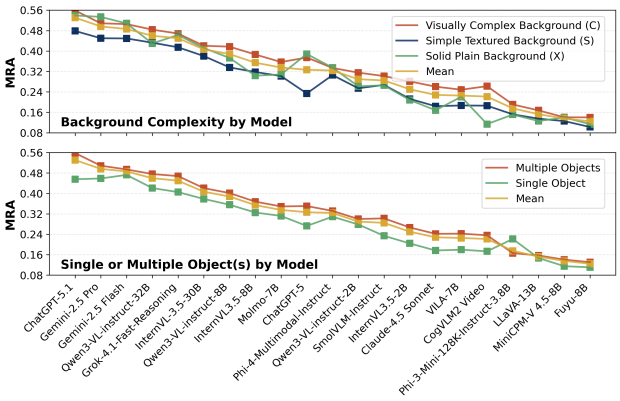


Figure 5. **Effect of scene context.** We plot the MRA (%) scores for all benchmark models on different categories, sorted in descending order according to their average MRA performance.

drops by 80%, and even the strongest model drops by 70%. Despite being given a numerically precise but altered prior, the outputs remain close to the original physical magnitudes implied by real-world experience, rather than those dictated by the provided priors.

Taken together, these results lead to a consistent conclusion: ***existing VLMs are not yet input-faithful quantitative reasoners.*** They only weakly exploit pixel-level information in videos, and they do not reliably condition on the exact numerical priors provided in the prompts. Instead, their quantitative kinematic inferences are dominated by internal, pre-trained world knowledge, with visual evidence and explicit priors acting more as soft hints than hard constraints.

5.3. Do Structured Prompts Help VLMs Reason?

Based on the above observation that VLMs rely more heavily on text prompts, we therefore further investigate whether a more structured "chain-of-thought" reasoning pattern can improve quantitative physical predictions. Instead of directly asking for the target quantity, we decompose each question into a four-step chain:

`(I) Pixel-level source property`: "What is [the ground-truth object's property] in pixels?"

`(II) Scale estimation`: "What is the proportional relationship between pixels and [a kinematic scale]?"

`(III) Pixel-level target property`: "What is [the inference target's property] in pixels?"

`(IV) World-level target property`: "What is [the inference target's real-world property]?"

We query each model sequentially on the same analysis subset as in subsection 5.2. If the model produces a parseable numerical answer at a given step, we append both the question and the extracted answer as additional context for the next step. If the model fails to generate a valid number at a step, subsequent prompts omit that step and its answer. We evaluate models' final answers with MRA.

As shown in Table 2, chain-of-thought (CoT) prompting is less helpful than anticipated. Among the 21 models we examined, only three show any improvement, with ChatGPT-5 and Fuyu-8B exhibiting a noticeable increase in MRA. For the remaining 19 models, including several strong open-weight systems, performance under our CoT protocol is *worse* than under direct zero-shot prompting, sometimes by a large margin. In other words, under our setup, explicitly spelling out pixel measurement, scale estimation, and rescaling does *not* systematically improve current VLMs' quantitative reasoning. Many models appear unable to reliably solve the intermediate numeric subproblems, so decomposing the task mainly amplifies and propagates early errors.

Our analysis suggests that existing VLMs can exploit visual cues, priors, and structured prompts to reason quantitatively, but do so in a rather brittle and often inconsistent manner. QUANTIPHY thus provides not only a benchmark for aggregate performance, but also a diagnostic tool for probing where quantitative physical reasoning succeeds, fails, and how it might be improved. We provide additional studies in the supplementary material Appendix A to further support our findings.

6. Discussion and Future Work

Summary of Findings. Our key finding is that state-of-the-art VLMs still fail to reliably connect visual observations with quantitative physical facts. This reflects a lack of *input faithfulness*: although models process videos, they rarely infer kinematic properties from pixel-level evidence. Instead, they rely heavily on parametric priors, often overriding explicit inputs and visual cues with memorized world knowledge. As a result, current systems behave more like approximate guessers than precise visual measurers, limiting their reliability for embodied agents.

Limitations. Our study has several limitations. The dataset focuses only on translational motion, excluding rotational dynamics, and uses a fixed camera perspective, making the task simpler than real-world settings with moving viewpoints. We also consider only rigid objects, not soft or deformable materials. Finally, the dataset remains relatively simple, with isolated motions rather than complex multi-object interactions.

Conclusion and Future Work. These limitations point to clear next steps. Future benchmarks should cover broader physical phenomena, including rotational dynamics, deformable objects, varying camera viewpoints, and complex multi-body interactions. Our findings may also motivate improved VLM training strategies, such as physics-informed objectives or pretraining on physics-rich data. Ultimately, this line of work aims to support the development of embodied AI agents that can reason about and interact with the physical world more effectively.

Acknowledgment

The authors thank Juze Zhang and Heng Yu for their help with lab data capturing. This work was partially funded by the NIH Grant R01AG089169, Stanford HAI Hoffman-Yee Award, Stanford HAI graduate fellowship, and UST.

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024. 7, 33

[2] Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025. 7, 33

[3] Kelsey R Allen, Kevin A Smith, and Joshua B Tenenbaum. Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, 117(47):29302–29310, 2020. 2

[4] Anthropic. Claude 4.5 sonnet [large language model]. https://www.anthropic.com/news/claude-sonnet-4-5, 2025. Accessed: 2026-03-26. 5, 7, 30

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2

[6] Alisson Azzolini, Hannah Brandon, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, Francesco Ferroni, Rama Govindaraju, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025. 2

[7] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 7, 33

[8] Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. PHYRE: A New Benchmark for Physical Reasoning, 2019. 2

[9] Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint arXiv:2503.06800*, 2025. 2

[10] Rohan Bavishi, Chris Clark, Arnav Mishra, et al. Fuyu-8b: A multimodal architecture for ai agents, 2023. Official model announcement; accessed 2026-03-26. 7, 33

[11] Steven S. Beauchemin and John L. Barron. The computation of optical flow. *ACM computing surveys (CSUR)*, 27(3):433–466, 1995. 2

[12] Zhongang Cai, Yubo Wang, Qingping Sun, Ruisi Wang, Chenyang Gu, Wanqi Yin, Zhiqian Lin, Zhitao Yang, Chen Wei, Hui En Pang, Xuanke Shi, Kewang Deng, Xiaoyang Han, Zukai Chen, Jiaqi Li, Xiangyu Fan, Hanming Deng, Lewei Lu, Bo Li, Ziwei Liu, Quan Wang, Dahua Lin, and Lei Yang. Holistic evaluation of multimodal llms on spatial intelligence. *arXiv preprint arXiv:2508.13142*, 2025. 4

[13] Zhe Chen and et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic capabilities. *arXiv preprint arXiv:2312.14238*, 2024. 33

[14] Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding, 2025. 2, 4

[15] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 91–104, 2025. 7, 33

[16] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. Palm-e: An embodied multimodal language model. 2023. 1

[17] Tobias Gerstenberg, Noah D Goodman, David A Lagnado, and Joshua B Tenenbaum. A counterfactual simulation model of causal judgments for physical events. *Psychological review*, 128(5):936, 2021. 2

[18] Andrea Giachetti, Marco Campani, and Vincent Torre. The use of optical flow for road navigation. *IEEE transactions on robotics and automation*, 14(1):34–48, 2002. 2

[19] Google. Gemini 2.5 flash [large language model]. https://deepmind.google/models/gemini/flash/, 2025. Accessed: 2025-11-14. 5, 7, 30

[20] Google. Gemini 2.5 pro [large language model]. https://deepmind.google/models/gemini/pro/, 2025. Accessed: 2025-11-14. 5, 7, 30

[21] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022. 1

[22] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11287–11297, 2021. 2

[23] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 7, 33

[24] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 2

[25] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How Far is Video Generation from World Model: A Physical Law Perspective, 2024. 2

[26] Shiqian Li, Kewen Wu, Chi Zhang, and Yixin Zhu. On the learning mechanisms in physical reasoning. *Advances*

in Neural Information Processing Systems, 35:28252–28265, 2022. 2

[27] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26689–26699, 2024. 7, 33

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 7, 33

[29] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 1

[30] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025. 7, 33

[31] OpenAI. Chatgpt version 5 [large language model]. https://openai.com/index/introducing-gpt-5/, 2025. Accessed: 2026-03-26. 5, 7, 30

[32] OpenAI. Chatgpt version 5.1 [large language model]. https://openai.com/index/gpt-5-1/, 2025. Accessed: 2026-03-26. 5, 7, 30

[33] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 4, 20

[34] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[35] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, et al. Grounding dino 1.5: Advance the" edge" of open-set object detection. *arXiv preprint arXiv:2405.10300*, 2024. 20

[36] Ilyse Resnick, Alexandra Davatzes, Nora S Newcombe, and Thomas F Shipley. Using relational reasoning to learn about scientific phenomena at unfamiliar scales. *Educational Psychology Review*, 29(1):11–25, 2017. 2

[37] Keattisak Sangsuwan and Mongkol Ekpanyapong. Video-based vehicle speed estimation using speed measurement metrics. *IEEE Access*, 12:4845–4858, 2024. 2

[38] Sketchfab, Inc. Sketchfab. https://sketchfab.com/, 2025. Accessed: 2025-11-14. 4

[39] Deepak* Sridhar, Kartikeya* Bhardwaj, Jeya Pradha Jeyaraj, Nuno Vasconcelos, Ankita Nayak, and Harris Teague. Video reasoning without training, 2025. 4

[40] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016. 2

[41] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024. 1

[42] Hsiao-Yu Tung, Mingyu Ding, Zhenfang Chen, Daniel M Bear, Chuang Gan, Joshua B Tenenbaum, Daniel L K Yamins, Judith Fan, and Kevin A Smith. Physion++: Evaluating Physical Scene Understanding that Requires Online Inference of Different Physical Properties. 2

[43] Quan Vuong, Sergey Levine, Homer Rich Walke, Karl Pertsch, Anikait Singh, Ria Doshi, Charles Xu, Jianlan Luo, Liam Tan, Dhruv Shah, et al. Open x-embodiment: Robotic learning datasets and rt-x models. In *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023*, 2023. 1

[44] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 7

[45] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17868–17879, 2024. 1, 28

[46] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024. 2

[47] xAI. Grok 4.1 fast reasoning [large language model]. https://docs.x.ai/grok-4.1, 2025. Accessed: 2025-12. 5, 7, 33

[48] Tiange Xiang, Adam Sun, Scott Delp, Kazuki Kozuka, Li Fei-Fei, and Ehsan Adeli. Wild2avatar: Rendering humans behind occlusions. *arXiv preprint arXiv:2401.00431*, 2023. 1

[49] Tiange Xiang, Adam Sun, Jiajun Wu, Ehsan Adeli, and Li Fei-Fei. Rendering humans from object-occluded monocular videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3239–3250, 2023.

[50] Tiange Xiang, Kuan-Chieh Wang, Jaewoo Heo, Ehsan Adeli, Serena Yeung, Scott Delp, and Li Fei-Fei. Neuhmr: Neural rendering-guided human motion reconstruction. In *2025 International Conference on 3D Vision (3DV)*, pages 1518–1528. IEEE, 2025. 1

[51] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 2

[52] Xinrun Xu, Pi Bu, Ye Wang, Börje F. Karlsson, Ziming Wang, Tengtao Song, Qi Zhu, Jun Song, Zhiming Ding, and Bo Zheng. DeepPHY: Benchmarking Agentic VLMs on Physical Reasoning. *arXiv preprint arXiv:2508.05405*, 2025. 2

[53] Jihan Yang, Shusheng Yang, Anjali 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. 4

[54] 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 Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025. 2, 4, 5

[55] Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025. 2

[56] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31, 2018. 2

[57] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019. 2

[58] Tianyu Yu, Zefan Wang, Chongyi Wang, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe, 2025. 7, 33

[59] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, pages 1–21. Springer, 2022. 2

[60] Zhicheng Zheng, Xin Yan, Zhenfang Chen, Jingzhou Wang, Qin Zhi Eddie Lim, Joshua B Tenenbaum, and Chuang Gan. Contphy: Continuum physical concept learning and reasoning from videos. *arXiv preprint arXiv:2402.06119*, 2024. 2